AGL1K: An Audio Geo-Localization Benchmark for Audio-Language Models

Anonymous ACL submission

Abstract

Geo-localization aims to infer the geographic origin of a given signal. In computer vision, geo-localization has served as a demanding benchmark for compositional reasoning and is relevant to public safety. In contrast, progress on audio geo-localization has been constrained by the lack of high-quality audio-location pairs. To address this gap, we introduce AGL1K, the first audio geo-localization benchmark for audio language models (ALMs), spanning 72 countries and territories. To extract reliably localizable samples from a crowd-sourced platform, we propose the Audio Localizability metric that quantifies the informativeness of each recording, yielding 1,444 curated audio clips. Evaluations on 16 ALMs show that ALMs have emerged with audio geo-localization capability. We find that closed-source models substantially outperform open-source models, and that linguistic clues often dominate as a scaffold for prediction. We further analyze ALMs reasoning traces, regional bias, error causes, and the interpretability of the localizability metric. Overall, AGL1K establishes a benchmark for audio geo-localization and may advance ALMs with better geospatial reasoning capability.

1 Introduction

Geo-localization aims to infer the geographic origin of a signal and offers a compelling alternative to standard classification because it requires compositional reasoning over diverse clues. Mapping observations to a single GPS coordinate not only requires perceptual inference but also broad world knowledge of geography and culture. In computer vision, this problem is typically studied as image geo-localization (Weyand et al., 2016; Regmi and Shah, 2019; Li et al., 2024). As for audio, it can also provide analogous evidence, including reverberation patterns, traffic density, and coastal wave dynamics. Beyond academic interest, audio geo-localization has clear societal value. For example, assisting investigations of audios disseminated by extremist actors can help fact-checkers verify the claimed location of viral content. These applications make audio geo-localization relevant to public safety. However, despite this potential, **a systematic benchmark for audio geo-localization remains absent**.

The lack of audio geo-localization benchmarks stems from two main factors. First, **there is no publicly available audio dataset with location annotations**. In contrast to image geo-localization, where progress has been enabled by large-scale geo-tagged data from social media platforms, no comparable resource exists for audio. Second, **the field lacks a quantitative notion of audio localizability to filter for geographically informative recordings**. Even if crowd-sourced platforms can provide large numbers of audio-location pairs, without such a measure it remains nontrivial to identify samples that carry meaningful geographic signals.

To systematically evaluate audio geo-localization capability in audio language models (ALMs), we introduce AGL1K, the first audio geo-localization benchmark for ALMs. AGL1K is curated from the crowd-sourced Aporee platform and filtered using our proposed Audio Localizability metric, which estimates the geographic informativeness of each recording by aggregating evidence from both positive and negative sound categories during inference. The resulting benchmark spans 72 countries across six continents and covers diverse acoustic scenes, including nature soundscapes, animal vocalizations, music, human-made sounds, and spoken conversations. **This diversity makes AGL1K a suitable testbed for assessing compositional reasoning in modern ALMs, with potential downstream relevance to public safety and misinformation detection.**

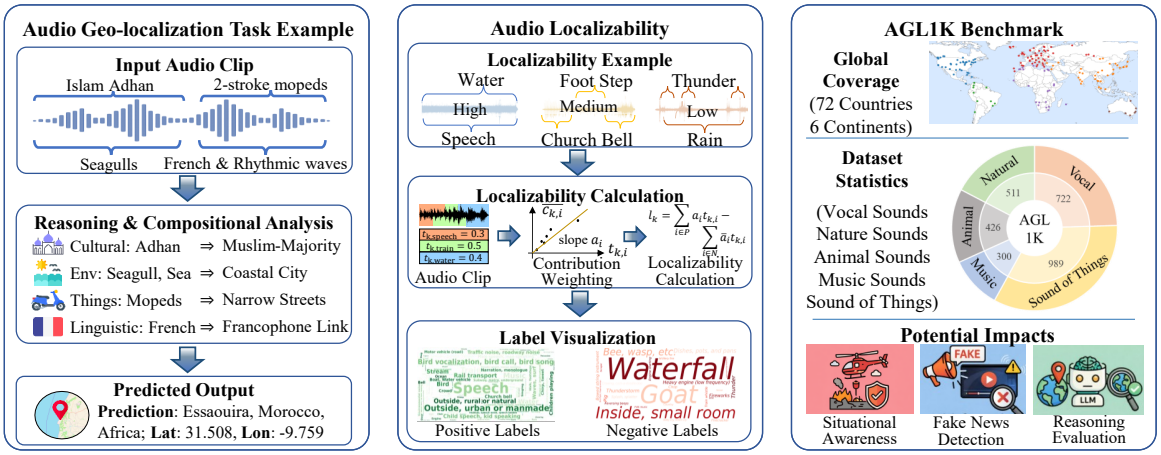

Figure 1: Overview of AGL1K, illustrating the audio geo-localization task, localizability, and composition.

Our comprehensive evaluation of 16 ALMs shows that current ALMs have begun to exhibit meaningful audio geo-localization capability, yet AGL1K remains challenging. The results reveal a clear capability hierarchy between closed- and open-source models. Leading ALMs (e.g., Gemini 3 Pro) demonstrate strong knowledge and reasoning, with failures increasingly driven not by missing information but by over-commitment to a single clue. In contrast, open-source models exhibit more fundamental limitations in fine-grained perception, which prevents reliable extraction of geographically informative signals.

Through detailed error analysis, we conclude three findings suggesting directions for improving future audio models. (1) **Enhance fine-grained perception**: Many open-source ALMs misidentify languages and other subtle acoustic clues, limiting perceptual sensitivity and downstream knowledge retrieval. (2) **Mitigate regional bias:** Systematic prediction imbalance such as over-predicting certain continents or regions, persists across models. (3) **Strengthen compositional reasoning**: As audio geo-localization requires integrating multiple weak clues, models must avoid relying on any single clue as decisive evidence.

In summary, this work makes the following contributions:

- **The first audio geo-localization benchmark for ALMs.** We introduce AGL1K, a benchmark comprising 1,444 user-uploaded audio clips from 72 countries, covering diverse acoustic scenes including nature soundscapes, animal vocalizations, music, human-made sounds, and spoken conversations.
- **A principled notion of Audio Localizability.** We propose an Audio Localizability, a quantitative measure of an audio's geographic informativeness. The metric aggregates informativeness from positive and negative sound labels during inference, enabling filtering localizable recordings.
- **A comprehensive evaluation of state-of-the-art ALMs.** We benchmark 16 ALMs and find the emergence of audio geo-localization to some extent. We further analyze representative models' reasoning traces, regional biases, and error causes to provide future insights for improving ALMs.

2 Related Work

2.1 Geo-Localization

Most existing geo-localization research focuses on image-based geo-localization, which can be broadly categorized into three paradigms. **Classification-based** approaches discretize the Earth's surface into predefined regions, but achieving fine-grained accuracy requires a large number of classes, making training and scalability challenging (Weyand et al., 2016; Clark et al., 2023; Müller-Budack et al., 2018; Seo et al., 2018). **Retrieval-based** methods embed images and geographic coordinates into a shared representation space and localize by nearest-neighbor search. However, they rely on massive, globally distributed, annotated image databases that are costly to construct and maintain (Regmi and Shah, 2019; Shi et al., 2019, 2020; Cepeda et al., 2023). More recently, **vision-language model-**

based methods exploit the world knowledge encoded in large multimodal models to directly predict locations, demonstrating promising performance (Li et al., 2024; Han et al., 2025; Wang et al., 2025b). In contrast, audio geo-localization remains unexplored. Existing benchmarks are limited in diversity, often focusing on narrow domains such as bird vocalizations (Chasmai et al., 2025).

2.2 Audio-Language Models

The development of ALMs has closely followed the advances in deep learning. Early breakthroughs such as Deep Speech 2 (Amodei et al., 2016) replaced traditional multi-stage pipelines with end-to-end neural networks, achieving near-human transcription performance. Subsequent progress in self-supervised and weak-supervised learning further advanced the field, exemplified by Whisper (Radford et al., 2023), which demonstrated multilingual transcription performance comparable to supervised methods. With the rise of large model pretraining, audio has emerged as a first-class modality within unified architectures. Recent foundation models such as GPT-4o (Achiam et al., 2023), Gemini3 (Google DeepMind, 2025), Qwen3-Omni (Xu et al., 2025b).

