MACHINE HEARING: CAN AUDIO LLMS GENERALIZE BEYOND HUMAN AUDITORY PERCEPTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal LLMs have made remarkable progress in perceiving the physical world, while their perceptual scope remains largely human-centric, focusing on what humans can directly see or hear. This progress raises the prospect of perception beyond human limits, such as extending hearing to sounds that are inaudible to humans. More broadly, the physical world is full of waves, vibrations, and oscillations, only a small fraction of which are directly audible to humans. This invites the extension of multimodal LLMs toward a broader form of machine hearing, which remains largely underexplored. To investigate this direction, we introduce the **M**achine **H**earing **B**enchmark (**MHB**), comprising 6,128 instances across 13 tasks and seven domains. MHB spans highly heterogeneous input regimes, with sampling rates from approximately 10^{-5} to 10^{6} Hz, sequence lengths spanning approximately 10^{1} to 10^{6} samples, and both single- and multi-channel inputs. Evaluations of current open-source and proprietary ALLMs reveal limited zero-shot transfer to machine hearing. Supervised adaptation on machine-hearing data alone does not close this gap, suggesting that the challenge extends beyond data exposure to a systematic mismatch between existing audio front ends and machine-hearing inputs. To address this mismatch, we introduce **MachEar**, a lightweight encoder extension that complements the pretrained Audio LLM for machine-hearing inputs. The results show that MachEar equips a pretrained Audio LLM with stronger machine-hearing capabilities while largely preserving its existing audio understanding performance.

1 INTRODUCTION

Advanced Audio Large Language Models (ALLMs) (Chu et al., 2024; Tang et al., 2024; Xu et al., 2025) have demonstrated strong capabilities in understanding speech, music, environmental sounds, and other forms of audio. However, their perceptual scope remains largely human-centric, focusing primarily on what humans themselves can hear. As multimodal intelligence continues to expand, a natural question is whether this boundary should remain tied to human auditory perception. The physical world is full of waves, vibrations, and oscillations, only a small fraction of which are directly audible to humans (Fig. 1). Earthquakes rumble through the ground, machines vibrate as they operate, radio and radar waves carry oscillatory patterns, biological systems pulse and resonate, and even astronomical phenomena vary in characteristic rhythms. Extending ALLMs to interpret such phenomena therefore opens a broader form of machine hearing, raising a fundamental question: can current ALLMs extend beyond human-audible audio to understand what humans cannot hear?

To systematically investigate this question, we introduce the **M**achine **H**earing **B**enchmark (**MHB**), comprising 6,128 instances across 13 tasks and seven domains: Astronomy, Earth Science, Healthcare, Electromagnetic, Underwater Acoustics, Bioacoustics, and Manufacturing. MHB covers a wide range of machine-hearing phenomena, spanning sampling rates from approximately 10^{-5} to 10^{6} Hz, sequence lengths spanning approximately 10^{1} to 10^{6} samples, and both single- and multi-channel inputs. This diversity allows us to examine whether current ALLMs can extend beyond conventional human-audible audio across fundamentally different input regimes.

We benchmark seven current ALLMs, including both open-source and proprietary models, on MHB to examine how far their existing hearing capabilities extend beyond conventional audio. The results reveal a striking gap. Even taking the best-performing Audio LLM for each task, performance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

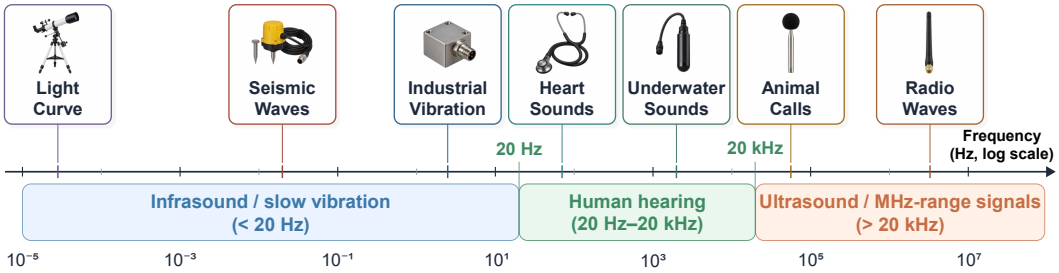


Figure 1: Conceptual illustration of waves, vibrations, and oscillations across characteristic frequency scales. Human-audible sound constitutes only a small subset of the broader signal space considered in machine hearing.

remains below random guessing on 9 of the 13 tasks. Moreover, the difficulty is not limited to incorrect predictions: existing models can fail to recognize or process machine-hearing inputs, with such failures particularly concentrated in regimes with extreme sampling rates or unusually short sequences. We further ask whether these limitations can be addressed simply by exposing ALLMs to machine-hearing data. Supervised adaptation substantially improves performance, but does not consistently eliminate the gap. The remaining failures instead point to systematic limitations in how ALLMs process inputs far outside conventional audio regimes.

Motivated by these findings, we introduce **MachEar**, a lightweight encoder extension designed to complement the pretrained audio encoder in low- and high-frequency regimes and on extremely short sequences. Equipping Qwen2.5-Omni with MachEar improves its weighted Macro-F1 on MHB from 67.02% to 72.30%, while largely preserving its ASR and audio captioning performance.

Summary of Contributions. 1) We construct the Machine Hearing Benchmark (**MHB**), which extends the evaluation of ALLMs beyond conventional human-audible audio to a broad range of machine-hearing phenomena across diverse domains and input regimes. **2)** We systematically benchmark current open-source and proprietary Audio LLMs on MHB, revealing limited transfer beyond human-audible audio and showing that supervised adaptation alone is insufficient to overcome these limitations. **3)** We introduce **MachEar**, a lightweight encoder extension that equips pretrained Audio LLMs with stronger machine-hearing capabilities while largely preserving their existing audio understanding performance.

2 Related Work

2.1 Audio Benchmarks

Existing audio benchmarks, such as SUPERB (Yang et al., 2021), SUPERB-SG (Tsai et al., 2022), HEAR (Turian et al., 2022), HARES (Wang et al., 2022), AIR-Bench (Yang et al., 2024b) and AudioBench (Wang et al., 2025), primarily evaluate speech, music, environmental sounds, and other forms of human-audible audio. Within this scope, these benchmarks evaluate models on downstream tasks such as automatic speech recognition (ASR), audio captioning, sound event recognition, and speech emotion recognition. As a result, current audio evaluation remains largely human-centric, with limited coverage of phenomena and input regimes beyond conventional human hearing. Whether Audio LLMs can extend beyond this boundary toward a broader form of machine hearing therefore remains largely underexplored. MHB complements existing audio benchmarks by extending evaluation beyond human-audible audio to diverse machine-hearing phenomena across a wide range of domains and input regimes, within a unified LLM-compatible framework.

2.2 Audio LLMs

Audio LLMs, such as Qwen2-Audio (Chu et al., 2024), SALMONN (Tang et al., 2024), Audio Flamingo 3 (Ghosh et al., 2025), and Qwen2.5-Omni (Xu et al., 2025), integrate audio perception with large language models to support audio understanding, and achieve strong performance across a range of audio tasks, including automatic speech recognition (Yang et al., 2024a), speech emotion understanding (Wang et al., 2024), and audio captioning (Chen et al., 2025). However, their audio front ends are predominantly built on encoders such as Whisper (Radford et al., 2023), WavLM (Chen et al., 2022), and EAT (Chen et al., 2024), which were developed for conventional

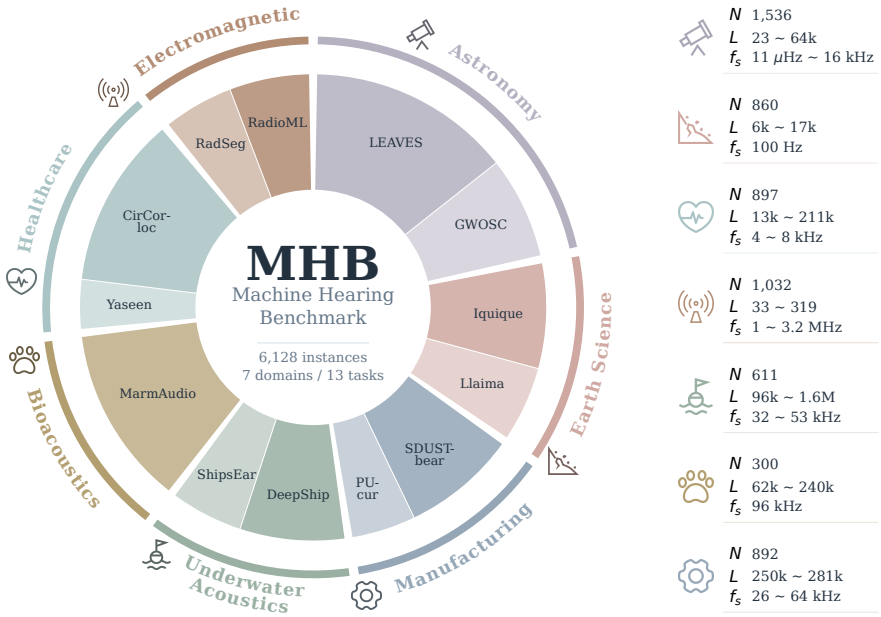


Figure 2: Overview of the Machine Hearing Benchmark (MHB), comprising 6,128 instances across 13 tasks and seven domains. For each domain, N, L, and f_s denote the number of instances, sequence length, and sampling rate, respectively.

human-audible audio, including speech, music, and environmental sounds. Consequently, the hearing capabilities of current ALLMs remain largely centered on the human-audible regime, while their ability to handle the broader and highly heterogeneous input regimes encountered in machine hearing has received limited study. Our work investigates this extension beyond human auditory perception, and further explores how existing ALLMs can be augmented for machine hearing while retaining their conventional audio capabilities.

