MRMAD: A Multi-Round Multi-Audio Benchmark for Evaluating Acoustic Degradation Perception in Large Audio-Language Models

Yize Li[1, 2*†] **Ningyuan Yang**[3*] **Sile Yin**[2] **Sindhuja Thogarrati**[2]
Sung-En Chang[2] **Andrew C. Singer**[3] **Xue Lin**[1] **Chuan-Che Huang**[2] **Shuo Zhang**[2†]
[1]Northeastern University [2]Bose Corporation [3]Stony Brook University

Abstract

Large audio-language models (LALMs) have shown promising progress in understanding speech, music, and general sound events, yet their ability to reason about how audio signals are degraded remains underexplored. Existing benchmarks primarily evaluate semantic understanding, event recognition, or high-level audio reasoning, leaving a basic question unanswered: **Do LALMs understand the differences in audio quality?** We introduce MRMAD, a Multi-Round Multi-Audio Degradation benchmark for evaluating audio degradation perception and understanding in LALMs. MRMAD spans speech, music, and sound, and frames evaluation as multi-turn dialogues across multiple audio inputs, requiring models to identify types of degradation, compare severity, and perceive corruption changes across turns. Unlike current single-turn audio-language benchmarks, MRMAD evaluates whether LALMs can maintain consistent degradation hypotheses with new evidence and comprehend low-level acoustic phenomena over multi-turn dialogues. Through a systematic evaluation of 18 representative LALMs from non-thinking to reasoning and Omni models, we find that current models often recognize coarse content while failing to diagnose, compare, or reason about degradations reliably. Human evaluations further reveal a significant perception gap between LALMs and human listeners. MRMAD thus exposes a critical yet overlooked aspect of audio-language understanding and provides a diagnostic foundation for building future LALMs that are robust to real-world acoustic conditions[1].

1 Introduction

Large Audio-Language Models (LALMs) and Omni language models (OLMs) have recently shown strong progress in audio-centered interaction, enabling models to perceive and reason over speech, music, and sound events through natural language interfaces (Tang et al., 2024; Chu et al., 2024; Georgiev et al., 2024; Kong et al., 2025; Li et al., 2025; Wang et al., 2025d; You et al., 2026; Li et al., 2026a; Xiao et al., 2026a; Sun et al., 2026a; Tong et al., 2026). Along with these models, numerous audio-language benchmarks have been proposed to evaluate audio understanding, instruction following, and complex reasoning across diverse auditory domains (Weck et al., 2024; Yang et al., 2024; Wang et al., 2025a; Sakshi et al., 2025; Ma et al., 2025; Wang et al., 2026a; Sun et al., 2026b; Yang et al., 2026a). These benchmarks have substantially advanced the evaluation of semantic audio understanding, including speech comprehension, sound event recognition, music reasoning, and audio-grounded question answering.

Despite this progress, a fundamental auditory capability remains underexplored—judging audio quality, both objective and perceptual. Real-world audio recordings are rarely pristine and can be affected by diverse degradations, including background noise, reverberation, codec compression, bandwidth limitation, and temporal discontinuities (Pascual et al., 2017; Lemercier et al., 2025; Yang et al., 2026b). For humans, recognizing those distortions is often a prerequisite for assessing audio quality and determining how audio should be restored or enhanced. For LALMs, however, it remains unclear whether they can reliably perceive these low-level quality differences rather than relying primarily on high-level semantic cues.

Existing benchmarks are not designed to answer this question. General audio-language benchmarks such as AIR-Bench (Yang et al., 2024), AudioBench (Wang et al., 2025a), MMAU (Sakshi et al., 2025), MMSU (Wang et al., 2026a), and MMAR (Ma et al., 2025) mainly evaluate content understanding, instruction follow-

*Equal contribution. Work done during an internship at Bose Corporation. †Correspondence to: Yize Li <li.yize@northeastern.edu>, Shuo Zhang <SHUO_ZHANG@bose.com>.

[1]Code and benchmark: https://github.com/Bose/MRMAD

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47028–47054
October 24-29, 2026 ©2026 Association for Computational Linguistics

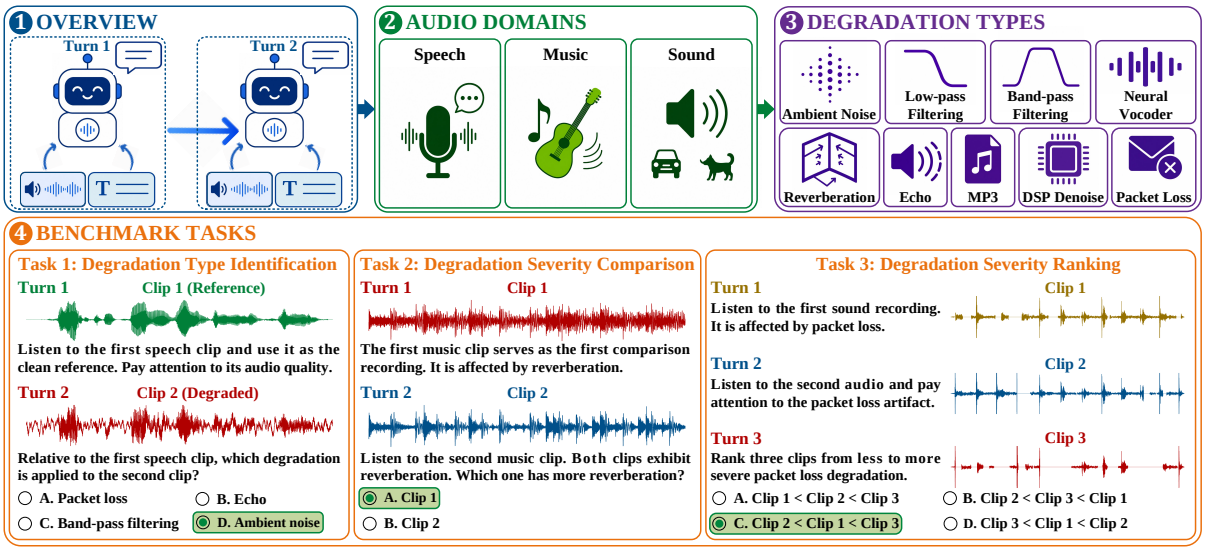


Figure 1: MRMAD benchmark overview. MRMAD evaluates multi-turn multi-audio degradation perception and quality understanding across three audio domains, nine degradation types, and three tasks.

ing, domain knowledge, or multi-step reasoning. Music-specific benchmarks, including MuChoMusic (Weck et al., 2024), focus on musical concepts and cultural or functional reasoning. Speech quality benchmarks such as QualiSpeech (Wang et al., 2025b) move closer to low-level perceptual assessment, but remain limited to speech and single-audio evaluation.

To systematically evaluate audio quality understanding, an intrusive setting with reference audio is needed for comparison. However, concatenating multiple audio clips into a single input would require models to perform accurate temporal localization to determine which segment corresponds to each prompt, introducing an additional confounding factor. A multi-round, multi-audio format is therefore better suited for this setting: each round provides one audio clip with its corresponding text prompt, enabling comparison through conversational context.

To address this gap, we introduce ***MRMAD***, a **M**ulti-**R**ound **M**ulti-**A**udio **D**egradation benchmark for evaluating audio quality understanding in LALMs. MRMAD comprises three basic tasks—*degradation type identification* (DTI), *degradation severity comparison* (DSC), and *degradation severity ranking* (DSR)—spanning across speech, music, and sound, while covering nine common degradation types. To the best of our knowledge, it is the ***first*** unified benchmark that explicitly targets audio quality perception and understanding in LALMs.

Our contributions are summarized as follows:

- We introduce MRMAD, the first multi-round, multi-audio benchmark for evaluating audio quality understanding in LALMs, complementing evaluations that focus on semantic audio understanding and general reasoning.
- We design three tasks comprising 8,400 multiple-choice questions across speech, music, and sound to assess whether models can identify degradation types, compare degradation severity, and rank degraded clips.
- We comprehensively evaluate 18 LALMs on MRMAD using a reproducible evaluation framework, revealing substantial limitations of current models in degradation-aware auditory perception, cross-audio comparison, and multi-turn severity reasoning.
- We conduct a small-scale evaluation comparing representative LALMs with human listeners, demonstrating that LALMs still notably underperform humans on our benchmark.

2 Related Work

Large Audio-Language Models. Recent LALMs have extended unimodal LLMs to the auditory modality by connecting audio encoders with instruction-tuned language models. Models such as SALMONN (Tang et al., 2024) and Qwen2-Audio (Chu et al., 2024) demonstrate broad capabilities across speech, music, and general sound. Large Audio Reasoning Models (LARMs) fur-

Table 1: Comparison between MRMAD and existing audio-language benchmarks in terms of benchmark size (number of questions), audio-domain coverage, multi-turn design, and evaluation focus.

Benchmark	Size	Domain			Multi-Turn	Degradation Perception & Quality Understanding
		Speech	Music	Sound		
AIR-Bench (Yang et al., 2024)	19k	✓	✓	✓	✗	✗
MuChoMusic (Weck et al., 2024)	1.2k	✗	✓	✗	✗	✗
AudioBench (Wang et al., 2025a)	100k+	✓	✓	✓	✗	✗
QualiSpeech (Wang et al., 2025b)	4.08k	✓	✗	✗	✗	✓
MMAU (Sakshi et al., 2025)	10k	✓	✓	✓	✗	✗
MMAR (Ma et al., 2025)	1k	✓	✓	✓	✗	▼ (Extremely Limited)
MMSU (Wang et al., 2026a)	5k	✓	✗	✗	✗	✗
MRMAD (*ours*)	8.4k	✓	✓	✓	✓	✓

ther emphasize complex auditory reasoning, long-audio understanding, and chain-of-thought supervision (Ghosh et al., 2024; Xie et al., 2025; Ghosh et al., 2025b). In parallel, OLMs such as Qwen-Omni (Xu et al., 2025a,b) and Gemini (Georgiev et al., 2024; Anil et al., 2025), move toward unified interaction across text, audio, and vision. Despite these advances, existing models are primarily optimized and evaluated for semantic understanding, instruction following, and high-level reasoning. Their ability to perceive and reason about low-level audio quality attributes remains insufficiently studied, especially in multi-turn, multi-audio conversational settings (Yang et al., 2026a).