A wide range of benchmarks has been proposed to evaluate ALMs, covering basic Speech-to-Text Translation (OpenAudioBench (Wang et al., 2025a)), Vocal Sound Classification (VocalSound (Gong et al., 2022)), as well as higher-level reasoning tasks such as spatial reasoning (StarBench (Liu et al., 2025)) and causal discovery (MECAT (Niu et al., 2025)). However, there is still no dedicated benchmark for assessing audio geo-localization ability in ALMs.

3 Audio Geo-Localization Benchmark

The Audio Geo-Localization Benchmark (AGL1K) is designed to evaluate the geographic compositional reasoning capabilities of audio-language models. We next describe how we obtain large-scale audio-location pairs (3.1), perform coarse quality control (3.2), identify reliably localizable recordings using our proposed Audio Localizability metric (3.3) and post-processing after obtaining ALMs' response (3.4).

3.1 Dataset Acquisition

Global-scale audio geo-localization has received limited attention, largely due to the scarcity of GPS-tagged audio recordings. We identify Aporee[1], a crowd-sourced platform where users upload geo-tagged audio worldwide. In collaboration with the Aporee team, we obtain tens of thousands of audio-location pairs released under an Apache license. Additional details about Aporee are provided in Appendix A.1.

3.2 Initial Filtering

Because Aporee recordings are user-uploaded, their quality varies substantially. To coarsely remove low-quality or uninformative samples, we apply four acoustic filters targeting common failure causes: low signal energy, noise-like content, clipping, and overly monotonous recordings. Specifically, we compute (1) *RMS Energy* to exclude extremely low-amplitude audio, (2) *Spectral Flatness* to remove noise-like signals, (3) *Clipping Ratio* to discard heavily clipped recordings, and (4) *Acoustic Complexity* to eliminate clips with limited temporal variation. These filters retain recordings with sufficient acoustic structure and potential geographic clues. Formal definitions are provided in Appendix A.3.

3.3 Audio Localizability

3.3.1 Localizability Calculation

While the initial acoustic filters remove low-quality recordings, they do not ensure that the remaining samples contain enough geo-informativeness evidence. To construct a benchmark with localizable inputs, we introduce a principled measure of **audio localizability**.

We model the localizability of an audio recording as the net contribution of *informative* (positive) versus *uninformative* (negative) sound categories. Intuitively, longer exposure to informative categories (e.g., language, place-specific human activities) should improve geo-localizability, whereas prolonged uninformative content (e.g., infant crying) can obscure relevant clues. Categories also contribute unequally: coast sounds may provide coarse evidence, while linguistic content often offers finer specificity. Motivated by this, we define the localizability of sample k as

$$l_k = \sum_{i \in P} a_i t_{k,i} - \sum_{i \in N} \bar{a}_i t_{k,i}. \quad (1)$$

where P and N denote the sets of positive and negative sound categories, respectively; $t_{k,i} \in [0, 1]$ is

[1] https://aporee.org/maps/

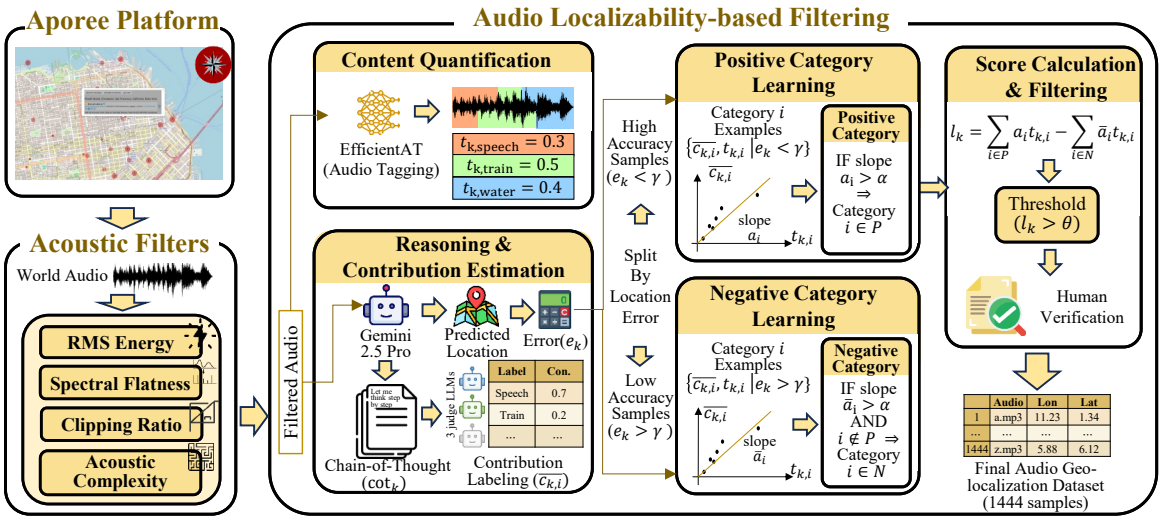

Figure 2: **Overview of the benchmark construction framework.** AGL1K is curated from the crowd-sourced Aporee platform. The recordings are first filtered using four acoustic filters, followed by our proposed Audio Localizability measure, which quantifies the geo-informativeness of each sample.

the fraction of time that category i is present in sample k; and a_i (resp. $\bar{a}_i$) quantifies the contribution strength of category i when it acts as a positive (resp. negative) category. We obtain $t_{k,i}$ by tagging each recording with EfficientAT (Schmid et al., 2023) under the AudioSet ontology (Gemmeke et al., 2017). Estimating category-level contributions is therefore central to assessing localizability.

Because audio geo-localization lacks established expert heuristics, we derive category contributions from the reasoning behavior of a strong ALM. For each sample k, we run a Gemini 2.5 to produce a predicted location and a chain-of-thought cot_k, and compute the distance error e_k between the prediction and the ground-truth coordinates. We then prompt three language models to assess, based on cot_k, the contribution of each detected audio category to the model's reasoning, using a five-level discrete scale. Averaging across the three judges yields $\bar{c}_{k,i}$, the estimated contribution of category i for sample k.

To determine which category acts as a positive versus a negative category, we fit simple linear relationships between category duration and judged contribution. For samples with $\{k|e_k < \gamma\}$, we fit $y = a_i x$ with $x = t_{k,i}$ and $y = \bar{c}_{k,i}$, and include category i in P if $a_i > \alpha$. For samples with $\{k|e_k \geq \gamma\}$, we fit $y = \bar{a}_i x$ and include category i in N if $\bar{a}_i > \alpha$ and $i \notin P$. Using Equation 1, we deem a recording highly localizable if $l_k > \theta$.

Applying the above criterion yields a pool of highly localizable recordings. From this pool, we manually curate 1,444 high-quality clips, balanced between samples with and without human speech, to form the final dataset used in our audio geo-localization benchmark. We set $\alpha = 1/3$,$\gamma = 1000$ km and $\theta = 1$ empirically.