3 Machine Hearing Benchmark

The Machine Hearing Benchmark (MHB) comprises 6,128 instances across 13 tasks from seven domains: Astronomy, Earth Science, Healthcare, Electromagnetic, Underwater Acoustics, Bioacoustics and Manufacturing. As summarized in Fig. 2, MHB spans sampling rates from approximately 10^{-5} to 10^{6} Hz, sequence lengths spanning approximately 10^{1} to 10^{6} samples, and single- and multi-channel signals. Across domains, these signals differ by several orders of magnitude in sampling rate and sequence length and are acquired through markedly different sensing devices, highlighting the heterogeneity of MHB.

3.1 Data Curation and Task Formulation

MHB is curated from 13 publicly available datasets and domain-specific data sources, supplemented by numerical simulations where appropriate. For each task, we retain the original input measurements and task semantics, with target labels derived from the corresponding annotations or metadata. We select tasks that can be naturally formulated as machine-hearing problems while preserving their original scientific or application objectives. Details about the data sources can be found in Appendix A.2.

To support unified evaluation with ALLMs and enable consistent evaluation across heterogeneous domains, each MHB instance consists of an input, a task-specific textual prompt, and a predefined label set. We preserve the original task definitions and label semantics while standardizing the prompt structure and response format across tasks, allowing different machine-hearing problems to be evaluated under a unified Audio LLM protocol. Representative prompt-response formulations are provided in Appendix E.

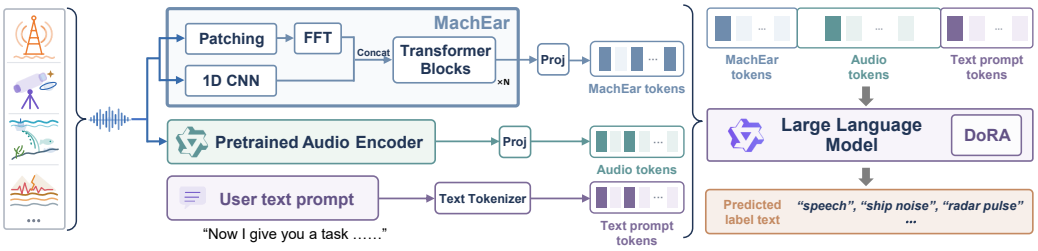


Figure 3: Overview of MachEar integrated with a pretrained audio pathway. Given an input signal, MachEar and the pretrained audio encoder process the same input in parallel to produce complementary representations. The resulting MachEar tokens, audio tokens, and text prompt tokens are concatenated and jointly fed into the LLM for prediction.

3.2 Domains and Tasks

The 13 tasks in MHB span seven domains and cover a broad range of waves, vibrations, oscillations, and acoustic phenomena. These include seismic waveforms, physiological recordings, bioacoustic and underwater acoustic signals, mechanical vibration and motor-current signals, electromagnetic and radar signals, gravitational-wave signals, and astronomical light curves.

In **Astronomy**, MHB includes gravitational-wave event detection and variable-star classification from stellar light curves. In **Earth Science**, MHB includes earthquake focal-depth classification from seismic waves and seismic event-type classification in volcanic environments. In **Healthcare**, MHB includes auscultation-location identification and cardiac disease classification from phonocardiogram recordings. In **Electromagnetic**, MHB includes modulation scheme classification from radio-frequency waveforms and waveform-type classification from radar pulse segments. In **Underwater Acoustics**, MHB includes vessel and sound-source classification from underwater acoustic recordings. In **Bioacoustics**, MHB includes marmoset vocalization-type classification. In **Manufacturing**, MHB includes bearing fault diagnosis from mechanical vibration and motor-current signals. Detailed task-level statistics are provided in Appendix A.1.

4 MachEar

The broad variation in machine-hearing inputs calls for an extension that can accommodate diverse input regimes while retaining the pretrained audio capabilities of existing Audio LLMs. We therefore introduce MachEar, a lightweight encoder extension that broadens the effective coverage of the pretrained audio encoder toward low- and high-frequency regimes and extremely short sequences. As illustrated in Fig. 3, MachEar processes the same input in parallel with the pretrained audio encoder, and the resulting MachEar tokens, audio tokens, and text prompt tokens are jointly fed into the LLM for prediction.

MachEar combines spectral and temporal features to handle diverse machine-hearing inputs. Specifically, FFT and 1D CNN features are extracted from the input through two parallel pathways and concatenated, followed by Transformer blocks and a projection layer that encode the full input sequence into a fixed number Q of compact MachEar tokens. The same architecture is applied across all machine-hearing domains without domain-specific modifications. More architectural and implementation details of MachEar are provided in Appendix C.1.

5 Experimental Setup

5.1 Baseline Models

We evaluate seven ALLMs, including three open-source and four proprietary models. To contextualize the performance of ALLMs, we also evaluate five foundation models pretrained on specialized signals from related domains. These models provide a complementary reference for examining how pretraining in related domains beyond speech transfers to machine-hearing tasks.

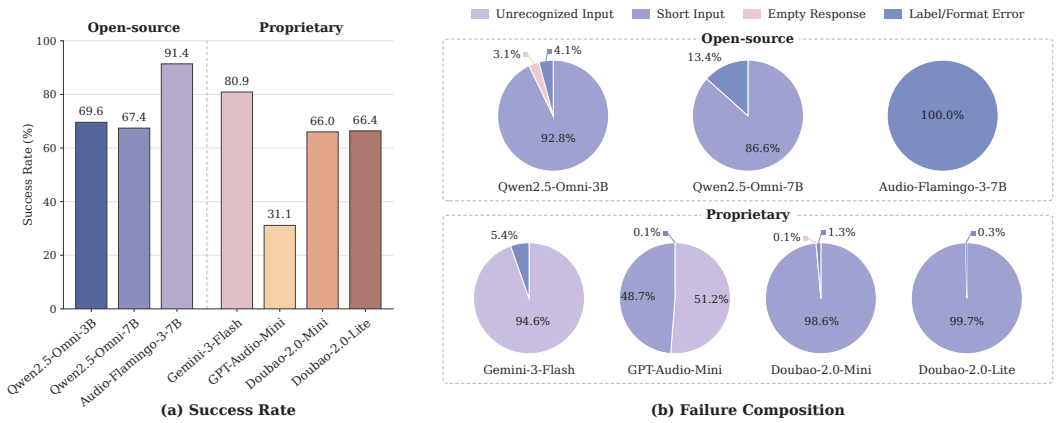


Figure 4: Response validity of Audio LLMs on MHB. (a) Overall success rate (SR), defined as the percentage of valid responses. (b) Composition of invalid responses across four failure categories.

ALLMs. We evaluate three open-source models, Qwen2.5-Omni-3B[1], Qwen2.5-Omni-7B[2] (Xu et al., 2025), and Audio-Flamingo-3-7B[3] (Ghosh et al., 2025), together with four proprietary models: Gemini-3-Flash[4], GPT-Audio-Mini[5], Doubao-2.0-Mini[6], and Doubao-2.0-Lite[6]. All Audio LLMs are evaluated zero-shot on MHB using a unified prompting protocol.

Domain-specific Foundation Models. We additionally evaluate five pretrained foundation models from related domains: Time-MoE[7] (Shi et al., 2025) for general time series, SpectrumFM[8] (Zhou et al., 2026) for radio spectrum, SeisLM[9] (Liu et al., 2024) for seismic waveforms, FISHER[10] (Fan et al., 2025) for industrial signals, and SPEAR[11] (Yang et al., 2025) for speech and audio. For each MHB task, each foundation model is kept frozen, and 5-nearest-neighbor classification is performed in its representation space using the corresponding training set as the reference set. Detailed model descriptions and experimental setup are provided in Appendix B.

5.2 Evaluation Metrics

We evaluate model performance using Macro-F1 and Success Rate (SR). Macro-F1 is used as the primary classification metric to account for class imbalance across tasks. Success Rate measures the percentage of evaluated samples for which an ALLM produces a valid and parseable response within the predefined task-specific label set. Representative invalid responses and the corresponding failure categories are provided in Appendix D.

6 Results and Discussion

6.1 Benchmarking ALLMs on MHB

Response Validity and Failure Modes. We first examine whether current ALLMs can reliably process MHB inputs and produce valid responses. The overall success rate of each model and the breakdown of its typical failure modes are reported in Fig. 4. Although success rates vary across models, invalid responses remain non-negligible across the board, indicating that reliable input handling and task-compliant generation are themselves part of the challenge on MHB. A notable fraction of invalid responses arises from short-input failures, which are concentrated in tasks with atypical

[1] https://huggingface.co/Qwen/Qwen2.5-Omni-3B
[2] https://huggingface.co/Qwen/Qwen2.5-Omni-7B
[3] https://huggingface.co/nvidia/audio-flamingo-3
[4] https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview
[5] https://platform.openai.com/docs/models/gpt-audio-mini
[6] https://www.volcengine.com/product/doubao/
[7] https://huggingface.co/Maple728/TimeMoE-200M
[8] https://github.com/ChunyuLiu188/SpectrumFM
[9] https://github.com/liutianlin0121/seisLM
[10] https://huggingface.co/jiangab/FISHER-small-0723
[11] https://huggingface.co/marcoyang/spear-large-speech-audio

Table 1: Zero-shot Macro-F1 (%) of Audio LLMs on MHB, with invalid responses treated as incorrect predictions. Random Guess is included as a reference.