Audio-Language Benchmarks. A growing number of benchmarks have been proposed for audio-language models. AIR-Bench (Yang et al., 2024) evaluates LALMs through chat-based tasks over diverse audio domains, while AudioBench (Wang et al., 2025a) covers speech understanding, audio scene understanding, and paralinguistic voice understanding. In addition, MMAU (Sakshi et al., 2025) increases difficulty with expert-level knowledge and multi-step reasoning, and MMAR (Ma et al., 2025) extends evaluation to multi-domain reasoning tasks. MMSU (Wang et al., 2026a) focuses on fine-grained spoken language understanding and reasoning, including linguistic and paralinguistic phenomena. For music, MuChoMusic (Weck et al., 2024) evaluates musical concepts, cultural context, and reasoning over music tracks. Closest to our focus, QualiSpeech (Wang et al., 2025b) investigates speech quality assessment with language descriptions and reasoning. However, these benchmarks primarily focus on content understanding, event inference, domain-specific knowledge, or general reasoning, rather than systematically evaluating audio quality understanding in multi-turn, multi-audio settings. Table 1 compares our MRMAD with recent audio-language benchmarks.

3 MRMAD Benchmark

3.1 Overview

MRMAD assesses whether LALMs can identify, compare, and rank common audio degradations across diverse audio domains. The benchmark consists of 8,400 multiple-choice questions paired with carefully constructed audio clips. Spanning speech, music, and sound, MRMAD covers nine degradation types: *ambient noise*, *low-pass filtering*, *band-pass filtering*, *MP3 compression*, *reverberation*, *echo*, *neural vocoder artifacts*, *DSP denoising artifacts*, and *packet loss*.

MRMAD contains three tasks: *degradation type identification* (DTI), *degradation severity comparison* (DSC), and *degradation severity ranking* (DSR). In DTI, models identify the degradation type in a degraded clip with access to its clean reference. In DSC, models compare two degraded clips and determine which one has higher or lower severity. In DSR, models rank three degraded clips by degradation severity. These tasks use a multi-turn, multi-audio format, requiring models to integrate information across rounds rather than making isolated single-clip predictions. This design enables a more comprehensive evaluation of fine-grained perceptual reasoning over audio quality. Table 2 summarizes the source datasets used to construct MRMAD across speech, music, and sound. Figure 2 illustrates the overall data construction pipeline.

3.2 Data Construction

3.2.1 Source Audio Data

We selected source datasets according to two main criteria: audio quality and domain diversity. First, since MRMAD aims to evaluate fine-grained audio quality understanding, the original audio clips should be of sufficiently high quality so that the

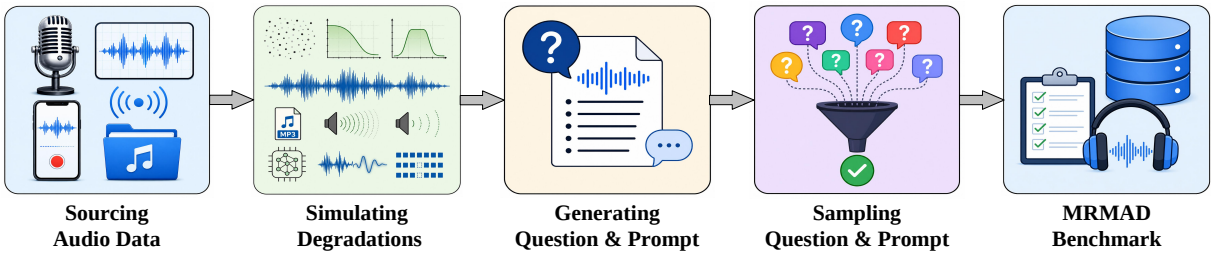


Figure 2: Data construction pipeline of MRMAD. For each collected audio clip, we simulate all domain-specific degradations at three severity levels. We then generate task-specific multi-turn questions and prompts for DTI, DSC, and DSR based on the degraded audio clips. Finally, we sample a balanced subset to form the benchmark.

Table 2: Summary of the benchmark composition across speech, music, and sound. For each audio type, we report the dataset sources and the number of questions.

Type	Dataset Sources	Num
Speech	VCTK (Yamagishi, 2012) LJSpeech (Ito and Johnson, 2017) HiFi-TTS (Bakhturina et al., 2021) LibriSpeech (Panayotov et al., 2015)	3,150
Music	URMP (Li et al., 2018) MUSDB18-HQ (Rafii et al., 2019) FMA-small (Defferrard et al., 2017)	2,800
Sound	FSD50K (Fonseca et al., 2021) TAU Urban 2019 (Mesaros et al., 2019)	2,450

introduced degradations are perceptually distinguishable. Second, to ensure broad coverage, we collected audio clips from multiple datasets across three domains. For speech, we sourced recordings of a diverse set of speakers from four high-quality speech corpora. Specifically, we included 108 speakers from VCTK (Yamagishi, 2012), 80 from LibriSpeech *dev-clean* and *test-clean* subsets (Panayotov et al., 2015), 3 from HiFi-TTS *clean* subset (Bakhturina et al., 2021), and 1 from LJSpeech (Ito and Johnson, 2017). For each speaker, we sampled approximately 28 utterances, resulting in 5,400 speech clips in total. For music, we sampled songs from MUSDB18-HQ (Rafii et al., 2019), URMP (Li et al., 2018), and FMA-small (Defferrard et al., 2017). To improve musical diversity, we selected at most one song from each artist or composer. For each song, we randomly extracted two non-overlapping 15-second segments, yielding 4,800 music clips from 2,400 tracks. For sound, we used TAU Urban Acoustic Scenes 2019 (Mesaros et al., 2019) and FSD50K (Fonseca et al., 2021). We sampled 5 clips from each of the 100 scene-location pairs in TAU 2019, and selected 3,700 high-quality, non-speech, non-music clips from a filtered FSD50K subset (Yang et al., 2026c).

Overall, we prepared 5,400 speech clips, 4,800 music clips, and 4,200 sound clips, following a 9:8:7 ratio that matches the number of degradation types in each domain. This design balances clean-audio coverage with domain-specific degradation diversity. For noise data, we collected background noise recordings from the DEMAND dataset (Thiemann et al., 2013). All clips are converted to mono and resampled to 16 kHz to match the input format commonly used by current LALMs.

3.2.2 Audio Degradations

MRMAD is constructed by applying a diverse set of audio degradations to the prepared clips. For speech, we included all 9 degradation types. For music, we excluded echo, since delayed repetitions can be intentional musical effects and are not always perceived as distortion. For sound, we excluded ambient noise and DSP denoising artifacts because many sound clips naturally contain ambience or noise-like content, making these degradations perceptually ambiguous.

Before applying degradations, each clip is first RMS-normalized to a consistent loudness level and saved as the reference audio. For each reference, we generated the degraded clips at three severity levels—*audible*, *medium*, and *strong*—using different parameter settings. To keep severity perceptually comparable across domains, we slightly adjusted parameter ranges for speech, music, and sound, which differ in spectral structure, temporal dynamics, and perceptual sensitivity. Further details are provided in Appendix A.

3.2.3 Question & Prompt Generation

We generated multiple-choice questions using a template-based prompting procedure. Each question follows a multi-round format, where models receive one audio clip and one text prompt in each

round. To reduce prompt-specific bias, we prepared multiple natural-language templates and randomly sampled one per question. We also appended an explicit answer constraint, requiring the model to output only a letter, such as A, B, C, or D. Examples and details are illustrated in Appendix F.

Degradation Type Identification. DTI evaluates whether LALMs can identify a degradation type by comparing a degraded clip with its clean reference. Each DTI question has two turns: the first provides the reference audio, and the second provides the degraded audio with four candidate degradation types. We use only the *strong* level to ensure the artifact is clearly perceptible. Three distractors are sampled from valid degradation types outside the correct answer's perceptual group. We categorize {*low-pass filtering, band-pass filtering, MP3 compression*} as spectral or spectro-temporal distortions; {*reverberation, echo*} as temporal effects; {*neural vocoder, DSP denoising*} as algorithm-induced processing artifacts; ambient noise and packet loss as two separate groups. This avoids overly ambiguous choices and encourages identification of the dominant degradation category.

Degradation Severity Comparison. DSC evaluates whether LALMs can compare the severity of the same degradation across two clips. Each question consists of two turns: the first provides a degraded clip and its degradation type, and the second provides another clip with the same degradation type but a different severity level for comparison. Both clips are generated from the same reference, with one at the *audible* level and the other at the *strong* level, ensuring a clear perceptual contrast. For each instance, we randomized the order of the two clips and whether the prompt asks for higher or lower severity, reducing fixed response patterns.

Degradation Severity Ranking. DSR evaluates whether LALMs can rank three audio clips by the severity of the same degradation type. Each question has three turns, presenting three degraded versions of the same reference at *audible*, *medium*, and *strong* levels. The clips are randomly ordered across turns, and the prompt asks for either ascending or descending severity ranking. Inspired by MuChoMusic (Weck et al., 2024), the answer space contains four ranking options: the correct ranking, two close distractors formed by swapping adjacent positions, and one easier distractor sampled from the remaining incorrect permutations. This design evaluates fine-grained severity reasoning beyond identifying only the most or least degraded clip.