3.3.2 Top Positive and Negative Categories

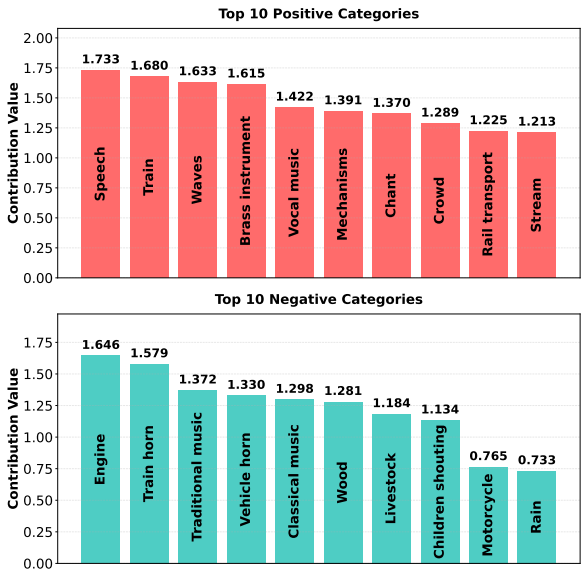

Figure 3: **Top Positive and Negative Categories.**

To evaluate the quality of the proposed localizability metric, we present the top 10 positively

and negatively contributing sound categories (Figure 3). Several clear patterns emerge: Speech is the most prominent positive contributor, highlighting the strong geographic specificity encoded in linguistic cues. Additionally, sounds with inherently regional distributions such as rail transport and waves also exhibit high localizability. In contrast, generic noise sources like engine or train horn, as well as globally pervasive natural sounds such as rain and wood, tend to have strong negative contributions, likely due to their ubiquity and limited geographic discriminability. Therefore, **these results demonstrate that the learned feature attributions align well with human intuition, supporting the interpretability and validity of the localizability metric.**

3.3.3 Localizability Examples

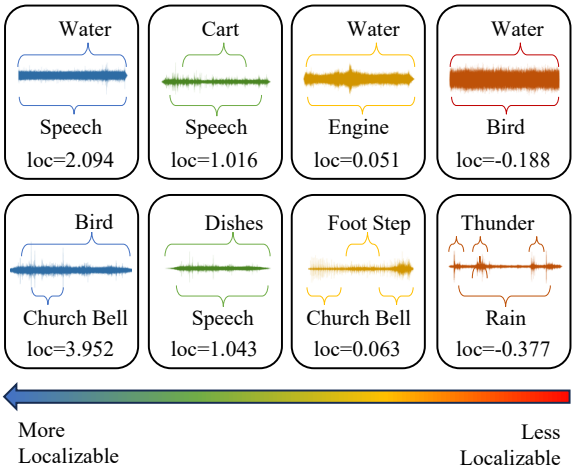

Figure 4: **Localizability Examples.** Localizability increases from right to left.

Figure 4 visualizes the computed localizability for eight example clips. Recordings dominated by thunder and rain yield the lowest scores, likely because these sounds are ubiquitous across most regions worldwide. Although clips containing footsteps or church bells are also weakly localizable, the presence of birdsong can substantially increase localizability, as species-specific vocalizations provide stronger geographic clues. For the clips in the second column, speech is present, but it is heavily masked by indoor foreground noises (e.g., carts and dishware), which markedly weakens geographic informativeness. Overall, **the localizability scores are consistent with human intuition and support our metric as a reliable basis for filtering localizable data.**

3.4 Post Processing

Following the above pipeline, we obtain AGL1K, a benchmark contains a broad range of audio categories including *nature sounds*, *animal vocalizations*, *music*, and *sounds of things* as well as many recordings that couple multiple categories and thus provide rich geo-informative clues. Geographically, the benchmark spans six continents and covers more than 72 countries and territories.

After collecting ALM predictions, we observed that models often use aliases or informal variants for the same region (e.g., *Oceania* being reported as *Australia*, or *United States* as *USA*). To ensure consistent evaluation, we normalize model outputs to a unified location schema by mapping such variants to canonical region names, thereby improving the reliability of our benchmark results.

In addition, we have developed an interactive audio geo-localization website, which can be deployed on Hugging Face Spaces to allow global users to explore and engage with the task. Please refer to Appendix A.2.

4 Experiments

In this section, we investigate the following research questions:

RQ1: How do current audio-language models perform on the proposed benchmark?

RQ2: How do audio-language models infer potential geographic locations from acoustic clues?

RQ3: Do audio-language models exhibit geographic prediction inequality across regions?

RQ4: What are the primary causes of model prediction errors in audio geo-localization?

4.1 Experimental Settings

4.1.1 Benchmarking Models

We benchmarked 8 state-of-the-art closed-source ALMs including: GPT-4o Audio Preview (Achiam et al., 2023), Gemini 3 Pro (Google DeepMind, 2025), Gemini 2.5 Pro (Comanici et al., 2025), Gemini 2.5 Flash, Gemini 2.5 Flash-Lite, Gemini 2.0 Flash-Thinking (Google, 2025), Gemini 2.0 Flash, Gemini 2.0 Flash-Lite. We also benchmarked 8 high-quality open-source ALMs including: Qwen3-Omni, Qwen2.5-Omni (Xu et al., 2025a), Phi-4-MM1 (Abouelenin et al., 2025), Kimi-Audio (KimiTeam et al., 2025), Gemma-3n-E4B-it (Team et al., 2025), MiniCPM-o-2.6 (Yao et al., 2024), Mimo-audio (Xiaomi, 2025), and Mimo-audio-think (Xiaomi, 2025).

Table 1: Audio geo-localization performance of different models. The best results within the closed-source and open-source model groups are separately bolded, while the second-best results are underlined in each group.

Model	Geo-score↑	Distance↓	Cont. acc.↑	Country acc.↑	City acc.↑	Reject rate↓	<1 acc.↑	<10 acc.↑	<500 acc.↑	Speech dis.↓	Non-speech dis.↓
RANDOM	112.16	9869.01	0.14	0.01	0.00	0.00	0.00	0.00	0.00	9930.88	9717.65
Closed-Source Models											
Gemini 3 Pro	**3031.82**	**2180.57**	**0.82**	**0.51**	**0.11**	0.01	**0.07**	**0.19**	**0.52**	**1548.29**	**3727.33**
Gemini 2.5 Pro	2826.95	2521.97	0.78	0.49	**0.11**	**0.00**	0.05	0.17	0.48	1634.22	4693.67
Gemini 2.5 Flash	2256.22	3558.37	0.65	0.39	0.07	0.01	0.02	0.12	0.36	2689.96	5682.74
Gemini 2.5 Flash-Lite	1687.97	4373.89	0.55	0.28	0.03	0.02	0.01	0.06	0.21	3742.97	5917.32
Gemini 2.0 Flash-Thinking	2494.46	2991.51	0.73	0.39	0.07	0.01	0.02	0.12	0.38	2265.47	4767.61
Gemini 2.0 Flash	2535.49	2906.31	0.73	0.40	0.08	0.01	0.02	0.13	0.41	2189.30	4660.31
Gemini 2.0 Flash-Lite	2417.14	3223.85	0.71	0.38	0.06	0.04	0.01	0.11	0.37	2296.28	5492.96
GPT-4o Audio Preview	1996.15	4067.87	0.61	0.37	0.05	0.04	0.02	0.10	0.30	2841.03	7069.09
Open-Source Models											
Qwen3-Omni	**1498.92**	5174.36	0.47	0.25	0.02	0.04	**0.01**	**0.05**	**0.20**	4741.73	6235.23
Qwen2.5-Omni	1317.74	5476.83	0.45	**0.26**	0.02	0.31	0.00	0.04	0.17	4540.06	7777.20
Phi-4-MM1	826.88	6462.43	0.33	0.08	0.01	0.15	0.00	0.01	0.08	6311.65	6831.08
Kimi-Audio	1336.20	5590.20	0.43	0.22	0.02	0.37	0.00	0.04	0.18	4634.32	7928.59
Gemma-3n-E4B-it	1161.54	5815.46	0.41	0.17	0.01	0.05	0.00	0.04	0.14	5402.73	6825.13
MiniCPM-o-2.6	1080.05	6600.83	0.44	0.22	0.02	0.54	0.00	0.01	0.14	5572.78	9115.74
Mimo-Audio	1445.39	**4853.25**	**0.54**	0.20	**0.03**	0.11	0.00	0.03	0.16	**4113.93**	6661.84
Mimo-Audio-Think	1447.30	5008.01	0.51	0.20	**0.03**	**0.02**	0.00	0.04	0.18	4599.73	**6006.78**

For open-source models except for Qwen3-Omni, we deploy them on a RTX 4090. Other models are accessed via their APIs with default settings.

4.1.2 Evaluation Metrics

The metrics includes mean distance error(km) δ, Geoscore based on Geoguessr game defined as $5000\exp(-\delta/1492.7)$, hierarchical continent/-country/city level accuracies, and thresholded accuracy at 1 km, 10 km, 500 km (denoted at $<*$) to capture both fine-grained and coarse localization performance. We also report a models reject rate (fraction of non-answers) as an indicator of robustness. Finally, to analyze reliance on speech clues, we compute average distance errors separately for speech and non-speech audio segments. The definition of each metric can be found in the Appendix A.4.