Access	Model	Astronomy		Earth Science		Healthcare	
		GWOSC	LEAVES	Iquique	Llaima	Yaseen	CirCor-loc
–	Random Guess	49.95	8.94	40.66	22.64	19.79	24.92
Open-source	Qwen2.5-Omni-3B	34.36	1.87	18.57	5.91	7.87	23.81
	Qwen2.5-Omni-7B	38.11	0.00	25.46	3.75	12.60	22.06
	Audio-Flamingo-3-7B	49.83	0.22	34.77	3.75	7.74	17.83
Proprietary	Gemini-3-Flash	35.62	0.00	31.88	23.85	16.05	15.68
	GPT-Audio-Mini	6.82	0.00	15.70	6.69	16.96	18.95
	Doubao-2.0-Mini	32.73	0.00	31.02	14.89	7.74	11.91
	Doubao-2.0-Lite	33.10	0.00	31.51	16.47	7.74	21.53

Access	Model	Electromagnetic		Underwater Acoustics		Bioacoustics	Manufacturing	
		RadSeg	RadioML	ShipsEar	DeepShip	Marm Audio	SDUST-bear	PU-cur
–	Random Guess	19.96	9.05	6.46	24.88	16.60	9.94	32.54
Open-source	Qwen2.5-Omni-3B	6.87	0.00	0.54	8.69	5.35	1.76	18.18
	Qwen2.5-Omni-7B	8.25	0.00	1.11	8.76	9.66	3.44	20.81
	Audio-Flamingo-3-7B	6.64	1.52	0.83	0.00	6.09	4.84	27.16
Proprietary	Gemini-3-Flash	10.60	6.59	8.92	16.26	14.52	11.46	27.51
	GPT-Audio-Mini	0.00	0.00	5.57	5.76	0.00	4.50	25.11
	Doubao-2.0-Mini	0.00	0.00	15.98	12.82	17.85	3.64	14.15
	Doubao-2.0-Lite	0.00	0.00	8.91	16.80	19.51	2.21	11.26

signal regimes, including the ultra-low-sampling-rate LEAVES task and the MHz-range RadSeg and RadioML tasks. Beyond this shared pattern, failure modes remain strongly model dependent: Gemini-3-Flash is dominated by unrecognized input often treating the provided input as absent and requesting the time-series, I/Q samples, or acoustic features, while Audio Flamingo 3 primarily produces label or format violations that arise from responses outside the predefined label set, such as "Tank" instead of "Tanker". GPT-Audio-Mini shows the lowest success rate, with its invalid responses almost evenly split between unrecognized-input and short-input failures. A detailed task-level breakdown of success rates and representative invalid responses are provided in Appendix D.

Performance. Table 1 reports task-level Macro-F1 results on MHB. Macro-F1 is computed over all MHB samples, with invalid responses treated as incorrect predictions. Even when taking the best ALLM result for each task, performance remains below random guessing on 9 of the 13 tasks. Particularly low scores are observed on tasks that combine extreme sampling rates with short sequences, including the ultra-low-rate LEAVES task and the MHz-range RadSeg and RadioML tasks. Notably, these are also among the tasks with frequent input-handling failures in the preceding success-rate analysis. Even when the models produce valid responses, their predictions remain largely uninformative. At the same time, relative strengths vary across models and domains, with no single ALLM consistently dominating MHB. Overall, these results show that current ALLMs struggle to generalize beyond human auditory perception, with performance remaining sensitive to both signal characteristics and task domain.

6.2 Benchmarking Domain Foundation Models on MHB

Beyond ALLMs, we further evaluate foundation models pretrained on related domains to provide a complementary view of MHB tasks. Table 2 summarizes their performance. First, results show that domain-specific pretraining can provide a strong advantage when the pretraining and target tasks are well aligned, as expected. For example, FISHER achieves the best performance on both Manufacturing tasks. However, this advantage is not universal: SeisLM leads on Iquique but not Llaima, and SpectrumFM does not rank first on either Electromagnetic task. Conversely, strong transfer is not restricted to nominally matched domains. SPEAR, an audio foundation model, remains competitive across many tasks outside conventional audio, while the general time-series model Time-MoE leads on LEAVES and RadSeg. Overall, these results suggest that domain specialization alone does not guarantee consistently strong transfer across heterogeneous MHB tasks.

Table 2: Macro-F1 (%) of domain-specific foundation models on MHB using frozen representations and 5-NN classification. Best and second-best results are shown in bold and underlined, respectively.

Model	Astronomy		Earth Science		Healthcare	
	GWOSC	LEAVES	Iquique	Llaima	Yaseen	CirCor-loc
Time-MoE	59.18	**67.07**	46.82	<u>94.11</u>	93.09	27.87
SpectrumFM	52.86	<u>61.70</u>	48.85	79.48	76.05	26.52
SeisLM	48.75	58.42	**60.42**	93.33	93.08	29.12
FISHER	<u>68.75</u>	0.00	0.00	0.00	**94.20**	<u>38.57</u>
SPEAR	**78.91**	44.82	<u>52.28</u>	**95.51**	<u>93.88</u>	**51.07**

Model	Electromagnetic		Underwater Acoustics		Bioacoustics	Manufacturing	
	RadSeg	RadioML	ShipsEar	DeepShip	MarmAudio	SDUST-bear	PU-cur
Time-MoE	**88.12**	2.82	7.62	20.96	15.61	22.80	36.08
SpectrumFM	82.92	<u>44.19</u>	13.68	32.10	22.45	32.26	21.51
SeisLM	<u>87.22</u>	**45.33**	20.08	<u>49.98</u>	29.11	55.09	33.42
FISHER	0.00	0.00	<u>23.17</u>	48.69	**92.57**	**73.93**	**53.69**
SPEAR	67.42	31.02	**25.96**	**58.69**	<u>77.55</u>	<u>60.57</u>	<u>53.34</u>

6.3 Can Supervised Adaptation Close the Machine-Hearing Gap?

This section investigates whether the poor performance of current ALLMs on MHB can be explained simply by a mismatch between their training data and machine-hearing domains. We therefore examine whether direct supervised adaptation on machine-hearing data is sufficient to close the gap. We take Qwen2.5-Omni-7B as the backbone, finetune the LLM with DoRA, and compare two configurations: `Whisper_frz`, where the pretrained Whisper encoder is frozen, and `Whisper`, where it is jointly finetuned on MHB training data. As shown in Fig. 5, supervised adaptation enables the model to make meaningful predictions on tasks where zero-shot ALLMs largely fail. Further finetuning the Whisper encoder improves performance on most tasks, showing that the pretrained audio front end itself needs to adapt to machine-hearing inputs. However, this adaptation does not resolve all limitations. Tasks such as LEAVES, RadSeg, and RadioML remain particularly challenging despite finetuning, and are characterized by sampling rates and sequence lengths far outside conventional audio regimes. Their persistent difficulty after task-specific adaptation suggests that the bottleneck is not merely data-domain mismatch, but architectural constraints of conventional audio front ends in handling the diverse input regimes of machine hearing. This motivates an architectural extension of the audio front end, which we pursue with MachEar in the next section.

6.4 Augmenting ALLMs with MachEar

To address the architectural limitations identified above, we introduce MachEar, a lightweight encoder extension for augmenting ALLMs with stronger machine-hearing capability. Our analysis of domain foundation models in Sec. 6.2 provides useful architectural clues. Models that remain effective on challenging MHB tasks often adopt explicit frequency-domain processing or convolutional modelling of temporal inputs, suggesting inductive biases that are better suited to input regimes beyond conventional audio. Guided by these clues, MachEar combines FFT-based spectral modelling with a 1D CNN to complement the pretrained Whisper encoder. Rather than replacing the pretrained audio encoder, MachEar operates alongside it, broadening the effective coverage of the audio front end toward low- and high-frequency regimes and extremely short sequences.

We evaluate two MachEar-based configurations using the same Qwen2.5-Omni-7B backbone: `MachEar`, which directly replaces the pretrained Whisper encoder with MachEar; and `Whisper + MachEar`, which adds MachEar to Qwen backbone and jointly finetunes both encoders. As shown in Fig. 5, `MachEar` alone outperforms the finetuned `Whisper` configuration on a substantial subset of MHB tasks, especially LEAVES and RadSeg that have shorter length. This indicates that its frequency-domain and convolutional design captures signal structure that is not fully modelled by audio front ends developed for conventional audio tasks. Accordingly, augmenting Qwen2.5-Omni-7B with MachEar yields more consistent gains across tasks, raising the overall weighted Macro-F1 from 67.02% to 72.30%. Together, these results indicate that MachEar provides complementary machine-hearing capacity beyond direct adaptation of the pretrained audio encoder.