3.2.4 Question & Prompt Sampling

After question generation, we sampled a balanced subset to ensure broad and controlled coverage in the final benchmark. It contains 8,400 questions following a 3:3:1 split across DTI, DSC, and DSR, with each instance corresponding to a unique reference clip. For speech, each task covers all 192 speakers with balanced questions per speaker. We also assigned an equal number of questions to each degradation type and restricted each speaker-degradation pair to appear at most once per task. For music, sampling is performed at the track level. DTI and DSC are track-disjoint and jointly cover all 2,400 tracks. To avoid segment overlap, DTI and DSC use one segment from each selected track, while DSR samples only a subset of the remaining segments. Each task also maintains balanced degradation coverage. For sound, sampling is performed at the clip level. Each selected question uses a unique reference clip. Similarly, we assigned an equal number of questions to each degradation for each task. Table 3 summarizes the overall statistics of the MRMAD benchmark.

Table 3: Summary statistics of MRMAD, including question counts, degradation coverage, domain and task distributions, and average audio and question lengths.

Statistics	Number
Total Questions	8,400
Audio domains	3
Degrad. Types (Speech : Music : Sound)	9 : 8 : 7
# Questions (Speech : Music : Sound)	9 : 8 : 7
# Questions (DTI : DSC : DSR)	3 : 3 : 1
Average Audio Length	9.1 s
# Speech : Music : Sound	4.7 : 15.0 : 8.1
Average Question Length	73.8 words
# Speech : Music : Sound	70.9 : 63.5 : 113.4

4 Experimental Setup

4.1 Models

We evaluated 18 recent open-source and closed-source models on MRMAD, with each model following its official inference configuration.

LALMs. SALMONN (Tang et al., 2024), Qwen2-Audio (Chu et al., 2024), Step-Audio-2-mini (Wu et al., 2025), Audio Flamingo 3 (Ghosh et al., 2025a), Audio Flamingo Next (Ghosh et al., 2026), MiMo-Audio-Instruct (Zhang et al., 2025), Voxtral-Mini (Liu et al., 2025), and GPT-Audio-1.5[2].

[2] https://developers.openai.com/api/docs/models/gpt-audio-1.5

Table 4: Performance comparison of open- and closed-source audio-language models on MRMAD across three tasks and audio domains. Average scores are computed as correct predictions divided by total samples. Best results are highlighted in ***bold***.

Models	Size	DTI (Task 1, %↑)				DSC (Task 2, %↑)				DSR (Task 3, %↑)			
		Speech	Music	Sound	Avg.	Speech	Music	Sound	Avg.	Speech	Music	Sound	Avg.
Random Guess	–	25.00	25.00	25.00	25.00	50.00	50.00	50.00	50.00	25.00	25.00	25.00	25.00
Open-Source		*Large Audio Language Models (LALMs)*											
SALMONN	7B	23.56	25.67	26.57	25.14	49.11	50.58	49.14	49.61	19.78	25.00	24.86	23.00
Qwen2-Audio	8.4B	24.22	26.25	28.00	26.00	41.78	45.42	50.48	45.53	28.22	26.50	26.86	27.25
Step-Audio-2-mini	8B	24.96	28.75	28.86	27.36	52.96	50.08	50.67	51.33	25.11	29.75	30.29	28.17
Audio Flamingo 3	7B	27.56	27.00	23.90	26.31	51.19	48.83	***51.52***	50.50	23.33	28.50	30.57	27.17
Audio Flamingo Next	8B	26.00	28.00	23.52	25.94	52.52	47.92	49.62	50.14	25.78	31.00	31.71	29.25
MiMo-Audio-Instruct	7B	***36.22***	***32.83***	***31.05***	***33.58***	***55.11***	***51.25***	51.05	***52.64***	***28.23***	***32.50***	***34.57***	***31.50***
Voxtral-Mini	4.7B	24.07	27.33	25.71	25.64	44.52	43.17	44.67	44.11	25.78	27.75	31.71	28.17
		Large Audio Reasoning Models (LARMs)											
Step-Audio-R1	33B	29.78	28.17	30.48	29.44	55.26	54.08	51.62	53.81	27.11	28.50	27.43	27.67
Audio-Reasoner	8.4B	18.81	23.08	24.57	21.92	43.04	32.58	39.62	38.56	15.78	10.75	14.29	13.67
Qwen3-Omni-Thinking	30B-A3B	***37.78***	***39.75***	***31.43***	***36.58***	***61.48***	***61.50***	***58.75***	***60.70***	***27.56***	***34.50***	***36.29***	***32.42***
Audio Flamingo 3-Think	7B	26.30	30.33	25.90	27.53	51.26	49.83	51.52	50.86	24.44	27.00	28.57	26.50
Audio Flamingo Next-Think	8B	26.22	24.08	19.62	23.58	49.63	48.42	52.48	50.06	25.78	25.75	24.00	25.25
MiMo-Audio-Think	7B	35.11	31.92	29.71	32.47	46.81	49.00	48.48	48.03	24.67	27.00	28.29	26.50
		Omni Language Models (OLMs)											
Qwen2.5-Omni	10.7B	27.33	30.50	29.90	29.14	50.30	49.58	50.67	50.17	***34.00***	***33.25***	***33.43***	***33.58***
AudioMCQ-Weak-To-Strong	10.7B	27.19	28.33	27.90	27.78	52.22	50.25	52.10	51.53	31.78	30.25	32.86	31.58
Omni-R1	10.7B	27.85	29.75	30.29	29.19	51.26	49.50	51.05	50.61	28.22	29.25	31.14	29.42
Qwen3-Omni-Instruct	30B-A3B	***31.70***	***42.83***	***31.33***	***35.31***	***52.74***	***52.25***	***52.95***	***52.64***	22.00	27.25	29.43	25.92
Closed-Source		*LALMs & LARMs & OLMs*											
GPT-Audio-1.5	–	26.81	30.00	27.43	28.06	49.14	48.17	23.11	33.28	23.11	24.75	20.29	22.83
Gemini 3 Flash	–	59.63	43.92	41.05	48.97	67.11	60.33	61.33	63.17	40.67	39.50	38.57	39.67
Gemini 3.1 Flash-Lite	–	36.89	32.92	30.57	33.72	52.22	53.58	51.52	52.47	34.00	30.25	31.14	31.92
Gemini 3.1 Pro	–	***84.81***	***83.58***	***91.05***	***86.22***	***88.82***	***92.50***	***92.10***	***90.81***	***58.44***	***69.50***	***65.71***	***64.25***
Gemini 3.5 Flash	–	73.56	55.50	49.90	60.64	84.30	76.50	72.00	78.11	55.56	50.00	46.57	51.08

LARMs. Step-Audio-R1 (Tian et al., 2025), Audio-Reasoner (Xie et al., 2025), Qwen3-Omni-Thinking (Xu et al., 2025b), Audio Flamingo 3-Think (Ghosh et al., 2025a), Audio Flamingo Next-Think (Ghosh et al., 2026) and MiMo-Audio-Think (Zhang et al., 2025).

OLMs. Qwen2.5-Omni (Xu et al., 2025a), AudioMCQ-Weak-To-Strong (He et al., 2026), Omni-R1 (Zhong et al., 2025), Qwen3-Omni-Instruct (Xu et al., 2025b), and various Gemini 3 models[3], including Gemini 3 Flash / 3.1 Flash-Lite / 3.1 Pro / 3.5 Flash (Anil et al., 2025).

4.2 Evaluation scheme

We evaluate MRMAD with an output-based multiple-choice protocol. For each instance, the model receives multi-round audio inputs, text prompts, and candidate options, and returns a single option label in the final turn. All models use the same multi-turn inference templates and answer parser for fair comparison. Predictions are extracted by matching either the option label or a unique option string with rule-based regular expressions. Further details are provided in Appendix C.

[3] https://aistudio.google.com/models/gemini-3 and models are collected before 06/2026.

5 Results and Discussions

5.1 Main Results

Table 4 presents the performance of open-source and closed-source models on MRMAD. Further analysis is provided in Appendix B. We summarize the main observations as follows.

MRMAD poses a substantial challenge to current LALMs. Most open-source models perform only marginally above random guessing on MRMAD. On DTI and DSR, most models remain close to the 25% random baseline, while on the DSC task, several models are near the 50% binary-choice baseline. This indicates that current audio-language models, despite their progress on general audio understanding and reasoning, still struggle to capture low-level degradation cues.

Closed-source models lead, but the advantage is not uniform. Gemini 3.1 Pro achieves the best performance across all three tasks, with 86.22% on DTI, 90.81% on DSC, and 64.25% on DSR. Compared with the best open-source model on each task,

Table 5: Results of human evaluation and LALM performance. Average scores reflect the proportion of correct predictions relative to the total sample count. The best results are indicated in ***bold***, and the second-best ones in *italics*.