4.2 Comparison Results (RQ1)

Modern audio-language models have begun to exhibit emergent capabilities in audio geo-localization. Table 2 reports the performance of all models on the Audio Geo-Localization Benchmark. It shows Gemini 3 Pro achieves the strongest overall results, with 19% of examples localized within 10 km, continent-level accuracy of 0.82, and country-level accuracy of 0.51. However, a pronounced gap exists between closed-source and open-source models: the best open-source model, Mimo-audio, records an average error of 4853 km. It is more than twice that of Gemini 3 Pro.

Audio geo-localization performance scales systematically with model capacity. Within the Gemini family, we observe a clear advance. Distance error decreases from 2992 km in Gemini 2.0 Flash-Thinking to 2181 km in Gemini 3 Pro, alongside steady gains in various accuracies. Notably, some Gemini 2.0 Flash variants outperform Gemini 2.5 Flash, implying that the distilled model may prioritize efficiency or task-specific objectives that do not directly align with audio geo-localization. However, reasoning-enhanced variants (e.g., thinking modes) do not yield consistent improvements, suggesting that explicit reasoning alone provides limited benefit.

Linguistic content serves as a dominant scaffold for audio-based geo-localization. For Gemini 3 Pro, the average localization error drops from 3727 km on non-speech audio to **1548 km** when speech is present. Similar gaps appear across all evaluated audio-language models from closed-source to open-source models. This result highlights the dominant role of linguistic clues, while purely environmental sounds remain more challenging.

Additional statistics and comparisons are provided in Appendix A.6.

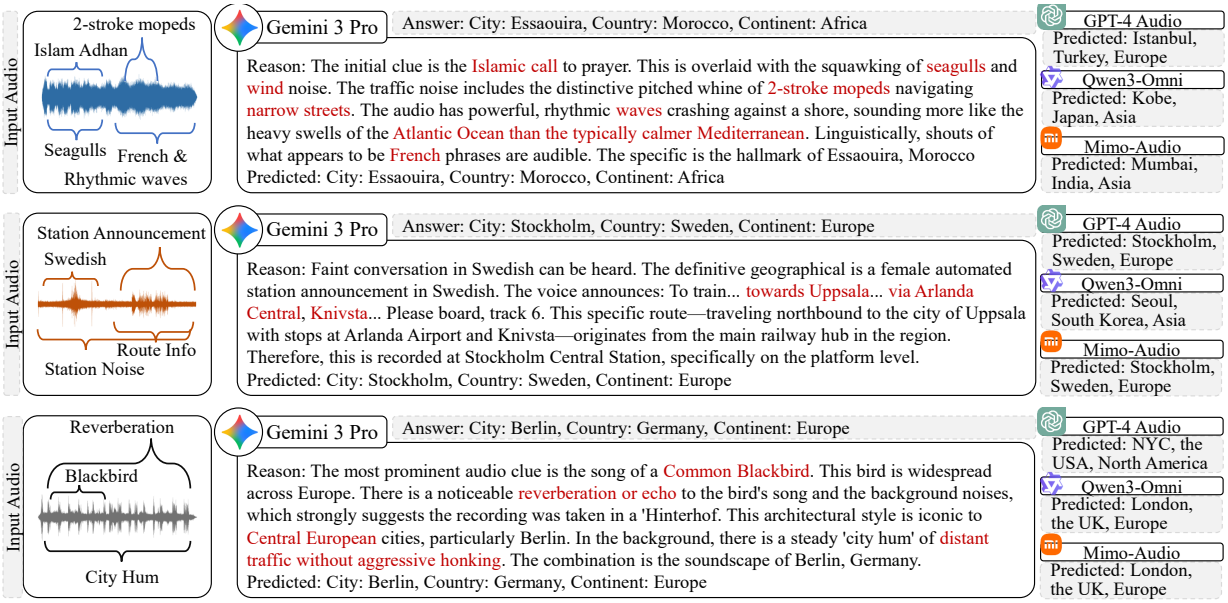

Figure 5: **Benchmark examples.** We select three representative audio samples and present the distribution of their audio clues, the reasoning process of Gemini 3 Pro, and predictions of three other ALMs.

4.3 Benchmark Examples (RQ2)

To qualitatively examine the reasoning behavior of modern audio models, we analyze Gemini 3 Pro's reasoning output across three representative scenarios, shown in Figure 5, along with the prediction of GPT-4o Audio, Qwen3-Omni, and Mimo-Audio. The complete reasoning traces are provided in the Appendix B.6.

The first case requires integrating linguistic and environmental signals. In addition to recognizing Islamic Adhan, Gemini 3 Pro detects coastal wind, seagull calls, French Speech, and motorcycle noise in narrow streets, jointly localizing the recording to Essaouira, Morocco. However, most other models overemphasize Adhan and lead to predictions of Istanbul or Mumbai.

The second case tests geographic inference from indirect clues. The audio references travel toward Uppsala via Arlanda and Knivsta, leveraging regional transit knowledge. Gemini 3 Pro, GPT-4o, and Mimo-Audio correctly infer Stockholm as the recording location.

This third case contains no spoken language. Instead, Gemini 3 Pro relies on non-verbal clues, including the call of a Common Blackbird, architectural style, and background traffic patterns, to correctly localize the scene to Berlin, Germany. Qwen3 and Mimo, however, identified the European bird and randomly guessed a place in Europe without further thinking.

Together, these case studies show that **leading audio models can exploit linguistic, environmental, and contextual clues to perform geographic reasoning**, and **our benchmark captures these compositional reasoning modes in a unified evaluation framework.**

4.4 Continent-Level Prediction Inequality (RQ3)

Figure 6 shows continent-level confusion (row-normalized). Gemini 3 Pro has the strongest diagonal, meaning the most stable continent recognition, especially for Africa/Asia/Europe and performance drops on Oceania/South America. GPT-4o Audio is less concentrated on the diagonal. It makes broader cross-continent mistakes, and Oceania is particularly unstable, often being predicted as other continents. Qwen3-Omni shows clear label collapse toward North America: many samples from other continents are predicted as North America, leading to poor results on Oceania and South America. Overall, these results indicate that **geographic prediction inequality is a persistent issue across audio-language models, highlighting the need to explicitly account for regional fairness in future work**.

4.5 Error Distribution (RQ4)

To analyze model failure causes, we annotate 300 country-level misclassifications for three represen-

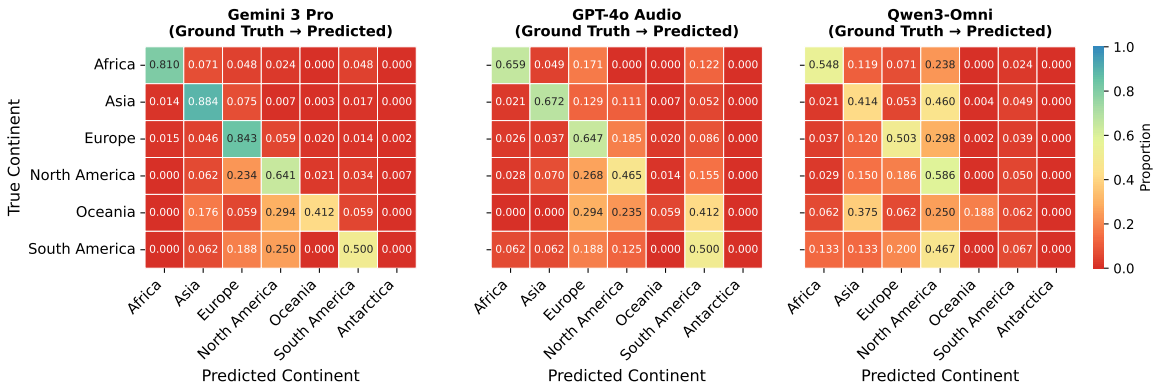

Figure 6: The continent-level prediction inequality in audio geo-localization. The i-j entry indicates that the truth is the portion of the continent in the i-th line that is predicted to be continent in the j-th row.