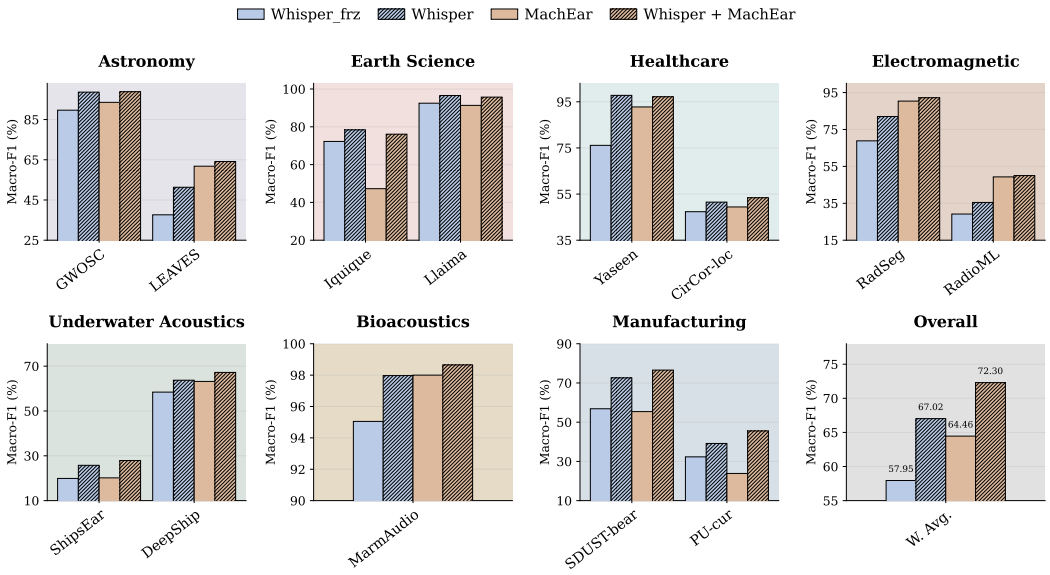


Figure 5: Task-level Macro-F1 (%) of four adaptation configurations on MHB. All use DoRA for LLM adaptation: `Whisper_frz` freezes Whisper; `Whisper` finetunes Whisper; `MachEar` replaces Whisper; and `Whisper + MachEar` jointly finetunes both encoders. *W. Avg.* denotes the test-set-size-weighted average.

6.5 Ablation Study and Analysis of MachEar

This section provides a more detailed analysis of where MachEar is most effective and how its main components contribute to its performance. We group the 13 MHB tasks according to sequence length and sampling rate, resulting in four input regimes:

- G1 – short high-rate inputs (signal length $<$ 1k, sampling rate $>$ 96kHz);
- G2 – short-to-medium low-rate inputs (signal length 1k–10k, sampling rate $<$ 2kHz);
- G3 – long low-rate inputs (signal length $>$ 10k, sampling rate $<$ 2kHz);
- G4 – long inputs within conventional audio sampling rates (signal length $>$ 10k, sampling rate 2–96kHz).

Based on this grouping, we compare Qwen2.5-Omni-7B with and without MachEar, and further analyze the contributions of MachEar's FFT and 1D CNN components.

Effectiveness across Input Regimes. We first examine how the benefit of augmenting Qwen2.5-Omni-7B with MachEar varies across the four input regimes by comparing `Whisper` with `Whisper + MachEar`. As shown in Fig. 6(a), adding MachEar provides the largest gains in G1 and G2, whereas its advantage becomes smaller for the longer-sequence G3 and G4 regimes. This pattern suggests that MachEar's complementary benefit is most pronounced for shorter sequences with sampling rates outside the conventional audio range, consistent with its design goal of compensating for the limitations of the original audio front end and extending the effective input range of ALLM.

Component Ablation. We next examine the contributions of the two main components of MachEar across the four signal regimes. To isolate the contribution of MachEar itself, we keep the pretrained Whisper encoder frozen across all ablation variants. As shown in Fig. 6(b), the 1D CNN contributes most clearly in the shorter-sequence regimes, with its removal reducing weighted Macro-F1 from 70.03% to 66.18% in G1 and from 78.04% to 63.76% in G2. In contrast, removing the FFT produces broader degradation across G2–G4, with drops of 7.38, 7.37, and 5.44 points, while having little effect on G1. These results reveal complementary roles of the two components of MachEar: the 1D CNN is particularly important for handling shorter sequences, whereas the FFT contributes more consistently across signals spanning a broader range of sampling rates.

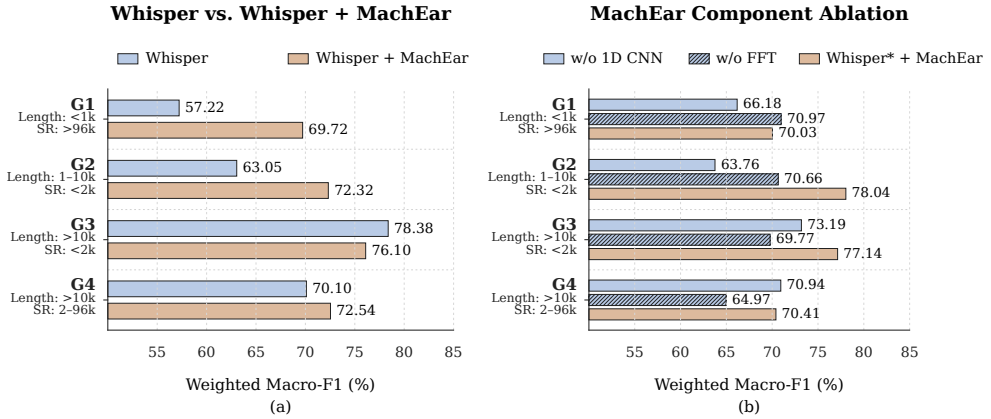


Figure 6: MachEar analysis across four input regimes. (a) `Whisper` vs. `Whisper + MachEar`. (b) Component ablations, where * denotes a frozen pretrained Whisper encoder. Results are test-set-size-weighted Macro-F1 (%) within each group.

6.6 Preservation of Conventional Audio Capabilities

We next evaluate whether extending the model toward machine hearing affects its performance on conventional audio tasks. We compare the machine-hearing-adapted `Whisper + MachEar` with the original Qwen2.5-Omni-7B on automatic speech recognition (ASR) and audio captioning tasks. As shown in Fig. 7, while degradation is expected after adapting the model to machine hearing, the observed drop on conventional audio tasks is surprisingly modest. The gap can be further narrowed by incorporating ASR and audio-captioning data during machine-hearing adaptation, as shown by `Whisper + MachEar (+audio)` in Fig. 7. A relatively small amount of conventional audio-task data, accounting for less than one third of the overall adaptation data, is sufficient to largely recover the degradation and achieves performance comparable to, or even better than, the original Qwen2.5-Omni-7B across the evaluated metrics. Overall, these results suggest that MachEar can extend ALLMs toward machine hearing without substantially disrupting their original audio capabilities.

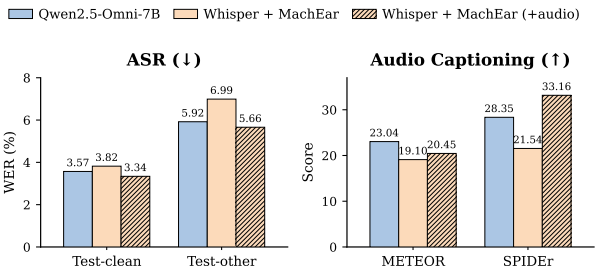


Figure 7: Performance on conventional audio tasks under different adaptation configurations. Lower WER is better, while higher METEOR and SPIDEr are better.

7 Conclusion

In this work, we investigate whether Audio LLMs can extend beyond human auditory perception toward a broader form of machine hearing. We introduce the Machine Hearing Benchmark (MHB), a diverse benchmark spanning 7 domains, 13 tasks and highly heterogeneous input regimes, to systematically evaluate this capability. Benchmarking a broad set of current open-source and proprietary ALLMs reveals that current ALLMs struggle to generalize beyond human hearing. More importantly, supervised adaptation with task-specific machine-hearing data improves performance but does not fully close the gap, showing that the limitation cannot be explained by data-domain mismatch alone. Instead, the persistent failures point to architectural constraints in conventional audio front ends when handling the diverse input regimes encountered in broader machine hearing. Motivated by this finding, we introduce MachEar, a lightweight encoder extension that can be readily attached to existing ALLMs. Results show that MachEar complements the pretrained audio encoder, particularly for extremely short sequences and inputs at low- and high-sampling-rate extremes. Importantly, these gains can be achieved while largely preserving conventional audio performance. Overall, our findings expose the current limits of ALLMs while demonstrating their emerging potential to extend toward machine hearing. More broadly, they point toward multimodal LLMs whose perceptual scope is no longer bounded by human senses, enabling them to perceive and understand aspects of the physical world inaccessible to humans.

AI USE STATEMENT

LLMs were employed solely to polish the writing and correct grammatical errors.

REFERENCES

ByteDance Seed. Doubao-Seed-2.0 model family. https://developer.volcengine.com/articles/7610285824933445675, 2026. Accessed: 2026-08-24.

João Paulo Canário, Rodrigo Fernandes de Mello, Millaray Curilem, Fernando Huenupan, and Ricardo Araujo Rios. Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification. *Data in Brief*, 30:105627, 2020. doi: 10.1016/j.dib.2020.105627.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. EAT: Self-supervised pre-training with efficient audio Transformer. *arXiv preprint arXiv:2401.03497*, 2024.