Models	DTI (Task 1, %↑)				DSC (Task 2, %↑)				DSR (Task 3, %↑)			
	Speech	Music	Sound	Avg.	Speech	Music	Sound	Avg.	Speech	Music	Sound	Avg.
Random Guess	25.00	25.00	25.00	25.00	50.00	50.00	50.00	50.00	25.00	25.00	25.00	25.00
Expert	***100.00***	***97.22***	***98.41***	***98.61***	***100.00***	***97.22***	***100.00***	***99.07***	***100.00***	***95.83***	***100.00***	***98.61***
Non-Expert	*86.42*	72.22	82.54	80.55	*97.53*	*93.06*	90.48	*93.98*	*88.89*	*91.67*	*80.95*	*87.50*
SALMONN	14.81	20.83	33.33	22.22	40.74	54.17	52.38	48.61	22.22	62.50	28.57	37.50
Audio Flamingo 3	25.93	25.00	19.05	23.61	66.67	41.67	52.38	54.17	22.22	75.00	28.57	41.67
MiMo-Audio-Instruct	29.63	29.17	19.05	26.39	44.44	45.83	42.86	44.44	33.33	37.50	28.57	33.33
Qwen3-Omni-Thinking	37.04	58.33	14.29	37.50	66.67	75.00	52.38	65.28	44.44	62.50	57.14	54.17
Gemini 3.1 Pro	85.19	*75.00*	*90.48*	*83.33*	77.78	87.50	*100.00*	87.50	66.66	75.00	71.42	70.83

it improves by 49.64, 30.11, and 30.67 absolute points, respectively. However, GPT-Audio-1.5 performs poorly, particularly on DSC, suggesting that closed-source deployment alone does not guarantee robust degradation-aware auditory perception.

Open-source models remain far from reliable audio degradation understanding. Among open-source models, Qwen3-Omni-Thinking achieves the best average performance on DTI and DSC, with 36.58% and 60.70%, respectively. Qwen2.5-Omni achieves the strongest open-source DSR result, with 33.58%. MiMo-Audio-Instruct is the best non-omni LALM, outperforming others across most tasks. Nevertheless, the large gap between these models and the Gemini 3 families indicates that existing open-source systems still lack reliable fine-grained perception and comparison capabilities of audio quality.

5.2 Reasoning vs. Non-Reasoning

We further investigate whether explicit reasoning improves degradation-aware audio understanding by comparing matched reasoning and non-reasoning variants. As shown in Table 4, the effect of reasoning is mixed rather than consistently positive. Qwen3-Omni-Thinking improves over Qwen3-Omni-Instruct on all three tasks, with gains of 1.27 points on DTI, 8.06 points on DSC, and 6.50 points on DSR, suggesting that reasoning can benefit models when the underlying audio representation is sufficiently informative. However, this trend does not generalize across model families: Audio Flamingo3-Think yields only marginal differences, while Audio Flamingo Next-Think and MiMo-Audio-Think underperform their non-reasoning counterparts on most task averages. Averaged over matched pairs, reasoning impacts performance by only -0.25, $+0.93$, and -0.79 points on DTI, DSC, and DSR, respectively. Moreover, Audio-Reasoner performs substantially below several non-reasoning baselines, indicating that reasoning-oriented training alone does not guarantee robustness on MRMAD. These results show that audio degradation understanding cannot be reduced to high-level reasoning alone. Models need to first encode fine-grained acoustic evidence, preserve these representations reliably across turns, and compare subtle perceptual cues across multiple audio inputs and dialogue contexts. Without degradation-sensitive audio representations, additional reasoning steps may amplify language priors rather than improve auditory grounding.

5.3 Human Evaluations

Although the labels in our benchmark are generated through controlled degradation rather than subjective annotation, human performance provides a valuable reference for perceptual difficulty and human-level perceptual capability. We therefore conduct a small-scale human evaluation on a subset of 168 questions (2% of MRMAD), selected to preserve the original distributions across domains and tasks. Six participants are invited: three audio-domain experts with professional experience in audio analysis or related fields, and three non-expert judges representing general listeners without specialized audio expertise. All participants independently evaluate the subset after a brief training session covering one example of each degradation type and audio type, along with a few example questions for each task. As shown in Table 5, experts achieve accuracies of 98.61%, 99.07%, and 98.61% across the three tasks, respectively, while non-expert listeners achieve 80.55%, 93.98%, and 87.50%. Except for Gemini 3.1 Pro, all models perform substantially below human evaluators, high-

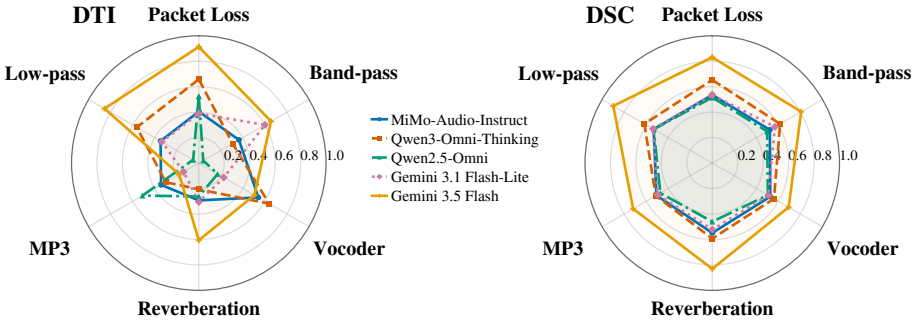


Figure 3: Degradation-wise performance of five models on DTI and DSC, averaged over speech, music, and sound. DTI exhibits stronger artifact-specific variation, while DSC is more stable across degradation types.

lighting a significant gap between human auditory perception and the capabilities of current LALMs. These results indicate that, although the tasks are tractable for human listeners, current LALMs remain substantially limited in their ability to distinguish fine-grained acoustic degradations.

5.4 Degradation Type Analysis

Figure 3 reveals that the DTI performance of different models varies substantially across degradation types. Stronger models, such as Gemini 3.5 Flash and MiMo-Audio-Instruct, perform better on packet loss and low-pass filtering, but exhibit lower accuracy on MP3 compression, band-pass filtering, and reverberation. Packet loss and low-pass filtering provide more salient perceptual cues through temporal discontinuities and broad high-frequency attenuation, suggesting that these models are more sensitive to coarse degradation cues than to subtle spectral, spatial, or artifact-specific distortions. In contrast, DSC is more uniform across degradation types, likely because the degradation category is given and the task reduces to within-type severity comparison. However, the persistent gap between closed-source and open-source models indicates that severity comparison still requires degradation-sensitive acoustic representations. Overall, the degradation-wise results show that MRMAD captures both task-level difficulty and artifact-specific weaknesses in current LALMs.

5.5 Why do LALMs Fail on MRMAD?

We conduct controlled diagnostic analyses to identify potential failure modes in LALMs. We focus on DTI that compares the clean reference audio with its degraded counterpart. All diagnostics are conducted on a fixed and balanced subset of the DTI split, covering six shared degradation types across speech, music, and sound, with 900 samples.

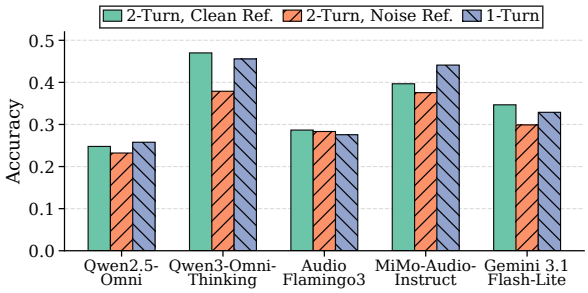
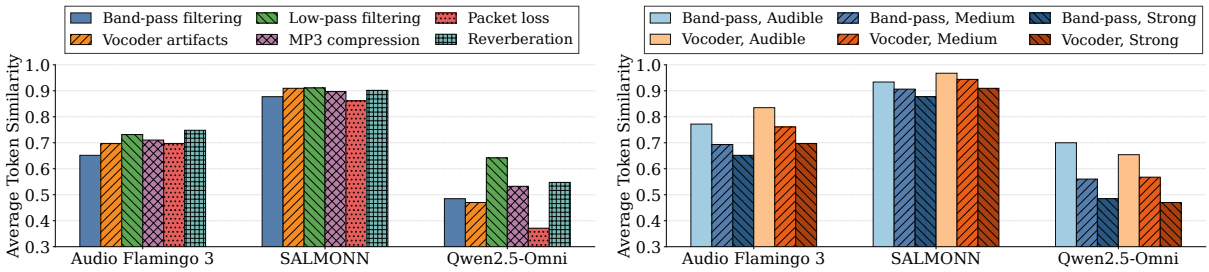


Figure 4: Performance on the DTI subset when the first-turn reference audio is either replaced with Gaussian noise or the entire first round is removed.

We select the representative models that reflect the overall performance trends. We analyze three factors: whether models effectively use the reference audio, whether the multi-turn format introduces additional difficulty, and whether audio tokenizers preserve degradation-relevant information.

Clean reference audio is under-utilized. In DTI, the clean reference in the first turn should provide essential evidence to identify the degradation in the second turn. We therefore compare the original setting with two variants: replacing the first-turn reference with Gaussian noise and even removing the first turn entirely. As shown in Figure 4, corrupting or removing the reference causes only limited performance drops for most models. Qwen2.5-Omni and Audio Flamingo 3 remain nearly unchanged, while MiMo-Audio-Instruct even performs better in the one-turn setting when the clean reference is removed. Qwen3-Omni-Thinking shows the strongest dependence on the clean reference, but its one-turn result is still close to the original setting. These results suggest that many models do not reliably perform explicit reference-to-degraded comparison; instead, they often rely on the degraded clip alone or on language priors. Importantly, this study does not imply that models ignore audio en-



(a) Average audio-token similarity between reference audio and strong-level degraded audio across six degradation types.

(b) Average audio-token similarity for band-pass filtering and neural vocoder artifacts across three degradation severity levels.

Figure 5: Token-level analysis of degradation-relevant information preservation in projected audio representations. Note that all three ALAMs apply Whisper-based audio encoders (Radford et al., 2023).

tirely, but rather that they under-utilize the reference audio required by the task.

Multi-turn context is not consistently beneficial. The one-turn variant removes the conversational dependency while preserving the degradation identification objective. If failures are mainly due to insufficient audio context, the original two-turn setting should consistently outperform the one-turn setting. However, Figure 4 shows that several models achieve comparable or higher accuracy when the first turn is deleted. This indicates that current LALMs may struggle to retain and compare audio evidence across turns, and that additional multi-turn context can introduce misalignment or memory burden rather than consistently improving reference-based reasoning (Laban et al., 2026; Xiao et al., 2026b).