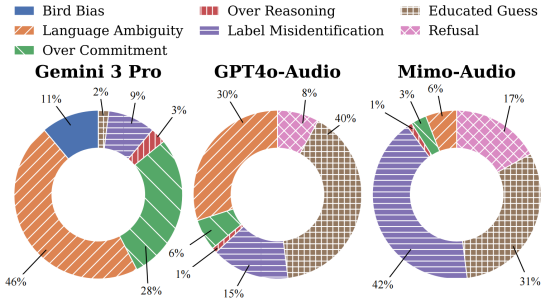

Figure 7: **Error distribution across three models.**

tative models. Errors are grouped into seven categories: Bird Bias, Language Ambiguity, Over-Commitment, Over-Reasoning, Label Misidentification, Educated Guess, and Refusal, capturing both perceptual and reasoning-related failures. The definition of each error is elaborated in the Appendix B.

Figure 7 shows that error distributions differ substantially across models. For Gemini 3 Pro, Language Ambiguity dominates, followed by Over-Commitment and Bird Bias. This reflects a tendency to exploit fine-grained acoustic clues, particularly language and bird vocalizations, which improves localization accuracy but also introduces distinctive errors. In contrast, GPT-4o Audio is primarily affected by Educated Guess and Language Ambiguity, indicating frequent fallback to uncertain predictions when informative clues are weak. Mimo-Audio exhibits a different pattern, with Label Misidentification accounting for the largest share of errors, followed by Educated Guess and Refusal, suggesting limitations in multilingual audio discrimination and confidence calibration.

In summary, this experiment suggests several actionable directions for improving audio geo-localization across model families. At the perception level, strengthening multilingual speech recognition is essential for reducing language-related errors. At the reasoning level, models should adopt more deliberate and evidence-aware reasoning to minimize reliance on uninformed educated guesses. At the fusion level, models should avoid over-dependence on single clues, such as language or bird vocalizations, and instead integrate multiple complementary audio signals. Together, **these findings provide practical guidance for enhancing the accuracy and robustness of audio-based localization systems.**

5 Conclusion

We introduce AGL1K, a benchmark for audio geo-localization. Built on user-uploaded recordings and filtered by our proposed audio localizability, the benchmark comprises 1,444 audio clips from 72 countries and territories across six continents, covering diverse acoustic scenes. Through the evaluation of 16 ALMs, we show that geo-localization ability has emerged in ALMs while AGL1K still remains challenging. Our analyses identify major error modes and suggest directions for improving ALMs, including strengthening fine-grained perception, reducing regional bias, and enhancing compositional reasoning. As audio geo-localization requires compositional reasoning on audio clues, we expect audio geo-localization to emerge as an important benchmark for evaluating ALMs' reasoning capability.

6 Limitations

- Due to the inherent distribution of the original data, the number of samples is uneven across continents. In particular, Europe, Asia, and North America are significantly overrepresented compared to other regions such as Africa, Oceania, and South America.
- The diversity of sound labels, as well as the distribution of positive and negative attribution results, is partially constrained by the choice of annotation models and the collected data.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse H. Engel, Linxi Fan, Christopher Fougner, Awni Y. Hannun, Billy Jun, Tony Han, and 20 others. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 173–182. JMLR.org.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mustafa Chasmai, Wuao Liu, Subhransu Maji, and Grant Van Horn. 2025. Audio geolocation: A natural sounds benchmark. *CoRR*, abs/2505.18726.

Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 2023. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23182–23190. IEEE.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Yuan Gong, Jin Yu, and James R. Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 151–155. IEEE.

Google. 2025. Gemini 2.0 Flash: Model Card. Model card (technical report), Google. Model card published April 15, 2025.

Google DeepMind. 2025. Gemini 3 Pro: Model Card. Model card (technical report), Google DeepMind. Model card published November 2025.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. 2025. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 814–825. ACM.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, Leilei Sun, Dahua Lin, and Jiaqi Wang. 2025. Star-bench: Probing deep spatio-temporal reasoning as audio 4d intelligence. *CoRR*, abs/2510.24693.

Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pages 575–592. Springer.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. 2025. MECAT: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *CoRR*, abs/2507.23511.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Krishna Regmi and Mubarak Shah. 2019. Bridging the domain gap for ground-to-aerial image matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 470–479. IEEE.

Florian Schmid, Khaled Koutini, and Gerhard Widmer. 2023. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 544–560. Springer.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10090–10100.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 2020. Where am I looking at? joint location and orientation estimation by cross-view matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4063–4071. Computer Vision Foundation / IEEE.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4297–4316. Association for Computational Linguistics.

Chun Wang, Xiaoran Pan, Zihao Pan, Haofan Wang, and Yiren Song. 2025b. GRE suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. *CoRR*, abs/2505.18700.

Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet - photo geolocation with convolutional neural networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 37–55. Springer.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

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. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

A Appendix

A.1 Introduction of Aporee Platform

Aporee, launched in 2006, provides geo-tagged audios by enabling users to voluntarily upload environmental sound recordings with associated GPS coordinates, resulting in a global sound map. The platform covers a wide range of acoustic environments, including urban areas, rural regions, and natural landscapes, and has been widely used for artistic, educational, and research purposes. Through collaboration with the Aporee team, we obtained access to their audio data under the Apache license.

A.2 Interactive Platform Construction

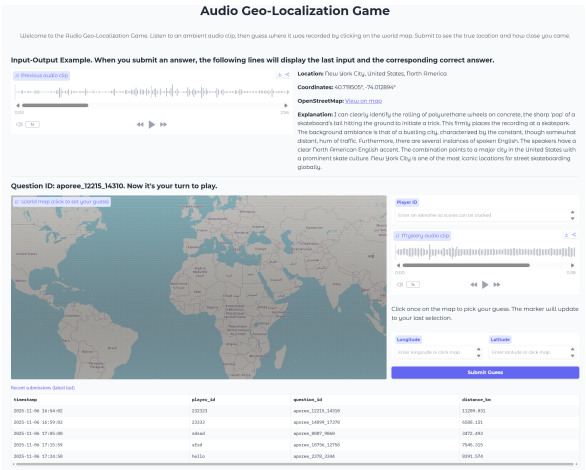

Figure 8: **The screenshot of our Interactive Platform**. Built by Hugging Face Gradio, our platform can be easily deployed to Hugging Face Space or locally.

To collect human performance data and encourage broader participation in audio geo-localization, we develop an interactive web-based platform and deploy it on HuggingFace. The interface presents users with an audio clip and an interactive map, and users are asked to infer the sounds location by clicking on the map. User predictions and anonymized identifiers are publicly logged on the HuggingFace Space, enabling analysis of human localization behavior and providing a reference baseline for comparison with model performance.

A.3 Formulation of the four acoustic filters

The four acoustic filters are defined as:

$$
\begin{aligned}
\mathrm{RMS} &= \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2}, \\
\mathrm{SF} &= \frac{\left(\prod_{k=1}^{K} P[k]\right)^{1/K}}{\frac{1}{K}\sum_{k=1}^{K} P[k]}, \\
\mathrm{CR} &= \frac{\#\{n \mid |x[n]| = x_{\max}\}}{N}, \\
\mathrm{AC} &= \sum_{f} \frac{\sum_{t=1}^{T-1} |E_{f,t+1} - E_{f,t}|}{\sum_{t=1}^{T} E_{f,t}},
\end{aligned}
\tag{2}
$$

where $x[n]$ denotes the audio amplitude at sample n, $P[k]$ the power spectral density at frequency bin k, and $E_{f,t}$ the energy of frequency band f at time frame t.

A.4 Definition of Evaluation Metrics

Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ denote the evaluation dataset, where x_i is an audio sample and $y_i = (c_i^{\text{cont}}, c_i^{\text{country}}, c_i^{\text{city}}, \ell_i)$ denotes its ground-truth continent, country, city, and geographic coordinates $\ell_i = (\phi_i, \lambda_i)$ (latitude and longitude). For a model prediction $\hat{y}_i = (\hat{c}_i^{\text{cont}}, \hat{c}_i^{\text{country}}, \hat{c}_i^{\text{city}}, \hat{\ell}_i)$, we define the following metrics.