Wenxi Chen, Ziyang Ma, Xiquan Li, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Kai Yu, and Xie Chen. SLAM-AAC: Enhancing audio captioning with paraphrasing augmentation and CLAP-refine through LLMs. In *Proc. ICASSP*, Hyderabad, 2025.

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. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. doi: 10.48550/arXiv.2407.10759.

Pingyi Fan, Anbai Jiang, Shuwei Zhang, Zhiqiang Lv, Bing Han, Xinhu Zheng, Wenrui Liang, Junjie Li, Wei-Qiang Zhang, Yanmin Qian, Xie Chen, Cheng Lu, and Jia Liu. FISHER: A foundation model for multi-modal industrial signal comprehensive representation. *arXiv preprint arXiv:2507.16696*, 2025. doi: 10.48550/arXiv.2507.16696.

Ya Fei, Ce Yu, Kun Li, Xiaodian Chen, Yajie Zhang, Chenzhou Cui, Jian Xiao, Yunfei Xu, and Yihan Tao. LEAVES: An expandable light-curve data set for automatic classification of variable stars. *The Astrophysical Journal Supplement Series*, 275(1):10, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Proc. NeurIPS*, San Diego, 2025.

Google. Gemini 3 Flash Preview model documentation. https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview, 2026. Accessed: 2026-08-24.

Baokun Han, Zujie Yang, Zongzhen Zhang, Huaiqian Bao, Jinrui Wang, Zongling Liu, and Shunming Li. A novel rolling bearing fault diagnosis method based on generalized nonlinear spectral sparsity. *Measurement*, 198:111131, 2022.

Zi Huang, Akila Pemasiri, Simon Denman, Clinton Fookes, and Terrence Martin. Multi-stage learning for radar pulse activity segmentation. In *Proc. ICASSP*, Seoul, 2024.

Muhammad Irfan, Zheng Jiangbin, Shahid Ali, Muhammad Iqbal, Zafar Masood, and Umar Hamid. DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Systems with Applications*, 183:115270, 2021. doi: 10.1016/j.eswa.2021.115270.

Charly Lamothe, Manon Obliger-Debouche, Paul Best, Régis Trapeau, Sabrina Ravel, Thierry Artières, Ricard Marxer, and Pascal Belin. A large annotated dataset of vocalizations by common marmosets. *Scientific Data*, 12(1):782, 2025.

Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *Proc. PHME*, Bilbao, 2016.

Tianlin Liu, Jannes Münchmeyer, Laura Laurenti, Chris Marone, Maarten V de Hoop, and Ivan Dokmanić. SeisLM: A foundation model for seismic waveforms. *arXiv preprint arXiv:2410.15765*, 2024. doi: 10.48550/arXiv.2410.15765.

Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, et al. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2524–2535, 2022. doi: 10.1109/JBHI.2021.3137048.

OpenAI. GPT Audio Mini model documentation. `https://platform.openai.com/docs/models/gpt-audio-mini`, 2026. Accessed: 2026-08-24.

Timothy J O'shea and Nathan West. Radio machine learning dataset generation with gnu radio. In *Proc. GRCon*, Boulder, 2016.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, Honolulu, 2023.

David Santos-Domínguez, Soledad Torres-Guijarro, Antonio Cardenal-López, and Antonio Pena-Gimenez. ShipsEar: An underwater vessel noise database. *Applied Acoustics*, 113:64–69, 2016. doi: 10.1016/j.apacoust.2016.06.008.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *Proc. ICLR*, Singapore, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *Proc. ICLR*, Vienna, 2024.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy T Liu, Cheng-I Lai, Jiatong Shi, et al. SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities. In *Proc. ACL*, Dublin, 2022.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. HEAR: Holistic evaluation of audio representations. In *Proc. NeurIPS 2021 Competitions and Demonstrations Track*, Conference held virtually, 2022.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *Proc. NAACL*, Albuquerque, 2025.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. BLSP-Emo: Towards empathetic large speech-language models. In *Proc. EMNLP*, Miami, 2024.

Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. Towards learning universal audio representations. In *Proc. ICASSP*, Singapore, 2022.

Jack Woollam, Andreas Rietbrock, Angel Bueno, and Silvio De Angelis. Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network. *Seismological Research Letters*, 90(2A):491–502, 2019.

Jack Woollam, Jannes Münchmeyer, Frederik Tilmann, Andreas Rietbrock, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Carlo Giunchi, Florian Haslinger, Dario Jozinović, et al. SeisBench—a toolbox for machine learning in seismology. *Seismological Research Letters*, 93(3):1695–1709, 2022.

Wen Wu, Ziyang Zhang, Liwei Liu, Xuenan Xu, Jimin Zhuang, Ke Fan, Qitan Lv, Junlin Liu, Chen Zhang, Zheqi Yuan, et al. SciTS: Scientific time series understanding and generation with LLMs. In *Proc. ICLR*, Rio de Janeiro, 2026.

Yuan Xie, Ji Xu, Jiawei Ren, and Junfeng Li. Adversarial multi-task underwater acoustic target recognition: Toward robustness against various influential factors. *The Journal of the Acoustical Society of America*, 156(1):299–313, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. doi: 10.48550/arXiv.2503.20215.

Guanrou Yang, Ziyang Ma, Fan Yu, Zhifu Gao, Shiliang Zhang, and Xie Chen. MaLa-ASR: Multimedia-assisted LLM-based ASR. In *Proc. INTERSPEECH*, Kos Island, 2024a.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proc. ACL*, Bangkok, 2024b.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. SUPERB: Speech processing universal PERformance benchmark. In *Proc. INTERSPEECH*, Brno, 2021.

Xiaoyu Yang, Yifan Yang, Zengrui Jin, Ziyun Cui, Wen Wu, Baoxiang Li, Chao Zhang, and Phil Woodland. SPEAR: A unified SSL framework for learning speech and audio representations. *arXiv preprint arXiv:2510.25955*, 2025. doi: 10.48550/arXiv.2510.25955.

Yaseen, Gui-Young Son, and Soonil Kwon. Classification of heart sound signal using multiple features. *Applied Sciences*, 8(12):2344, 2018.

Fuhui Zhou, Chunyu Liu, Hao Zhang, Wei Wu, Qihui Wu, Tony QS Quek, and Chan-Byoung Chae. SpectrumFM: A foundation model for intelligent spectrum management. *IEEE Journal on Selected Areas in Communications*, 44:4471–4488, 2026. doi: 10.1109/JSAC.2025.3644783. URL https://doi.org/10.1109/JSAC.2025.3644783.

Table 3: Task-level statistics of MHB, including the number of instances, input sequence length, sampling rate, number of channels, and data source for each task.

Domain	Name	# Instances	Sample Length Range	Sampling Rate(Hz)	Channel Range	Data Source
Astronomy	GWOSC	512	6.4e4	1.6e4	1	SciTS (Wu et al., 2026)
	LEAVES	1024	23∼1.3e3	1.1e-5	1	LEAVES (Fei et al., 2024)
Earth Science	Iquique	502	1.4e4∼1.7e4	1.0e2	3	Iquique (Woollam et al., 2019)
	Llaima	358	6.0e3	1.0e2	1	Llaima (Canário et al., 2020)
Healthcare	Yaseen	200	1.3e4∼2.7e4	8.0e3	1	Yaseen (Yaseen et al., 2018)
	CirCor-loc	697	2.1e4∼2.1e5	4.0e3	1	CirCor (Oliveira et al., 2022)
Electromagnetic	RadSeg	482	33∼3.2e2	3.2e6	2	RadSeg (Huang et al., 2024)
	RadioML	550	1.3e2	1.0e6	2	RadioML 2016.10A (O'shea & West, 2016)
Underwater Acoustics	ShipsEar	84	1.6e6	5.3e4	1	ShipsEar (Santos-Domínguez et al., 2016)
	DeepShip	527	9.6e4	3.2e4	1	DeepShip (Irfan et al., 2021)
Bioacoustics	MarmAudio	300	6.2e4∼2.4e5	9.6e4	1	MarmAudio (Lamothe et al., 2025)
Manufacturing	SDUST-bear	572	2.6e5	2.6e4	1	SDUST (Han et al., 2022; Fan et al., 2025)
	PU-cur	320	2.5e5∼2.8e5	6.4e4	1	PU (Lessmeier et al., 2016)
Overall	-	6,128	23∼1.6e6	1.1e-5∼3.2e6	1∼3	-

A More Details of MHB

A.1 Task Description

Table 3 presents task-level statistics, including sequence-length and sampling-rate ranges and the corresponding data sources. We give a detailed description of each task below.

All tasks in MHB are formulated as classification problems: each model receives a sensor-recorded signal together with a textual instruction and generates the corresponding class label in text form.

Astronomy. Astronomy studies celestial objects and physical phenomena in the Universe. Two tasks are involved:

- **GWOSC:** This task detects the presence of gravitational-wave events in time-series signals. It is formulated as a binary classification, with `normal` indicating the absence of an event and `anomaly` indicating its presence.
- **LEAVES:** This task identifies the variability category of a celestial object based on its brightness variations over time. It distinguishes seven categories: Cepheid, Delta Scuti, Eclipsing Binary, Long-Period Variable, Non-variable, Rotational Variable, and RR Lyrae.