Audio tokenizers often suppress degradation information. We further analyze whether the audio representations passed to the language model preserve sufficient degradation-relevant information. Given a reference clip x and its degraded counterpart $\tilde{x}$, let $\mathbf{Z}, \tilde{\mathbf{Z}} \in \mathbb{R}^{N \times d}$ denote their projected audio-token sequences after the final audio projection layer, where N is the number of aligned tokens and d is the embedding dimension. We compute the average token similarity as $\bar{s}(x, \tilde{x}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{z}_i^\top \tilde{\mathbf{z}}_i}{\|\mathbf{z}_i\|_2 \|\tilde{\mathbf{z}}_i\|_2}$. Lower similarity indicates greater sensitivity to degradation-induced acoustic changes, while consistently high similarity suggests invariance to low-level quality distortions. We consider models with audio encoders from the Whisper family (Radford et al., 2023). As shown in Figure 5, SALMONN maintains high similarity between clean and strongly degraded audio across degradation types, suggesting that its projected tokens are largely invariant to artifact-level changes. Audio Flamingo 3 presents moderate sensitivity, whereas Qwen2.5-Omni exhibits lower and more degradation-dependent similarity. The degradation severity analysis is consistent with this pattern: Qwen2.5-Omni shows a clearer decrease in similarity score from audible to strong severity for band-pass filtering and vocoder artifacts, while SALMONN remains highly similar across severity levels. These results reveal that some LALMs may fail on MRMAD because degradation cues are attenuated before reaching the language model, limiting the evidence available for type identification and severity reasoning.

These diagnostics indicate that failures are from coupled grounding and representation bottlenecks. Models often under-utilize the clean reference and do not consistently benefit from multi-turn audio context, while their audio tokens may attenuate degradation cues before language decoding.

6 Conclusion

We propose MRMAD, a multi-round multi-audio benchmark for evaluating degradation-aware auditory understanding in LALMs across audio domains. It complements existing benchmarks by targeting low-level quality perception through DTI, DSC, and DSR. We show that current LALMs remain limited in fine-grained degradation perception and multi-audio severity reasoning, due to unstable reasoning gains, degradation-dependent behavior, weak cross-turn grounding, and degradation-insensitive audio representations. Furthermore, human evaluations reveal a clear limitation in their ability to identify acoustic degradation. Therefore, MRMAD highlights degradation-aware perception as a critical missing capability for present audio-language modeling.

Limitations

First, MRMAD is designed as an evaluation benchmark, and we do not propose novel training meth-

ods or architectural solutions to improve LALMs. Accordingly, while our analysis identifies limitations in degradation-aware perception, cross-turn grounding, and audio-token representations, addressing these issues is left to future work. Second, several strong closed-source models, particularly the Gemini 3 family, are also among the strongest general-purpose multimodal models. Since their detailed architectures, training data, and audio-processing pipelines are not publicly available, we cannot fully disentangle whether their gains arise from audio-specific modeling, broader multimodal capabilities, scale, or proprietary training procedures. Then, the evaluated open-source LALMs do not invoke any external tools without agentic intervention (Tong et al., 2026; Wang et al., 2026b,c; Chen et al., 2026). Finally, due to the limited availability of expert human annotators for audio quality assessment, we only report human performance from a relatively small-scale evaluation to primarily quantify the discrepancy between LALMs and human auditory perception.

Ethics Statement

MRMAD is merely constructed from publicly available audio datasets and simulated degradations. The benchmark is intended for evaluating audio-language models on perceptual degradation understanding, not for identifying speakers, inferring private attributes, or making high-stakes decisions. All audio sources are used in accordance with their original licenses and usage terms. The benchmark may reveal weaknesses in current LALMs, but it does not introduce methods for generating harmful content or manipulating real-world audio systems.

References

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, and 1331 others. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-Fi Multi-Speaker English TTS Dataset. *arXiv preprint arXiv:2104.01497*.

Liyang Chen, Hongkai Chen, Yujun Cai, Sifan Li, Qingwen Ye, and Yiwei Wang. 2026. Audiorouter: Data efficient audio understanding via rl based dual reasoning. *arXiv preprint arXiv:2602.10439*.

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

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Heming Fu, Weici Pan, Liangkai Zhou, Zeyu Zhang, Zhenhua Liu, and Shan Lin. 2025. Mallm: Multi-agent decision-making with llms for multi-user edge-sensor environments. *ACM SIGMETRICS Performance Evaluation Review*, 53(2):3–8.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, and 1117 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. 2026. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025a. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025b. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19358–19405. PMLR.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, and 1 others. 2026. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*, volume 2026, pages 110411–110432.

Kaiyuan Hou, Yunqi Guo, Heming Fu, Hongkai Chen, Zhenyu Yan, Guoliang Xing, and Xiaofan Jiang. 2024. Improving on-device llms' sensory understanding with embedding interpolations. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*.

Qingqiao Hu, Weimin Lyu, Meilong Xu, Kehan Qi, Xiaoling Hu, Saumya Gupta, Jiawei Zhou, and Chao Chen. 2025. Loc-path: Learning to compress for pathology multimodal large language models. *arXiv preprint arXiv:2512.05391*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. 2025. Token reduction should go beyond efficiency in generative models – from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pages 54738–54778.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. 2025. Diffusion models for audio restoration: A review. *IEEE Signal Processing Magazine*, 41(6):72–84.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. 2018. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535.

Yuqi Li, Junhao Dong, Jiao Liu, Piotr Koniusz, Hansheng Zeng, Chuanguang Yang, Junming Liu, Yingli Tian, Tingwen Huang, and Hao Wu. 2026a. Evolving multimodal models for physical dynamics: A multi-objective neuroevolution approach. *IEEE Transactions on Evolutionary Computation*.

Yuqi Li, Xi Xiao, Yunbei Zhang, Lin Zhao, Yu Li, Aiden Zhao, Tianyang Wang, Hao Xu, and Yingli Tian. 2026b. Rethinking layer-wise information allocation for vision foundation model adaptation. *arXiv preprint arXiv:2607.21973*.

Yuqi Li, Zijie Zhou, Zhiyuan Peng, Junhao Dong, Haochen You, Renye Yan, Shiping Wen, Yingli Tian, and Tingwen Huang. 2025. A preference-driven methodology for efficient code generation. *IEEE Transactions on Artificial Intelligence*.

Hanzhong Liang, Jinghao Shi, Xiang Shen, Zixuan Wang, Vera Wen, Ardalan Mehrani, Zhiqian Chen, Yifan Wu, and Zhixin Zhang. 2025. Embedding-based retrieval in multi-modal content moderation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4264–4268.

Ju Lin, Yun Wang, Kaustubh Kalgaonkar, Gil Keren, Didi Zhang, and Christian Fuegen. 2021. A time-domain convolutional recurrent network for packet loss concealment. In *ICASSP*. IEEE.

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

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2019. TAU Urban Acoustic Scenes 2019, Development Dataset.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, and 28 others. 2026. Multifinben: Benchmarking large language models for multilingual and multimodal financial application. pages 16934–16963.

Barlian Henryranu Prasetio, Delfi Olivia Yusuf, Dahnial Syauqy, and Syahrul Chilmi. 2024. Spectral gating for noise reduction in speech stress recognition system. In *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 149–155. IEEE.

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*, pages 28492–28518. PMLR.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2019. MUSDB18-HQ - an uncompressed version of musdb18.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, FranÃğois Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pages 84929–84964.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Jinuo Sun, Yang Xiao, Sung Kyun Chung, Qiuchi Hu, Gongping Huang, Eun-Jung Holden, and Ting Dang. 2026a. Activation steering for accent adaptation in large audio language models. *arXiv preprint arXiv:2603.05813*.

Yirong Sun, Yanjun Chen, Xin Qiu, Gang Zhang, Hongyu Chen, Daokuan Wu, Chengming Li, Min Yang, Dawei Zhu, Wei Zhang, and Xiaoyu Shen. 2026b. Sonicbench: Dissecting the physical perception bottleneck in large audio language models. *arXiv preprint arXiv:2601.11039*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. Demand: a collection of multi-channel recordings of acoustic noise in diverse environments.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*.

Siqian Tong, Li Xuan, Yiwei Wang, Baolong Bi, Yujun Cai, Shenghua Liu, Yuchen He, and HAO Chengpeng. 2026. AuTAgent: A reinforcement learning framework for tool-augmented audio reasoning. In *Forty-third International Conference on Machine Learning*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy 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 (Volume 1: Long Papers)*, pages 4297–4316.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2026a. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pages 31374–31410.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025b. Qualispeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23588–23609.

Yucheng Wang, Jing Peng, Hanqi Li, Chenghao Wang, Wenming Tu, Yu Xi, Zhaokai Sun, Kai Yu, and Shuai Wang. 2026b. Audio-mind: An auditable agentic framework for audio understanding. *arXiv preprint arXiv:2605.28480*.

Yuwen Wang, Tian-Hao Zhang, Minghao Cai, Yilin Ren, Ziyang Jiang, Xin Wang, Zhichao Wang, Pan Zhou, Kun Zhan, and Xinyuan Qian. 2026c. Hear, invoke, and understand: A skill-calling multimodal agent for large audio language models. *arXiv preprint arXiv:2608.01881*.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. 2025c. Filter-and-refine: A MLLM based cascade system for industrial-scale video content moderation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*.