A.4.1 Mean Distance Error (km)

We compute the great-circle distance between predicted and ground-truth coordinates using the Haversine formula:

$$
d_i = \mathrm{Haversine}(\ell_i, \hat{\ell} * i). \tag{3}
$$

If the model refuses to answer, we directly set d_i as 10000. The mean distance error is

$$
\mathrm{DistErr} = \frac{1}{|\mathcal{A}|}\sum * i \in \mathcal{A} d_i, \tag{4}
$$

where $\mathcal{A} \subseteq 1, \ldots, N$ denotes the set of all samples.

A.4.2 Hierarchical Accuracy

We report accuracy at three geographic levels:

$$
\begin{aligned}
\mathrm{Acc}_L &= \frac{1}{|\mathcal{A}|}\sum_{i \in \mathcal{A}} \mathbb{I}\left[\hat{c}_i^{L} = c_i^{L}\right], \\
L &\in \text{continent}, \text{country}, \text{city},
\end{aligned}
\tag{5}
$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Table 2: Further Comparison Results.

Model	q25↓	q50↓	q75↓	Animal Distance↓	Music Distance↓	Nature Distance↓	Things Distance↓
RANDOM	6253.81	10062.38	13639.9	10181.77	10667.35	10259.51	10515.23
Closed-Source Models							
Gemini 3 Pro	**55**	**480.19**	**1888.6**	**3143.16**	**3231.5**	**3669.51**	**3765.42**
Gemini 2.5 Pro	97.59	548.1	2848.37	4635.67	3476.44	4620.69	4541.46
Gemini 2.5 Flash	279.21	1213.16	6782.78	5962.53	3893.8	5918.76	5353.91
Gemini 2.5 Flash-Lite	664.15	2295.73	8275.64	5976.78	4092.25	6247.15	5911.77
Gemini 2.0 Flash Thinking	263.39	883.79	5544.65	4576.98	4889.07	4839.28	4808.03
Gemini 2.0 Flash	218.81	886.77	5119.63	4510.66	4319.07	4673.66	4462.81
Gemini 2.0 Flash-Lite	280.3	1002.83	5888.17	5936.17	3900.27	5883.09	5426.24
GPT-4o Audio Preview	415.21	1715.7	7723.63	7960.98	4849.44	7696.62	6613.44
Open-Source Models							
Qwen3-Omni	**633.98**	5627.17	**8821.64**	**5920.09**	5048.03	6064.08	6484.72
Qwen2.5-Omni	983.09	5952.77	10000	7723.77	5694.41	7871.21	7854.18
Phi-4-MM1	1797.82	6485.13	10000	6905.37	6066.34	6728.47	6640.15
Kimi-Audio	829.04	5945.86	10000	8274.65	6692.15	8148.38	7601.32
gemma-3n-E4B-it	1263.42	7020.47	9441.25	7203.56	6063.93	7160.14	6794.96
MiniCPM-o-2.6	1355.05	10000	10000	9517.41	7424.07	9345.83	9070.55
Mimo-Audio	893.83	**3450.54**	9208.35	7318.02	4552	7293.23	6091.1
Mimo-Audio-think	854.09	3627.95	9036.6	6046.2	**4419.81**	**6028.08**	**6001.72**

A.4.3 Thresholded Distance Accuracy ($< \tau$)

To capture coarse-to-fine localization performance, we compute thresholded accuracy at multiple distance thresholds $\tau \in 1, 10, 100, 1000$ km:

$$\text{Acc}_{<\tau} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbb{I}\big[d_i < \tau\big]. \tag{6}$$

These correspond to the columns denoted as < 1, < 10, < 100, and < 1000 in Table 2.

A.4.4 Reject Rate

Some models may abstain from answering. We define the reject rate as

$$\text{RejectRate} = 1 - \frac{|\mathcal{A}_a|}{N}, \tag{7}$$

which measures the fraction of samples for which the model does not return a prediction and $\mathcal{A}_a$ denotes the set of samples for which the model produces a valid prediction (i.e., not rejected).

A.4.5 Speech and Non-Speech Distance Error

To analyze reliance on linguistic clues, we partition the dataset into speech samples $\mathcal{D} *$ speech and non-speech samples $\mathcal{D}_{\text{non-speech}}$. We then compute mean distance error separately:

$$\begin{aligned} \text{DistErr}_{\text{speech}} &= \frac{1}{|\mathcal{A}_{\text{speech}}|} \sum_{i \in \mathcal{A}_{\text{speech}}} d_i, \\ \text{DistErr}_{\text{non-speech}} &= \frac{1}{|\mathcal{A}_{\text{non-speech}}|} \sum_{i \in \mathcal{A}_{\text{non-speech}}} d_i. \end{aligned} \tag{8}$$

A.5 Definition of Main Sound Labels

Nature refers to the Natural Sound label in Audioset. This is defined as: Sounds produced by natural sources in their normal soundscape, excluding animal and human sounds.

Animal refers to the Animal label in Audioset. This is defined as: All sound produced by the bodies and actions of nonhuman animals.

Music refers to the Music label in Audioset. This is defined as: Music is an art form and cultural activity whose medium is sound and silence. The common elements of music are pitch, rhythm, dynamics, and the sonic qualities of timbre and texture.

Things refers to the Sounds of Things label in Audioset. This is defined as: Set of sound classes referring to sounds that are immediately understood by listeners as arising from specific objects (rather than being heard more literally as "sounds").

A.6 Further Comparison Results

We report the 25th, 50th, and 75th percentile localization errors (denoted as q25, q50, and q75, respectively) for each model in Table 2. Additionally, we break down the errors across four subcategories: Animal, Music, Nature, and Sound of Things. Notably, Gemini 3 Pro achieves localization errors under 50 km at the 25th percentile across all categories, highlighting its remarkable

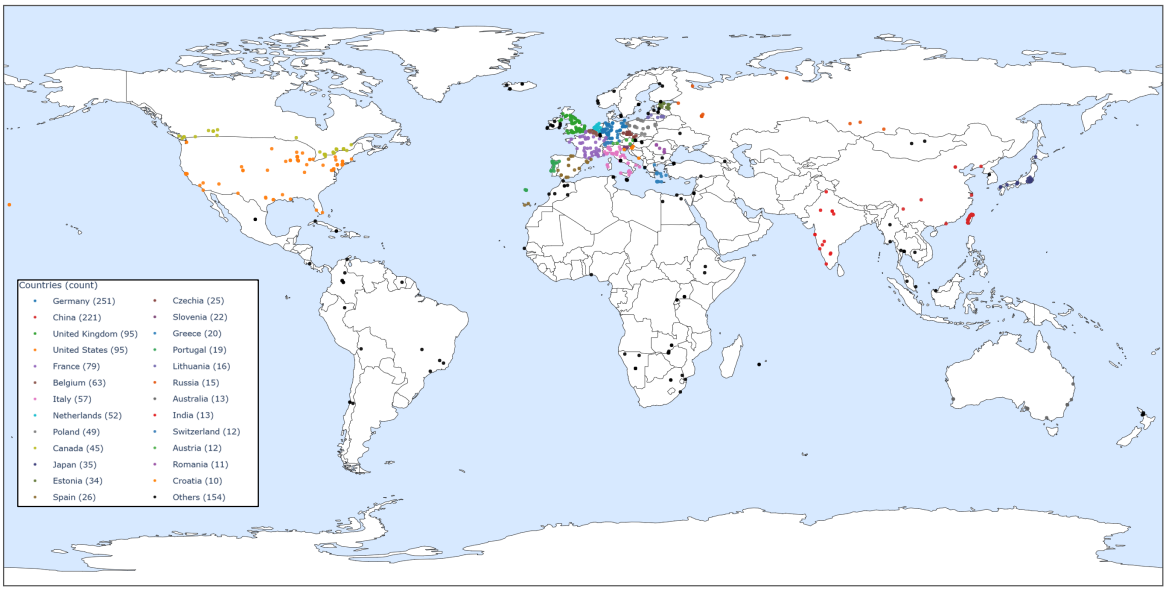

Figure 9: **The Global Distribution of AGL1K.** Locations of all 1,444 audios across 74 countries. Each marker denotes an image coordinate, and colors indicate country (top contributors listed; Others aggregates the remainder)

audio grounding capabilities. In contrast, the best-performing open-source model still exhibits an error of 634 km at the same percentile, underscoring the significant performance gap between proprietary and open-source models. The relatively lower errors in Music and Animal categories may be attributed to the presence of linguistic or distributional cues in musical content and animal vocalizations, whereas Nature and Sound of Things categories tend to lack such informative structure.