Earth Science. Earth science studies the structure of Earth and its physical processes. Two tasks are involved:

- **Iquique:** This task classifies the focal depth of an earthquake source based on seismic recordings. It distinguishes three categories: shallow (0–70 km), intermediate (70–300 km) and deep (300–700 km).
- **Llaima:** This task identifies seismic event types in a volcanic environment. It distinguishes four categories: Volcano-tectonic, Tremor, Tectonic, and Long-period.

Healthcare. Healthcare focuses on assessing and maintaining human health. Two tasks are involved:

- **Yaseen:** This task classifies heart sound patterns based on phonocardiogram (PCG) recordings. It distinguishes five categories: Aortic Stenosis, Mitral Regurgitation, Mitral Stenosis, Mitral Valve Prolapse, and Normal.
- **CirCor-loc:** This task identifies the auscultation location from a PCG recording. It distinguishes four locations: aortic valve, tricuspid valve, pulmonary valve, and mitral valve.

Electromagnetic. Electromagnetic signal analysis studies electromagnetic waveforms and their applications in sensing and communication. Two tasks are involved:

- **RadSeg:** This task identifies the waveform type of a pulsed radar signal segment. It distinguishes five categories: Barker codes, Polyphase Barker codes, Frank codes, Linear Frequency-Modulated (LFM) pulses, and Coherent Unmodulated Pulse Trains.
- **RadioML:** This task identifies the modulation type of radio signals from their I/Q waveforms. It distinguishes 11 analog and digital modulation schemes, including PSK, QAM, FSK, PAM, AM, and FM variants.

Underwater Acoustics. Underwater acoustics studies sound and acoustic phenomena in underwater environments. Two tasks are involved:

- **ShipsEar:** This task identifies the vessel or sound-source type from underwater acoustic recordings. It distinguishes 12 categories, including different types of fishing, service, recreational, and commercial vessels, as well as natural ambient noise.
- **DeepShip:** This task identifies the type of vessel from underwater acoustic recordings. It distinguishes four categories: Cargo, Passengership, Tanker, and Tug.

Bioacoustics. Bioacoustics studies biological sounds and their use in understanding and identifying living organisms. One task is involved:

- **MarmAudio:** This task identifies the vocalization type of marmoset calls. It distinguishes six categories: Twitter, Tsik, Phee, Trill, Infant cry, and Seep.

Manufacturing. Manufacturing involves monitoring industrial equipment and diagnosing machine conditions to ensure reliable operation. Two tasks are involved:

- **SDUST-bear:** This task performs bearing fault diagnosis based on vibration signals. It distinguishes 10 operating conditions, including the normal condition and multiple inner-race, outer-race, and rolling-element fault settings.
- **PU-cur:** This task performs bearing fault diagnosis based on motor current signals. It distinguishes three categories: healthy, inner-race fault (IR), and outer-race fault (OR).

A.2 Dataset Description

This section provides description of the datasets contained in MHB.

GWOSC (Wu et al., 2026) is a synthetic gravitational-wave dataset designed for gravitational-wave event detection. It is constructed by injecting gravitational-wave templates generated through numerical relativity simulations into real detector strain data obtained from the Gravitational Wave Open Science Center (GWOSC), thereby preserving realistic detector noise. The noise pool consists of 4-second segments sampled at 16 kHz and extracted from high-quality detector intervals. Waveform templates spanning typical binary black hole mass ranges are scaled to different matched-filter signal-to-noise ratios and superimposed on these noise segments.

LEAVES (Fei et al., 2024) is a large-scale light-curve dataset designed for the automatic classification of variable stars. It is constructed by merging open data from three major astronomical surveys: the All-Sky Automated Survey for SuperNovae (ASAS-SN), Gaia, and the Zwicky Transient Facility (ZTF). The dataset comprises 977,953 light curves from variable stars and 134,592 from nonvariable sources. All variable stars are annotated with seven superclasses: Eclipsing binaries, Rotational variables, RR Lyrae, Cepheids, Long-period variables, Delta Scuti, and Nonvariable.

Iquique (Woollam et al., 2019) is a seismic waveform dataset containing earthquake recordings from the aftershock sequence following the 2014 M_w 8.1 Iquique earthquake in northern Chile. The dataset contains 13,400 examples of picked seismic arrivals, all of which correspond to earthquake events. The recordings are collected from three-component seismic stations and sampled at 100 Hz. We obtain the dataset through SeisBench (Woollam et al., 2022).

Llaima (Canário et al., 2020) is a volcanic seismic dataset designed for the classification of seismic events. It contains 3,592 signals recorded at the LAV station of Llaima Volcano in Chile between 2010 and 2016. The signals correspond to the vertical component of seismic recordings sampled at 100 Hz and were processed to retain the representative segment of each event. Each signal is

annotated by specialists from the Observatorio Vulcanológico de los Andes Sur (OVDAS) into one of four categories: Volcano-Tectonic (VT), Long Period (LP), Tremor (TR), and Tectonic (TC).

Yaseen (Yaseen et al., 2018) is a phonocardiogram (PCG) dataset designed for heart sound classification. It contains 1,000 recordings collected from books and online sources, with all signals standardized to a sampling rate of 8 kHz. The dataset is evenly divided into five categories with 200 recordings per class: Normal, Aortic Stenosis, Mitral Regurgitation, Mitral Stenosis, and Mitral Valve Prolapse. These categories cover normal heart sounds and four common valvular heart diseases.

CirCor (Oliveira et al., 2022) is a large-scale pediatric phonocardiogram (PCG) dataset containing heart sound recordings collected from multiple auscultation locations. It comprises over 5,000 recordings from 1,568 subjects, with recordings acquired at the main cardiac auscultation sites, including the aortic, pulmonary, tricuspid, and mitral valve areas. The original dataset provides detailed annotations for heart murmurs and auscultation locations. In MHB, we use only the auscultation location information and formulate a four-class classification task over the aortic, tricuspid, pulmonary, and mitral locations.

RadSeg (Radar Segmentation Dataset) (Huang et al., 2024) is a synthetic radar dataset originally designed for sample-wise segmentation of interleaved radar pulse activities. It contains five radar signal types with signal-to-noise ratios ranging from -20 to 20 dB at a resolution of 0.5 dB, where each sequence consists of 32,768 complex baseband IQ samples acquired at 3.2 MHz. In MHB, we use the sample-wise annotations to extract signal segments containing a single radar class, thereby reformulating the original segmentation task as radar segment classification.

RadioML 2016.10A (O'shea & West, 2016) is a synthetic radio signal dataset designed for automatic modulation classification. RadioML 2016.10A contains 220,000 I/Q signal samples spanning 11 modulation types, including eight digital modulations and three analog modulations. Each sample consists of 128 complex baseband I/Q observations, and the signal-to-noise ratio ranges from -20 to 18 dB in increments of 2 dB. The 11 modulation classes are 8PSK, AM-DSB, AM-SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, QPSK, and WBFM.

ShipsEar (Santos-Domínguez et al., 2016) is a real-world underwater acoustic dataset designed for vessel sound recognition, containing recordings of 11 vessel types together with natural ambient noise. We follow the 12-class train–test split introduced by Xie et al. (Xie et al., 2024)[12], which partitions the original recordings at the file level to reduce overlap between training and test data. Following this split and the corresponding segmentation protocol, we obtain 625 signal segments, including 541 training samples and 84 test samples, covering all 12 sound-source categories.

DeepShip (Irfan et al., 2021) is a large-scale real-world underwater acoustic dataset designed for vessel classification. It contains 47 hours and 4 minutes of recordings from 265 vessels across four classes: Cargo, Passengership, Tanker, and Tug, collected under diverse sea states and noise conditions. We follow a publicly released unofficial train–test split[13], which partitions the data at the recording level to avoid overlap between training and test segments. Following this split and our segmentation procedure, we obtain 56,255 signal segments, including 39,371 training samples and 16,884 test samples.

MarmAudio (Lamothe et al., 2025) is a large-scale dataset of common marmoset vocalizations designed for call-type classification. It contains more than 800,000 audio segments totaling approximately 253 hours, recorded at a sampling rate of 96 kHz. Among them, approximately 215,000 calls are annotated with six vocalization types: Infant cry, Phee, Seep, Trill, Tsik, and Twitter.

SDUST (Han et al., 2022) is a rolling-bearing vibration dataset collected on a dedicated fault-diagnosis test bench at Shandong University of Science and Technology. It contains vibration signals recorded under normal operation and multiple inner-race, outer-race, and rolling-element fault conditions, with a sampling rate of 25.6 kHz. In MHB, we use the processed SDUST bearing subset released with FISHER/RMIS (Fan et al., 2025), which defines a 10-class fault-diagnosis task consisting of the normal condition (NC) and inner-race (IF), outer-race (OF), and rolling-element (RF) faults at three fault sizes: 0.2, 0.4, and 0.6.

[12] https://github.com/xy980523/ShipsEar-An-Unofficial-Train-Test-Split

[13] https://github.com/ZhuPengsen/Method-for-Splitting-the-DeepShip-Dataset

PU (Lessmeier et al., 2016) is a bearing condition-monitoring dataset from Paderborn University containing synchronously recorded motor current and vibration signals under diverse bearing conditions. The original dataset comprises measurements from 32 bearing experiments, including healthy bearings and bearings with artificially induced or naturally developed damage, recorded under four operating conditions. In MHB, we use the motor-current subset processed in FISHER/RMIS (Fan et al., 2025) and formulate a three-class bearing fault diagnosis task consisting of healthy, inner-race fault (IR), and outer-race fault (OR).