Zixuan Wang, Yu Sun, Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Luna Dong, Zhuolin Hao, Hongyu Xiong, and Yang Song. 2025d. Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1104–1112.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *ISMIR*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Yang Xiao, Eun-Jung Holden, and Ting Dang. 2026a. Adapting where it matters: Depth-aware adaptation for efficient multilingual speech recognition in low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 16776–16788.

Yang Xiao, Siyi Wang, Han Yin, Hong Jia, Vidhyasaharan Sethu, Eun-Jung Holden, and Ting Dang. 2026b. Why can't they remember? uncovering representation and retrieval bottlenecks in multi-turn acoustic memory. *arXiv preprint arXiv:2605.27039*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

Siheng Xiong, Oguzhan Gungordu, James C Kerce, and Faramarz Fekri. 2026a. Adaptive information control for search-augmented llm reasoning. *arXiv preprint arXiv:2602.01672*.

Siheng Xiong, Ali Payani, and Faramarz Fekri. 2026b. Enhancing language model reasoning with structured multi-level modeling. In *International Conference on Learning Representations*, volume 2026, pages 36557–36610.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 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. *arXiv preprint arXiv:2509.17765*.

Junichi Yamagishi. 2012. English multi-speaker corpus for cstr voice cloning toolkit. *URL http://homepages. inf. ed. ac. uk/jyamagis/page3/page58/page58. html*.

Chih-Kai Yang, Yun-Shao Tsai, Yu-Kai Guo, Ping-Le Tsai, Yen-Ting Piao, Hung-Wei Chen, Ting-Lin Hsiao, Yun-Man Hsu, Ke-Han Lu, and Hung-yi Lee. 2026a. Mugen: Evaluating and improving multi-audio understanding of large audio-language models. *arXiv preprint arXiv:2603.09714*.

Ningyuan Yang, Yize Li, Diego A Cuji, Ryan M Corey, Pu Zhao, Xue Lin, and Andrew C Singer. 2026b. A survey of advancing audio super-resolution and bandwidth extension from discriminative to generative models. *arXiv preprint arXiv:2605.16681*.

Ningyuan Yang, Sile Yin, Li-Chia Yang, Bryce Irvin, Xiao Quan, Marko Stamenovic, and Shuo Zhang. 2026c. Fsd50k-solo: Automated curation of single-source sound events. *arXiv preprint arXiv:2605.13931*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

Ruochu Yang, Milind Rao, Harshavardhan Sundar, Anirudh Raju, Aparna Khare, Srinath Tankasala, Di He, and Venkatesh Ravichandran. 2025. On retrieval of long audios with complex text queries. In *Proc. Interspeech 2025*, pages 2660–2664.

Zuyao You, Zhesong Yu, Mingyu Liu, Bilei Zhu, Yuan Wan, and Zuxuan Wu. 2026. Gamma: Towards joint global-temporal music understanding in large multimodal models. *arXiv preprint arXiv:2605.00371*.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo,

Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, Chun Chen, and 80 others. 2025. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. 2025. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*.

Qingsong Zou, Jingyu Xiao, Qing Li, Zhi Yan, Yuhang Wang, Li Xu, Wenxuan Wang, Kuofeng Gao, Ruoyu Li, and Yong Jiang. 2025. Queryattack: Jailbreaking aligned large language models using structured non-natural query language. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5725–5741.

Qingsong Zou, Zhi Yan, Zhiyao Xu, Kuofeng Gao, Jingyu Xiao, and Yong Jiang. 2026. Smartbench: Evaluating llms in smart homes with anomalous device states and behavioral contexts. *arXiv preprint arXiv:2603.06636*.

Appendix

A Degradation Simulation Details

This section provides a detailed introduction to the nine degradation types used in MRMAD and their parameter configurations.

A.1 Ambient Noise

Ambient noise is simulated by mixing the reference audio with background noise recordings randomly sampled from the DEMAND dataset (Thiemann et al., 2013). For speech degradation, we exclude the `OMEETING` scene because it contains speech-like meeting recordings, which could introduce confounding speech content rather than purely background noise. The degradation severity is controlled by the signal-to-noise ratio (SNR), where lower SNR values indicate stronger noise corruption. Specifically, we sample SNRs from [10, 15] dB for the audible level, [0, 5] dB for the medium level, and [-10, -5] dB for the strong level.

A.2 Low-Pass Filtering

Low-pass filtering is simulated by attenuating high-frequency content above a sampled cutoff frequency. Lower cutoff frequencies correspond to stronger bandwidth limitation and thus more severe degradation. For speech, we sample cutoff frequencies from [4250, 5500] Hz for the audible level, [2250, 3000] Hz for the medium level, and [750, 1500] Hz for the strong level. We use domain-specific cutoff ranges for music and sound because speech, music, and sound events differ in spectral bandwidth and perceptual reliance on high-frequency content. This keeps the three severity levels perceptually comparable across domains.

A.3 Band-Pass Filtering

Band-pass filtering is simulated by preserving a frequency band around a sampled center frequency while attenuating frequencies outside the passband. Let f_c denote the center frequency and r_b denote the bandwidth fraction. The resulting passband is defined as

$$\left[f_c\left(1-\frac{r_b}{2}\right),\ f_c\left(1+\frac{r_b}{2}\right)\right], \tag{1}$$

where smaller r_b values retain a narrower frequency range and therefore produce stronger degradation.

For each reference audio clip, we keep f_c fixed across the three severity levels and vary only r_b. This isolates severity changes to bandwidth restriction, avoiding confounds from different affected frequency regions that may be perceived differently. For speech and music, we sample r_b from [1.50, 1.75] for the audible level, [0.85, 1.25] for the medium level, and [0.30, 0.60] for the strong level. For sound events, we use larger r_b at each severity level than those used for speech and music. This preserves more spectral content, since sound events often have diverse and sparse spectral patterns, making overly narrow band-pass filtering less consistently interpretable.

A.4 MP3 Compression

MP3 compression is simulated by encoding and decoding the reference audio using the `libmp3lame` codec. The degradation severity is controlled by the bitrate, where lower bitrates introduce stronger codec artifacts such as distortion, high-frequency loss, and temporal smearing. For all audio types, we use 32 kbps for the audible level, 16 kbps for the medium level, and 8 kbps for the strong level.

A.5 Reverberation

Reverberation is simulated using algorithmic room effects with three parameters: room size, damping, and wet ratio. The room size controls the

perceived spatial extent of the reverberant environment, damping controls the decay of high-frequency reflections, and the wet ratio controls the amount of reverberated signal mixed with the reference audio. To define the three severity levels, we increase the room size and wet ratio while decreasing damping, producing progressively stronger room coloration and longer reverberant tails. We use the same parameter configuration for speech and music. For sound events, we use larger room size and wet ratio values, together with lower damping values, than those used for speech and music at the same level. This makes reverberation more perceptible for short or transient sound events, where weaker room effects may be less distinguishable.

A.6 Echo

Echo is simulated using a multi-tap delay model. For each reference audio clip, we first sample a base delay, an initial tap weight, a decay factor, and the number of delay taps. The delay times are then set as integer multiples of the base delay, and the tap weights decay exponentially across taps. We keep these delay-related parameters fixed across the three severity levels and vary only the wet ratio, so that differences across levels are mainly caused by the amount of delayed signal rather than changes in echo timing. The wet ratio is sampled from [0.20, 0.30] for the audible level, [0.50, 0.60] for the medium level, and [0.80, 0.90] for the strong level. Higher wet ratios introduce more prominent delayed repetitions, resulting in more severe echo degradation.

A.7 Neural Vocoder Artifacts

Neural vocoder artifacts are simulated by reconstructing audio from corrupted log-mel representations. Given a reference waveform, we first convert it to a log-mel spectrogram, add Gaussian noise to the mel features, and then synthesize the waveform using a neural vocoder. The degradation severity is controlled by the standard deviation of the added mel-domain noise, denoted by σ_{mel}. Larger σ_{mel} values introduce stronger distortions in the reconstructed waveform, producing artifacts such as spectral smearing, unnatural timbre, and vocoder-like reconstruction errors.

For speech, we employ a 16 kHz HiFi-GAN vocoder from SpeechBrain (Ravanelli et al., 2021, 2024), which directly reconstructs waveforms from the mel-spectrograms. For music and sound events, we use a 24 kHz BigVGAN vocoder (Lee et al., 2022) because it is better suited for general audio reconstruction with broader spectral content. In this pipeline, the 16 kHz reference audio is first resampled to 24 kHz, converted to the mel-spectrograms, corrupted in the mel domain, reconstructed by BigVGAN at 24 kHz, and finally resampled back to 16 kHz. We use different σ_{mel} ranges across audio types because the vocoders and source domains differ in spectral complexity and reconstruction robustness.

A.8 DSP Denoising Artifacts

DSP denoising artifacts are simulated by first adding background noise to the reference audio and then applying spectral-gating-based noise reduction (Prasetio et al., 2024). Spectral gating estimates a frequency-dependent noise threshold from the spectrogram and applies a mask to suppress components below the threshold. While this reduces background noise, aggressive suppression can introduce perceptual artifacts such as over-suppression, signal distortion, and musical noise.

The degradation severity is mainly controlled by the noise-reduction strength α. Larger α values suppress more estimated noise but also introduce stronger denoising artifacts. We sample α from [0.950, 0.960] for the audible level, [0.980, 0.985] for the medium level, and [0.997, 1.000] for the strong level. For each reference clip, we use the same added noise segment and the same SNR across the three severity levels, so that severity differences are primarily caused by the denoising strength rather than differences in the noise condition. Other denoising parameters, including the time constant and frequency-mask smoothing, are also shared across severity levels.