A.7 Further Experimental Settings

We accessed closed-source models via OpenRouter (https://openrouter.ai/models) and invoked Qwen3-Omni through Alibaba Clouds Bailian API (https://cn.aliyun.com/product/bailian). Please refer to the respective websites for detailed pricing information. Open-source models were deployed on a single NVIDIA RTX 4090 GPU, with each model requiring approximately 12 GPU hours for inference.

B Definition of Each Error Type

Bird Bias denotes errors arising from region-specific bird vocalizations.

Language Ambiguity refers to cases in which a language is spoken across multiple geographic regions, leading the model to select an incorrect location despite correctly identifying the language.

Over-Commitment captures situations in which the models intermediate reasoning supports multiple plausible hypotheses, but it nevertheless commits to a single overconfident prediction.

Over-Reasoning describes cases in which the model initially infers the correct answer but subsequently overrides it due to unnecessary or misguided additional reasoning. Label Misidentification corresponds to factual inference errors, most commonly incorrect language identification (e.g., classifying French as German).

Educated Guess refers to instances in which the model reports insufficient information and effectively responds at random.

Refusal covers cases in which the model declines to provide an answer.

B.1 Global Distribution of AGL1K

The Figure 9 shows the global distribution of AGL1K.

B.2 Consistency on three contribution-labeling LLM

To assess the reliability and instruction-following consistency of large language models in contribution evaluation, we measure cross-model agreement among Gemini, Claude, and Qwen using three complementary metrics (Figure 10). At the sample level, pairwise cosine similarity of contribution vectors ranges from **0.65 to 0.68**, indicating substantial directional agreement on which audio

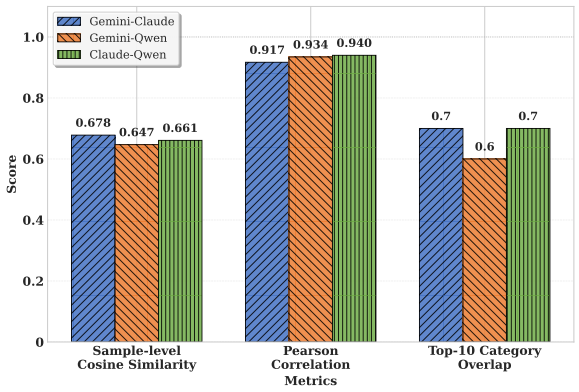

Figure 10: Model Similarity Comparison Across Different Metrics.

clues matter for individual examples. When aggregating contributions across categories, the Pearson correlation exceeds **0.91** for all model pairs, demonstrating highly consistent global attribution patterns. In addition, the Top-10 category overlap remains above **0.60**, showing that models agree on the majority of the most influential sound categories.

Together, these results indicate that contribution judgments are stable across different LLMs, supporting the robustness of our ensemble-based annotation strategy while allowing for minor model-specific preferences.

B.3 Continent-Distribution of Localizability

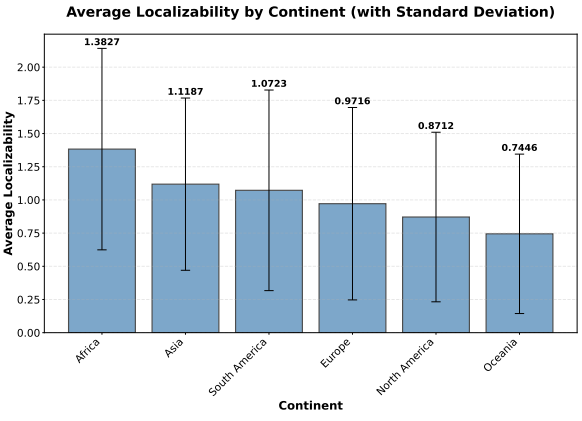

Figure 11: Continent-Distribution of Localizability.

We analyze the localizability of audio across different continents, as shown in Figure 11. We observe that Africa, Asia, and South America exhibit higher average localizability scores, whereas Europe, North America, and Oceania show lower values. One possible explanation lies in the linguistic and cultural distinctiveness of the former group: regions within Africa, Asia, and South America often feature more unique acoustic environments and languages, which may provide richer cues for geo-localization. In contrast, countries in Europe, North America, and Oceania often share English as a dominant language, potentially leading to less discriminative audio patterns and thus reduced localizability.

Moreover, the observed localizability rankings partially align with the actual model performance. For instance, Gemini 3 Pro achieves the highest prediction accuracy in Asia, followed by Europe, Africa, North America, South America, and Oceania. which correlates with the average localizability scores across continents. This suggests that localizability, as a metric, captures intrinsic properties of audio that influence model performance and may serve as a useful predictor of geo-localization difficulty across regions.

B.4 Human Annotator

This study involved human annotators at two key stages. First, during dataset refinement, after filtering about 3,000 samples using the localizability metric, a human annotator was tasked with selecting examples that were genuinely geo-locatable. The annotator, a PhD student in computer science, was instructed to consider the model's output, the audio content, and the ground-truth location to assess whether city-level inference was plausibly achievable.

Second, in the error analysis stage, another human annotator was responsible for labeling failure cases according to predefined error categories. For speech-related errors, the annotator consulted a PhD student in linguistics to determine whether language recognition was incorrect. Annotators were compensated at a rate of 100 RMB per hour and were informed that their annotations would be used in this study.

B.5 Prompts

B.5.1 Benchmark Prompt

Benchmark Prompt

You are an expert audio analyst with exceptional ability to identify geographical locations from environmental sounds. Listen carefully to the provided audio record-

ing and identify the geographical location where it was recorded.
You MUST provide your best guess regardless of uncertainty. Even if you feel unsure, you must still make an educated guess based on the audio characteristics, ambient sounds, languages spoken, cultural indicators, or any other audio clues you can detect.
Your response MUST be in the following JSON format (replace the values with your predictions):
{
"reason": "Detailed explanation of your reasoning process. Describe what audio clues you heard (e.g., language, accent, background sounds, environmental noise, cultural indicators, weather sounds, traffic patterns, animal sounds, etc.) and how they led you to this specific location prediction. Be specific about the evidence that supports your geographical conclusion."
"city": "Your predicted city name",
"country": "Your predicted country name",
"continent": "Your predicted continent name",
"longitude": predicted longitude value,
"latitude": predicted latitude value
}
Important requirements:

- 1. You MUST respond with a complete JSON object - no explanations outside the JSON, no refusals, no uncertainty statements
- 2. If you're unsure about specific coordinates, provide your best numerical estimate
- 3. All string values should be in English
- 4. Longitude should be between -180 and 180
- 5. Latitude should be between -90 and 90
- 6. The "reason" field must contain your detailed step-by-step analysis of the audio clues
- 7. Even if the audio seems generic, you must still provide a location guess with reasoning

Listen to the audio and provide your JSON response now:

B.5.2 Contribution Calculation Prompts

Contribution Calculation Prompts

You are analyzing an audio geolocalization reasoning text. Your task is to identify which sound categories from the provided list contributed to the location prediction, and assess their contribution level (0.0 to 1.0).
Available sound categories:
{categories}
Please analyze the reasoning text and identify which categories are mentioned or implied, and how much each category contributed to the location prediction.
Output your response as a JSON object where:
- Each key is a category name (exactly as listed above)
- Each value is a contribution score between 0.0 and 1.0, where:
* 0.0 = not mentioned/not relevant
* 0.1-0.3 = mentioned but minor contribution
* 0.4-0.6 = moderate contribution
* 0.7-0.9 = significant contribution
* 1.0 = primary/dominant contribution
Only include categories that have a contribution > 0.0. Output ONLY the JSON object, no additional text or explanation.
Example output format:
{
"Speech": 0.8,
"Music": 0.3,
"Animal": 0.2
}

B.6 Sample Output

In this section, we present the complete outputs of selected models for three representative cases. For each audio clip, we additionally generate a corresponding image using NanoBanana Pro to facilitate the interpretation and understanding of the audio.