B EVALUATED MODELS

This section provides corresponding model families, citations, and access points.

B.1 AUDIO LLMS

Qwen2.5-Omni (Xu et al., 2025) is an end-to-end multimodal model supporting text, image, audio, and video inputs, with text and speech generation capabilities. We evaluate the 3B and 7B variants on MHB using the checkpoints `Qwen/Qwen2.5-Omni-3B` and `Qwen/Qwen2.5-Omni-7B`. *Access:* `https://huggingface.co/Qwen/Qwen2.5-Omni-3B` and `https://huggingface.co/Qwen/Qwen2.5-Omni-7B`.

Audio Flamingo 3 (Ghosh et al., 2025) is a fully open 7B large audio-language model designed for audio understanding and reasoning across speech, sound, and music. We evaluate Audio Flamingo 3 on MHB using the official checkpoint `nvidia/audio-flamingo-3`. *Access:* `https://github.com/NVIDIA/audio-flamingo` and `https://huggingface.co/nvidia/audio-flamingo-3`.

Gemini 3 Flash (Google, 2026) is a multimodal model from Google supporting text, image, video, audio, and PDF inputs with text generation. We evaluate Gemini 3 Flash on MHB through the Gemini API using the model identifier `gemini-3-flash-preview`. *Access:* `https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview`.

GPT Audio Mini (OpenAI, 2026) is a cost-efficient audio-capable model from OpenAI supporting both text and audio inputs and outputs. We evaluate GPT Audio Mini on MHB through the OpenAI API using the model identifier `gpt-audio-mini`. *Access:* `https://developers.openai.com/api/docs/models/gpt-audio-mini`.

Doubao-Seed-2.0 (ByteDance Seed, 2026) is a family of general-purpose multimodal models from ByteDance, offering variants with different performance, latency, and cost trade-offs. We evaluate the Mini and Lite variants on MHB through the Volcano Engine API, using the model identifiers `doubao-seed-2-0-mini-260215` and `doubao-seed-2-0-lite-260215`, respectively. *Access:* `https://www.volcengine.com/product/doubao/` and `https://www.volcengine.com/docs/82379/1795150`.

B.2 DOMAIN-SPECIFIC FOUNDATION MODELS

Time-MoE (Shi et al., 2025) is a time-series foundation model designed for general-purpose forecasting across diverse temporal data. We evaluate the 200M-parameter variant on MHB using the official checkpoint `Maple728/TimeMoE-200M`. *Access:* `https://huggingface.co/Maple728/TimeMoE-200M`.

SpectrumFM (Zhou et al., 2026) is a spectrum foundation model designed for intelligent spectrum management and pretrained with self-supervised objectives including masked reconstruction and next-slot signal prediction. We evaluate SpectrumFM on MHB using the official pretrained checkpoint released for RML2018. *Access:* `https://github.com/ChunyuLiu188/SpectrumFM`.

SeisLM (Liu et al., 2024) is a foundation model for seismic waveforms pretrained on the large-scale SeisBench waveform collection. We evaluate the official pretrained SeisLM base model on MHB. *Access:* `https://github.com/liutianlin0121/seisLM`.

FISHER (Fan et al., 2025) is a foundation model for unified representation of heterogeneous industrial signals, including sound, vibration, and voltage signals. We evaluate the 21.4M-parameter small variant on MHB using the official checkpoint `jiangab/FISHER-small-0723`. *Access:* `https://huggingface.co/jiangab/FISHER-small-0723`.

SPEAR (Yang et al., 2025) is a self-supervised foundation model designed to learn unified representations across speech and general audio. We evaluate the 327M-parameter Large speech–audio variant on MHB using the official checkpoint `marcoyang/spear-large-speech-audio`. *Access:* `https://huggingface.co/marcoyang/spear-large-speech-audio`.

C Implementation Details of MachEar

C.1 MachEar Architecture and Token Fusion

Given an input signal $\mathbf{x}$, MachEar and the pretrained Whisper encoder process it in parallel. MachEar consists of two parallel pathways: an FFT-based spectral pathway and a dilated 1D CNN temporal pathway. Both pathways are aligned to Q token positions, with $Q = 64$ in all experiments. Their representations, $\mathbf{F}_{\mathrm{FFT}}$ and $\mathbf{F}_{\mathrm{CNN}}$, are concatenated along the feature dimension, projected to a shared hidden space, and processed by Transformer blocks to produce a compact sequence of MachEar tokens:

$$\mathbf{H}_{\mathrm{M}} = \mathrm{Proj}_{\mathrm{out}}\left(\mathrm{Transformer}\left(\mathrm{Proj}_{\mathrm{in}}\left[\mathbf{F}_{\mathrm{FFT}}; \mathbf{F}_{\mathrm{CNN}}\right]\right)\right). \tag{1}$$

For the FFT pathway, a signal of length L is divided into Q non-overlapping patches with patch length $P = \lfloor L/Q \rfloor$, and an FFT representation is computed for each patch. The 1D CNN instead operates directly on the full waveform, and its temporal features are adaptively pooled to the same Q token positions. For inputs shorter than Q, a global representation is used and invalid trailing positions are masked.

In parallel, the original Whisper encoder produces audio tokens $\mathbf{H}_{\mathrm{W}}$, while the textual prompt is embedded as $\mathbf{H}_{\mathrm{T}}$. The three sequences are concatenated and jointly fed into the LLM:

$$\mathbf{H}_{\mathrm{input}} = \mathrm{Concat}_{\mathrm{seq}}\left(\mathbf{H}_{\mathrm{M}}, \mathbf{H}_{\mathrm{W}}, \mathbf{H}_{\mathrm{T}}\right). \tag{2}$$

The same MachEar architecture and tokenization scheme are used for all machine-hearing tasks without domain-specific modifications.

C.2 Preprocessing and Finetuning Details

Input preprocessing. Multi-channel input signals are flattened into a one-dimensional sequence before encoding. All input signals are truncated to at most 240,000 samples. To better exploit the pretrained representations of Whisper, signals sampled between 2 and 96 kHz are resampled to 16 kHz before entering the Whisper pathway, while those outside this range retain their original sampling rates to avoid excessive changes in sequence length. In contrast, MachEar processes the truncated waveform at its original sampling rate.

Pathway adaptation and integration. We consider four configurations corresponding to Fig. 5. `Whisper_frz` uses the pretrained Whisper encoder while keeping it frozen, whereas `Whisper` jointly finetunes the Whisper encoder with the LLM. `MachEar` directly replaces the pretrained Whisper encoder with MachEar, while `Whisper + MachEar` adds MachEar alongside the pretrained audio pathway.

D Response Validity and Failure Analysis

In this section, we present a detailed task-level breakdown of success rates and representative invalid responses.

Table 4: Task-level success rates (%) of Audio LLMs on the 13 MHB tasks. A response is considered successful only if it matches the predefined task-specific label set and required output format.

Access	Model	Astronomy		Earth Science		Healthcare	
		GWOSC	LEAVES	Iquique	Llaima	Yaseen	CirCor-loc
Open-source	Qwen2.5-Omni-3B	100.00	7.13	100.00	100.00	100.00	80.92
	Qwen2.5-Omni-7B	100.00	7.13	73.71	100.00	91.50	83.79
	Audio-Flamingo-3-7B	100.00	100.00	100.00	100.00	100.00	100.00
Proprietary	Gemini-3-Flash	100.00	5.47	100.00	96.37	100.00	100.00
	GPT-Audio-Mini	4.88	0.00	40.24	33.52	100.00	90.82
	Doubao-2.0-Mini	99.80	0.39	100.00	100.00	100.00	95.98
	Doubao-2.0-Lite	100.00	0.39	98.80	100.00	100.00	99.86

Access	Model	Electromagnetic		Underwater Acoustics		Bioacoustics	Manufacturing	
		RadSeg	RadioML	ShipsEar	DeepShip	Marm Audio	SDUST-bear	PU-cur
Open-source	Qwen2.5-Omni-3B	52.70	0.00	100.00	100.00	99.67	100.00	100.00
	Qwen2.5-Omni-7B	52.07	0.00	100.00	100.00	99.33	100.00	100.00
	Audio-Flamingo-3-7B	100.00	100.00	100.00	0.00	100.00	100.00	100.00
Proprietary	Gemini-3-Flash	86.93	80.00	100.00	100.00	94.33	100.00	100.00
	GPT-Audio-Mini	0.00	0.00	86.90	15.75	1.67	75.70	41.88
	Doubao-2.0-Mini	0.00	0.00	100.00	99.81	100.00	100.00	100.00
	Doubao-2.0-Lite	0.00	0.00	100.00	100.00	100.00	100.00	100.00

Task-level success rates. Table 4 reports the success rate of each Audio LLM on the 13 MHB tasks considered in our adaptation experiments. A response is considered valid only when it conforms to the predefined task-specific label set and output format. The overall success rates reported in Fig. 4 are computed over all 6,128 samples rather than by averaging task-level rates. Task-level success rates vary substantially across input regimes. In particular, LEAVES, RadSeg, and RadioML exhibit frequent input-handling failures for several models, whereas many tasks within more conventional audio regimes achieve near-perfect response validity.