A.9 Packet Loss

Packet loss is simulated by dividing the reference audio into fixed-length packets and dropping selected packets according to a two-state Markov process (Lin et al., 2021). Each packet is assigned to either a normal state N or a loss state L. The parameter p_N denotes the probability of remaining in the normal state, while p_L denotes the probability of remaining in the loss state. The theoretical packet loss rate is given as

$$\rho = \frac{1 - p_N}{(1 - p_N) + (1 - p_L)}. \tag{2}$$

When a packet is marked as lost, its samples are replaced with zeros, producing discontinuities similar

to missing audio frames in packet-based transmission. We use 30 ms packets for all audio types. To construct the three severity levels, we choose lower p_N values and higher p_L values for stronger degradations, which increases both the frequency of loss events and the duration of loss bursts.

B Additional Analysis

Severity ranking is the most difficult task. DSR yields the lowest scores for nearly all models. Even the best open-source model reaches only 33.58%, slightly above the 25% random baseline, and the best closed-source model drops substantially from its DTI and DSC performance. This suggests that ranking multiple degraded clips requires more than detecting the presence of an artifact; models must preserve severity-sensitive acoustic evidence across multiple audio inputs and conversational turns.

Model scaling law alone does not explain performance. Although larger models such as Qwen3-Omni-Thinking and Step-Audio-R1 generally perform better than several smaller baselines, parameter count is not a sufficient predictor of MRMAD performance. MiMo-Audio-Instruct remains competitive with 7B parameters, while Audio-Reasoner performs substantially below several models of similar or smaller scale. This suggests that MRMAD evaluates capabilities that depend not only on model size but also on degradation-sensitive audio representations, training scale, and the ability to ground language outputs in acoustic evidence.

C Multi-turn Inference Template

Figure 6 shows the two-turn and three-turn templates for multi-round multi-audio evaluations. In the two-turn setting, the model first receives the Turn-1 audio a_1 together with its text prompt t_1 and generates an intermediate response r_1. The second turn preserves initial textual and audio inputs and the model's response as conversation history, then provides the Turn-2 audio a_2 and prompt t_2 for the final generation. The three-turn template extends this procedure by adding another audio-text pair (a_3, t_3) while retaining all previous user and assistant turns. This design enables the model to accumulate information across multiple audio clips before making the last decision. The final response is parsed by $\mathrm{ExtractChoice}(\cdot)$ to obtain the predicted option $\hat{y}$ from the candidate set $\mathcal{Y}$.

D Token-level Similarity

Similarity across degradation types. Figure 7 shows token-level similarity curves between clean and strongly degraded audio across six degradation types. SALMONN exhibits consistently high similarity for most tokens, with curves for different degradations quickly converging to near-identical values. This suggests that its projected tokens largely preserve high-level audio content while attenuating degradation-specific differences. In contrast, Audio Flamingo 3 shows lower and more variable similarity across token positions and degradation types, indicating that its representation retains more degradation-induced variation. However, the curves remain noisy and partially overlapping, suggesting that degradation categories are not clearly separated at the token level.

Similarity across degradation severity levels. Figures 8 and 9 compare token similarity across audible, medium, and strong severity levels for band-pass filtering and neural vocoder artifacts. For SALMONN, severity ordering appears only in the early tokens: stronger degradations yield lower similarity near the beginning of the sequence, but all severity levels quickly saturate toward high similarity. This suggests that severity information may be compressed or lost after projection, limiting the evidence available for downstream severity reasoning. Audio Flamingo 3 shows a clearer and more persistent monotonic trend, with stronger degradations generally corresponding to lower token similarity across much of the sequence. Nevertheless, the separation remains imperfect and varies across token positions, indicating that severity information is only partially and non-uniformly encoded.

Overall, the token-level analysis supports the failure patterns observed in MRMAD. Models with highly invariant projected tokens may lack sufficient low-level evidence for DTI and DSR, even when the text prompt explicitly asks for such judgments. Thus, improving degradation-aware LALMs requires audio encoders and projection modules that preserve fine-grained spectral, temporal, and degradation-specific cues rather than only semantic audio content.

E Response Examples

Tables 6–8 show representative examples of correct and incorrect model responses across the DTI, DSC, and DSR tasks. The correct examples suggest that models succeed when their intermediate

descriptions are aligned with the perceptual cues required by the task. For DTI, the model links the loss of high-frequency detail and muffled timbre to low-pass filtering. For DSC, it compares both clips along the same degradation dimension and recognizes that the second speech clip contains louder and more intrusive ambient noise. For DSR, it preserves severity information across three turns and correctly orders the clips according to the strength of the low-pass filtering artifact.

The incorrect examples reveal several limitations. In the DTI task, Qwen2-Audio loses task grounding and produces an irrelevant response, leading to an incorrect degradation-type prediction. In the DSC task, even the detailed reasoning produced by Qwen3-Omni-Thinking can disagree with the annotated severity, especially for complex neural-vocoder artifacts whose perceptual cues may be ambiguous. In the DSR task, Gemini 3.1 Pro can correctly describe individual clips but still sometimes fails to maintain a consistent global ranking or map the final ordering to the correct option. These results indicate that multi-round degradation reasoning requires not only artifact recognition, but also stable cross-turn memory, calibrated severity perception, and accurate option selection.

F Question & Prompt Template

To reduce prompt-specific bias and repeated wording patterns, we adopt diversified templates for question generation. This design is motivated by prior work showing that LLM behavior is sensitive to adversarial prompt variations, prompt wording, structured query forms, and structured reasoning formulations (Zhu et al., 2024; Xiong et al., 2024; Sclar et al., 2024; Zou et al., 2025; Liang et al., 2025; Hu et al., 2025; Zou et al., 2026; Xiong et al., 2026a,b; Li et al., 2026b). We therefore use task-specific prompt pools to encourage models to rely more on audio perception and task reasoning rather than superficial textual cues. The initial template pools are generated using ChatGPT-5.4 Thinking[4] and then manually reviewed and curated, following recent work that leverages LLMs and multimodal LLMs for task-specific annotation, data synthesis, and evaluation construction. (Long et al., 2024; Hou et al., 2024; Yang et al., 2025; Wang et al., 2025c; Fu et al., 2025; Peng et al., 2026).

The DTI templates are provided in Table 9, the DSC templates in Tables 10 and 11, and the DSR templates in Tables 12 and 13. For each instance, `{audio_type}` is determined by the audio domain, while `{degradation_label}` is determined by the audio's degradation type. For DTI, the final Round 1 text prompt string is directly sampled from the Round 1 text prompt pool, while the final Round 2 text prompt string is formed by concatenating a sampled Round 2 text prompt with a sampled answer constraint. The choices `{A}`–`{D}` contain the correct degradation type and three distractors sampled from degradation types outside the correct answer's perceptual group, with option orders randomized to avoid answer-position bias.

For DSC, the final Round 1 text prompt string is sampled from the Round 1 prompt pool, with the degradation type explicitly mentioned to guide the comparison. The final Round 2 text string concatenates a sampled Round 2 introduction prompt, a degradation-specific prompt, the fixed option format, and a sampled answer constraint. The comparison phrase `{cmp}` is randomly sampled from the higher- or lower-severity phrase pool, and the order of the two clips is randomized to avoid positional and answer-option bias. Since different degradations involve distinct perceptual cues, such as high-frequency attenuation for low-pass filtering, reverberant tails for reverberation, and temporal discontinuities for packet loss, we use diversified degradation descriptions instead of repeatedly using only `{degradation_label}`. This makes the comparison objective more precise and reduces ambiguity, while encouraging LALMs to reason about the specified audio degradation rather than relying on generic quality differences, repeated wording patterns, positional bias, or textual priors.

For DSR, we extend DSC from two to three clips. Round 1 and Round 2 prompts are sampled from their respective pools, with the degradation type specified. The Round 3 text prompt string concatenates an introduction prompt, a degradation-specific ranking prompt, a fixed multiple-choice option format, and an answer constraint. The paired ranking phrases `{low_word}` and `{high_word}` are randomly sampled according to the ranking direction, and the clip order is randomized to avoid positional and answer-option bias. Inspired by MuChoMusic (Weck et al., 2024), the answer space `{A}`–`{D}` contains the correct ranking, two harder adjacent-swap distractors, and one easier distractor sampled from the remaining incorrect permutations, enabling fine-grained severity reasoning beyond identifying the most or least degraded clip.

[4] https://chat.openai.com/

Algorithm 1 Two-turn inference template

Require: Turn-1 audio a_1, Turn-1 text prompt t_1
Require: Turn-2 audio a_2, Turn-2 text prompt t_2
Require: LALM $\mathcal{M}$, candidate option set $\mathcal{Y}$

Turn 1
Construct Turn-1 conversation:

$$\mathcal{C}_1 = \big[\mathrm{User}(a_1, t_1)\big].$$

Generate Turn-1 response:

$$r_1 = \mathcal{M}(\mathcal{C}_1).$$

Turn 2
Construct Turn-2 conversation by preserving Turn-1 context:

$$\mathcal{C}_2 = \big[\mathrm{User}(a_1, t_1), \mathrm{Assistant}(r_1), \mathrm{User}(a_2, t_2)\big].$$

Generate Turn-2 response:

$$r_2 = \mathcal{M}(\mathcal{C}_2).$$

Extract the predicted option:

$$\hat{y} = \mathrm{ExtractChoice}(r_2), \quad \hat{y} \in \mathcal{Y}.$$

return $\hat{y}, r_1, r_2$

Algorithm 2 Three-turn inference template

Require: Turn-1 audio a_1, Turn-1 text prompt t_1
Require: Turn-2 audio a_2, Turn-2 text prompt t_2
Require: Turn-3 audio a_3, Turn-3 text prompt t_3
Require: LALM $\mathcal{M}$, candidate option set $\mathcal{Y}$