B.6.1 Example 1-Gemini 3 Pro

Example1

{
reason": "The audio recording provides a distinct sonic fingerprint pointing to the Maghreb region of North Africa. The most prominent initial clue is the Adhan (Islamic call to prayer) performed in a style typical of the region, immediately placing the location in a Muslim-majority country. This is overlaid with the loud, persistent squawking of seagulls and significant wind noise, which strongly suggests a coastal city on a major body of water. The traffic noise includes the distinctive high-pitched whine of 2-stroke mopeds or scooters navigating narrow streets, a common soundscape in Moroccan medinas. In the second half of the recording, the sound shifts to powerful, rhythmic waves crashing against a shore, sounding more like the heavy swells of the Atlantic Ocean than the typically calmer Mediterranean. Linguistically, shouts of 'Yallah' (Arabic) and what appears to be French phrases are audible, reflecting the bilingual heritage of Morocco. The specific combination of intense wind, overwhelming seagull presence, the Adhan, and heavy Atlantic surf is the hallmark of Essaouira, Morocco, famously known as the 'Wind City of Africa' with its historic fortified port."
"city": "Essaouira",
"country": "Morocco",
"continent": "Africa",
"longitude": -9.7595,
"latitude": 31.5085
}

B.6.2 Example 1-GPT-4o Audio

Example 1

{
reason": "The audio recording features a clear Adhan, the Islamic call to prayer, which is typically heard in Muslim-majority countries. The background noises include traffic and urban sounds, suggesting a city environment. The presence of the Adhan combined with the urban noise leads to the conclusion that this might be a city in a country where Islam is predominant. Considering well-known cities, Istanbul in Turkey is a plausible guess due to its size, significance in the Muslim world, and blend of modernity and tradition. The city is often busy with traffic, as reflected in the audio, and the geographical location aligns with these audible cues."
"city": "Istanbul",
"country": "Turkey",
"continent": "Europe",
"longitude": 28.9784,
"latitude": 41.0082
}

B.6.3 Example 1-Qwen3-Omni

Example 1

{
reason": "The audio contains a mix of sounds that strongly suggest a specific geographical location. The most prominent sound is a loud, high-pitched, and sustained horn, characteristic of a Japanese train or subway announcement. This is followed by a rhythmic, metallic clatter of

wheels on tracks, typical of a train moving on rails. The background also features the sound of a large, powerful engine, possibly from a ferry or a large boat, with the sound of waves. The combination of these soundstrain/subway, ferry, and ocean wavespoints to a coastal city with a well-developed public transportation system. The language spoken in the background is Japanese. The combination of these elements, particularly the specific sound of the train horn and the coastal environment, is highly characteristic of Japan. The city of Kobe, Japan, is a major port city with a significant train network and is located on the coast of the Seto Inland Sea, which would produce the sounds of both a train and a ferry."
"city": "Kobe",
"country": "Japan",
"continent": "Asia",
"longitude": 135.198,
"latitude": 34.693
}

B.6.4 Example 1-Mimo-Audio

Example 1

{
reason": "The audio contains several distinct clues that point to a specific geographical location. The primary clue is the prominent sound of a train horn, which is repeated multiple times throughout the recording. This suggests the recording was made near a railway line. Additionally, there are background sounds of traffic, including cars and possibly buses, indicating an urban environment. The ambient noise level is high, with a mix of mechanical and human-made sounds, which is typical of a bustling city. While there are no clear spoken words or languages that could definitively identify the location, the combination of train sounds and urban traffic noise strongly suggests a major city with an active rail network. One possible candidate is Mumbai, India, which is known for its dense urban environment and extensive railway system. The sounds of trains and traffic are consistent with the city's infrastructure. However, without more specific audio cues, this remains an educated guess."
"city": "Mumbai",
"country": "India",
"continent": "Asia",
"longitude": 72.8,
"latitude": 19.0
}

B.6.5 Example 2-Gemini 3 Pro

Example 2

{
"reason": "The audio environment is a large, reverberant hall with the distinct sounds of rolling suitcases, closing doors, and murmuring crowds, indicating a busy train station. At 0:48, faint conversation in Swedish can be heard. The definitive geographical evidence begins at 1:08 with a standard electronic chime followed by a female automated station announcement in Swedish. The voice announces: 'Till tåg [number]... mot Uppsala... via Arlanda C, Knivsta... Var god stig ombord, spår 6' (To train... towards Uppsala... via Arlanda Central, Knivsta... Please board, track 6). This specific routetraveling northbound to the city of Uppsala with stops at Arlanda Airport (Arlanda C) and Knivstaoriginates from the main railway hub in the region. Therefore, this is recorded at Stockholm Central Station, specifically on the platform level.",
"city": "Stockholm",
"country": "Sweden",

"continent": "Europe",
"longitude": 18.0583,
"latitude": 59.3301
}

B.6.6 Example 2-Qwen3-Omni

Example 2

{
"reason": "The audio contains several key clues pointing to a specific location. The most definitive clue is the announcement made in Korean, which is a language spoken in Korea. The announcement is for a 'KTX' train, which is the high-speed rail service operated by Korail in South Korea. The announcement also mentions 'Seoul Station' (), which is a major railway hub in Seoul. The background sounds include the distinct rumbling and clatter of a train moving on tracks, the sound of doors closing, and the general ambiance of a train station. The combination of the Korean language, the mention of KTX and Seoul Station, and the train sounds strongly indicates that this recording was made at Seoul Station in South Korea.",
"city": "Seoul",
"country": "South Korea",
"continent": "Asia",
"longitude": 126.9781,
"latitude": 37.5326
}

B.6.7 Example 3-Gemini 3 Pro

Example 3

{
"reason": "The most prominent audio clue is the distinct, fluty, and melodious song of a Common Blackbird (Turdus merula). This bird is widespread across Europe and is famous for its adaptability to urban environments. The acoustic characteristics of the recording are equally telling: there is a noticeable reverberation or echo to the bird's song and the background noises, which strongly suggests the recording was taken in a 'Hinterhof'a large, enclosed inner courtyard formed by block-style apartment buildings. This architectural style is iconic to Central European cities, particularly Berlin. In the background, there is a steady 'city hum' of distant traffic without aggressive honking, and the mechanical sounds at the beginning (a heavy rumble and metallic squeal) resemble a tram (StraSSenbahn) or urban train passing nearby. The combination of the specific 'courtyard acoustics,' the pervasive blackbird song, and the sounds of European public transport infrastructure creates a sonic signature that is widely recognized as the soundscape of a residential neighborhood in Berlin, Germany.",
"city": "Berlin"
"country": "Germany",
"continent": "Europe",
"longitude": 13.405,
"latitude": 52.52
}

B.6.8 Example 3-GPT-4o Audio

Example 3

{
"reason": "The audio recording contains ambient nature sounds, with prominent bird calls and chirping that seems specific to a suburban or park area. At points, there are also distant traffic noises, indicating

a nearby roadway or urban environment. The type of birds and the overall soundscape suggest a temperate climate, possibly in the northern hemisphere. Given the generic nature sounds without distinctive markers like specific language or cultural audio indicators, a specific location prediction is difficult. However, based on the temperate nature, and commonality of such parks and environments, I am considering a well-known global city with expansive parks and temperate climate conditions. This leads me to a plausible guess of New York City, with Central Park being an exemplary location for such an environment. This choice is speculative and based on the assumption of a large city with notable green spaces and a presence of city and nature sounds.",
"city": "New York City"
"country": "United States",
"continent": "North America",
"longitude": -73.9654,
"latitude": 40.7829
}

B.6.9 Example 3-Qwen3-Omni

Example 3

{
"reason": "The audio contains a prominent and complex bird song, characteristic of a song thrush (Turdus philomelos), which is common in Europe. The background is dominated by the sounds of urban traffic, including cars and a distinct 'honk' of a vehicle horn. The combination of these specific bird calls and the nature of the traffic noise strongly suggests a location in Western Europe. The overall soundscape is typical of a city park or a green space within a European city. The specific combination of the song thrush's song and the urban environment points towards a location in the United Kingdom, where this bird is widespread and the traffic patterns are consistent with the audio. London is a major city in the UK with such environments.",
"city": "London"
"country": "United Kingdom",
"continent": "Europe",
"longitude": -0.1278,
"latitude": 51.5074
}