Failure categories and representative cases. We categorize invalid responses into four types. **Unrecognized Input** denotes cases where the model behaves as if the provided signal were absent and requests the input again. **Short Input** denotes failures caused by minimum-duration or encoder-length constraints. **Empty Response** denotes cases with no usable textual output. **Label/Format Error** denotes responses that do not match the predefined label set or required output format. These categories affect response validity only; Macro-F1 is still computed over all MHB samples, with invalid responses treated as incorrect predictions.

Representative examples of each failure type are shown in Table 5. Outputs are reproduced verbatim or shortened only for presentation, without changing the failure semantics.

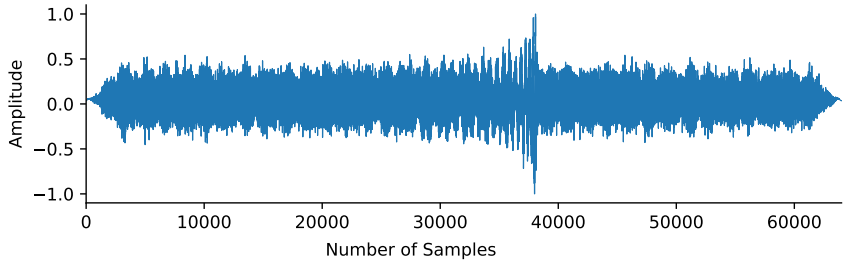
E CASE EXAMPLES

In this section, we present representative case examples from MHB. For visualization, the corresponding machine-hearing signals are plotted alongside the task instructions and model responses.

Table 5: Representative examples of the four invalid-response categories. Each example is selected from a different model and task.

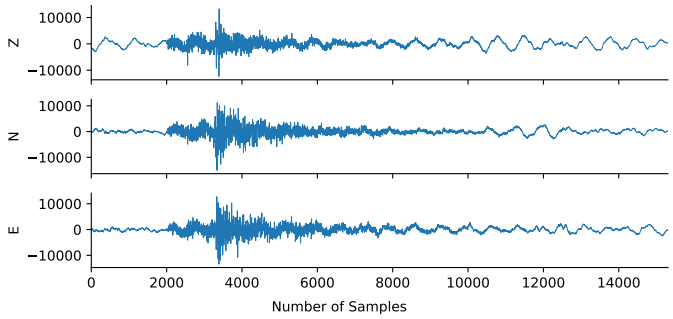
Failure Type	Model	Task	Representative Output / Error	Why Invalid
Unrecognized Input	Gemini-3-Flash	RadioML	"Please provide the in-phase (I) and quadrature (Q) data components ..."	Requests the I/Q data again although the signal has already been provided.
Short Input	Qwen2.5-Omni-3B	LEAVES	"Calculated output size ... 0. Output size is too small."	The input is too short for the audio encoder.
Empty Response	Doubao-2.0-Mini	GWOSC	*[empty output]*	No usable textual response is returned.
Label/Format Error	Audio-Flamingo-3-7B	DeepShip	GT: `Tanker`; Response: `Tank`	`Tank` does not exactly match the predefined label `Tanker`.

Astronomy: gravitational wave events



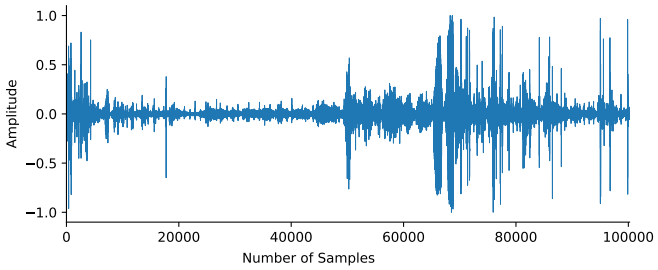
Question:
Please analyze the provided time-series data and determine whether a gravitational wave event has occurred. Classify the result as 'normal' (no gravitational wave event detected) or 'anomaly' (a gravitational wave event is detected).
Answer:
The classification result is: anomaly.

Earth Science: earthquake source depth



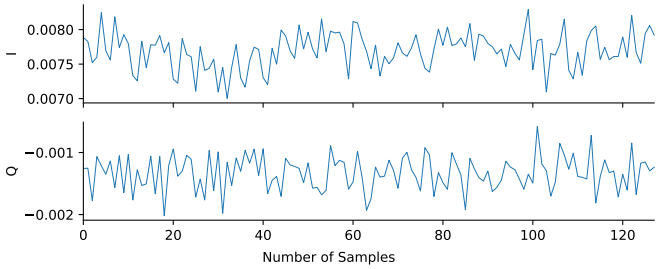
Question:
Now I give you a focal depth classification task. Classify the focal depth of the earthquake source into one of the following categories: shallow (0-70 km), intermediate (70-300 km), or deep (300-700 km). Analyze the recording and determine which category it belongs to.
Answer:
The focal depth of this seismic event is classified as shallow.

Healthcare: auscultation location from a PCG recording



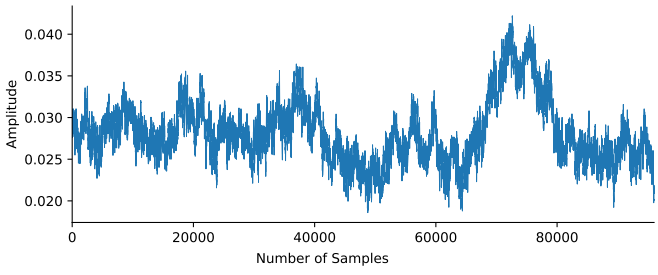
Question:
This is a phonocardiogram (PCG) recording of heart sounds, with a single channel, a sampling rate of 4000 Hz, and a duration of 25.06 seconds. Based on this PCG signal, please identify the auscultation location (aortic valve, tricuspid valve, pulmonary valve, or mitral valve).
Answer:
This recording is identified as pulmonary valve.

Electromagnetic: radio signal modulation type



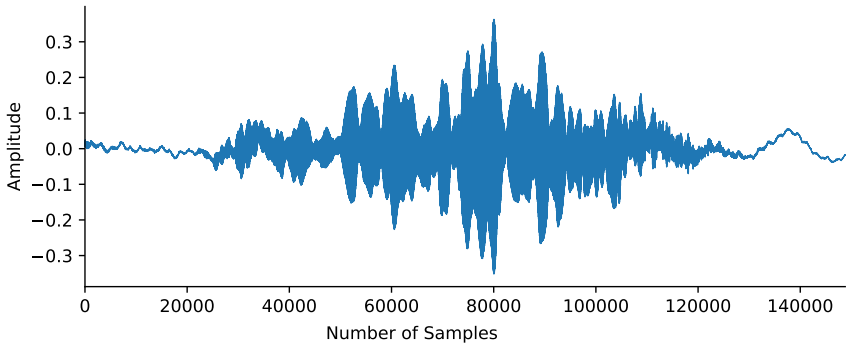
Question:
Given this two-channel I/Q sequence, identify the modulation scheme. The answer must be one of: 8PSK, AM-DSB, AM-SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, QPSK, WBFM.
Answer:
The modulation type is AM-SSB.

Underwater Acoustics: vessel type



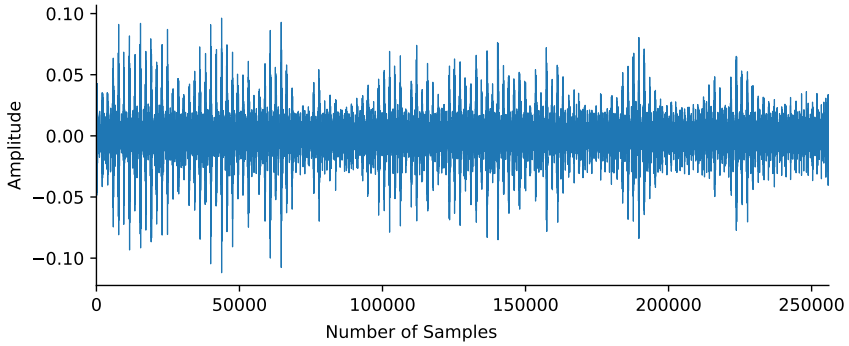
Question:
Given an underwater acoustic recording from the DeepShip dataset, which contains real-world underwater recordings of commercial vessels collected in the Strait of Georgia, classify the vessel into one of the following 4 categories: Cargo, Passengership, Tanker, Tug. Analyze the recording and determine which category it belongs to.
Answer:
The recording is classified as Tanker.

Bioacoustics: vocalization type of marmoset calls



Question:
Now, I need your assistance to complete a marmoset classification task. The provided audio is a recording of a marmoset call. It is known that the provided data corresponds to one of the following categories: Twitter, Tsik, Phee, Trill, Infant_cry or Seep. Please determine which of the following call types it belongs to.
Answer:
This audio is classified as Phee.

Manufacturing: bearing fault diagnosis from vibration



Question:
Now, I need your assistance to complete a fault classification task. This is a bearing fault diagnosis task. It is known that the provided vibration signal sample corresponds to one of the following categories: NC, IF0.2, IF0.4, IF0.6, OF0.2, OF0.4, OF0.6, RF0.2, RF0.4, RF0.6. Please identify the classification result represented by this data.
Answer:
The classification result is: RF0.4.