Turn 1
Construct Turn-1 conversation:

$$\mathcal{C}_1 = \big[\mathrm{User}(a_1, t_1)\big].$$

Generate Turn-1 response:

$$r_1 = \mathcal{M}(\mathcal{C}_1).$$

Turn 2
Construct Turn-2 conversation by preserving Turn-1 context:

$$\mathcal{C}_2 = \big[\mathrm{User}(a_1, t_1), \mathrm{Assistant}(r_1), \mathrm{User}(a_2, t_2)\big].$$

Generate Turn-2 response:

$$r_2 = \mathcal{M}(\mathcal{C}_2).$$

Turn 3
Construct Turn-3 conversation by preserving previous contexts:

$$\mathcal{C}_3 = \big[\mathrm{User}(a_1, t_1), \mathrm{Assistant}(r_1), \mathrm{User}(a_2, t_2), \mathrm{Assistant}(r_2), \mathrm{User}(a_3, t_3)\big].$$

Generate Turn-3 response:

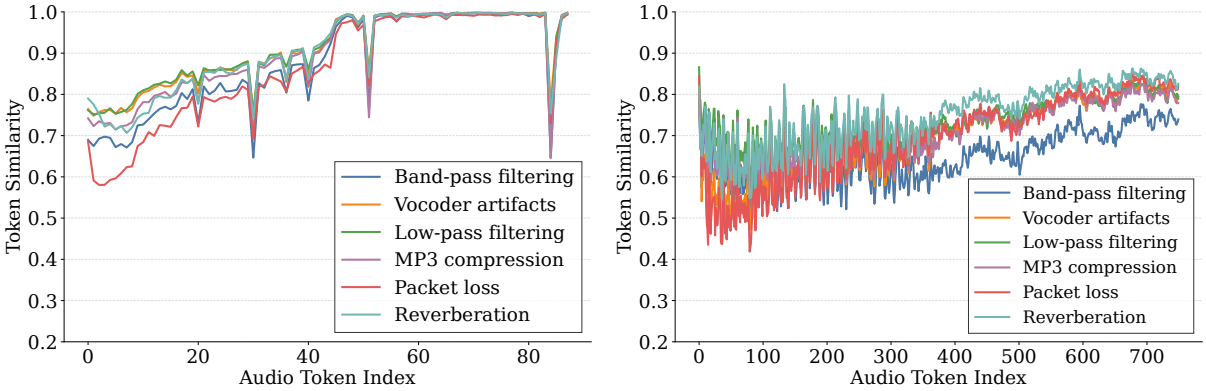
$$r_3 = \mathcal{M}(\mathcal{C}_3).$$

Extract the predicted option:

$$\hat{y} = \mathrm{ExtractChoice}(r_3), \quad \hat{y} \in \mathcal{Y}.$$

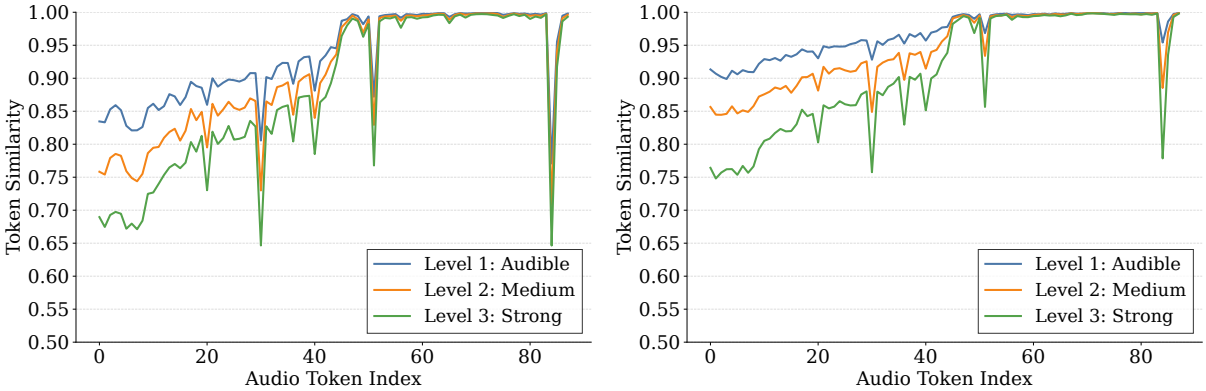
return $\hat{y}, r_1, r_2, r_3$

Figure 6: Two-turn and three-turn inference templates for multi-turn audio-language evaluation.



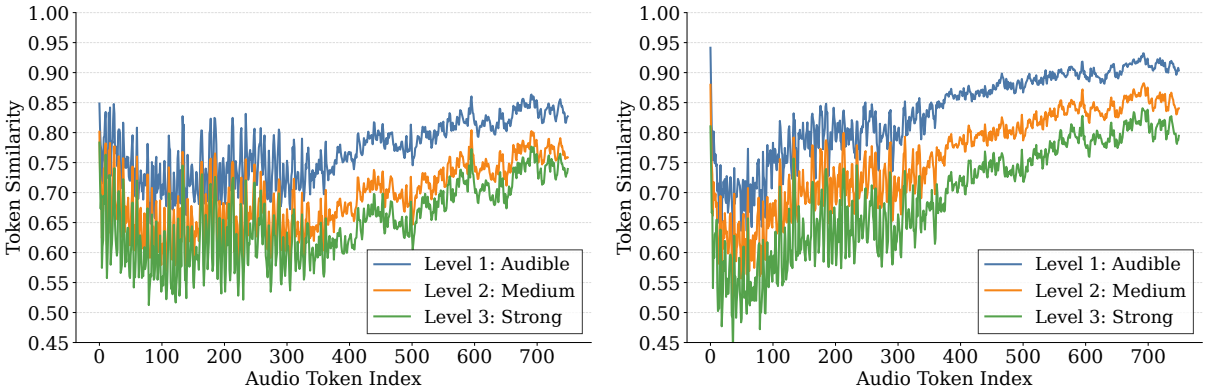
(a) SALMONN. (b) Audio Flamingo 3.

Figure 7: Projected audio-token similarity curves across six degradation types at the strong severity level.



(a) Band-pass filtering. (b) Neural vocoder artifacts.

Figure 8: Projected audio-token similarity curves across three severity levels for SALMONN.



(a) Band-pass filtering. (b) Neural vocoder artifacts.

Figure 9: Projected audio-token similarity curves across three severity levels for Audio Flamingo 3.

Correct Response	Wrong Response
Turn-1 Audio `Music/FMA-small/no_degrade.wav`	**Turn-1 Audio** `/Speech/VCTK/no_degrade.wav`
Turn-1 Text Prompt Please listen carefully to the first music clip and treat it as the reference recording. Pay attention to its clarity, timbre, and overall acoustic characteristics.	**Turn-1 Text Prompt** The first clip serves as the reference recording. Listen closely to its timbre, fidelity, and acoustic quality.
Turn-1 Output The music in the audio is instrumental with a fast tempo, synthesiser articulation, electronic arrangement, and video game sound. It has a loud, resounding, boomy, deep, and heavy vibe with an energetic and youthful aura. The music belongs to the Synthpop/Technopop genre.	**Turn-1 Output** I'm sorry, but I cannot provide answers on political matters. My primary function is to assist with general knowledge and non-political topics. If you have any other inquiries, please feel free to ask.
Turn-2 Audio `/Music/FMA-small/lowpass_level_3.wav`	**Turn-2 Audio** `/Speech/VCTK/ambient_noise_level_3.wav`
Turn-2 Text Prompt Please compare the second audio with the first reference audio. Which option best matches the most noticeable degradation in the second clip?	**Turn-2 Text Prompt** Please listen to the second speech clip. Relative to the first reference clip, which degradation best describes the main artifact you hear in the second clip?
Please choose one of the following options: A. reverberation B. ambient noise C. low-pass filtering D. packet loss	Please choose one of the following options: A. ambient noise B. low-pass filtering C. HiFi-GAN artifacts D. echo
Your answer must be A, B, C, or D only.	Output only a single letter: A, B, C, or D.
Turn-2 Output: C	**Turn-2 Output: D**
Ground Truth: C	**Ground Truth: A**

Table 6: Examples of correct and incorrect responses from Qwen2-Audio on the DTI task.

Correct Response	Wrong Response
Turn-1 Audio	**Turn-1 Audio**
`Speech/LibriSpeech/ambient_noise_level_1.wav`	`Sound/FSD50K/hifigan_level_1.flac`
Turn-1 Text Prompt	**Turn-1 Text Prompt**
Use the first speech clip as the first comparison sample and listen carefully to the ambient noise artifact.	Please listen to the first clip carefully. It is processed with a neural vocoder introducing vocoder reconstruction artifacts.
Turn-1 Output	**Turn-1 Output**
<think> Got it, let's tackle the problem. First, the user wants me to identify the ambient noise artifact in the first speech clip. Then, I analyzed the audio by listening carefully to the first clip multiple times to isolate background sounds from the speech; during the first pass, I heard a female voice speaking with a distinct low-frequency hum or drone present throughout the recording, and during the second pass focusing on the hum, I confirmed it was a continuous low-pitched sound resembling an electrical hum from equipment or room resonance rather than intermittent noise like wind or traffic, with the hum most noticeable during pauses between phrases such as after 'seats' and before 'at the breakfast table'. Next, my reasoning is that this hum is the primary ambient noise artifact because it's a consistent low-frequency drone that persists across the entire clip without interruption, distinguishing it from transient background sounds. Finally, the ambient noise artifact is a low-frequency hum or drone. </think> The ambient noise artifact in the first speech clip is a low-frequency hum or drone.	<think> Got it, let's tackle the problem. First, the user wants me to listen to the first audio clip and describe the artifacts introduced by a neural vocoder. Then, the audio starts with a high-frequency, sibilant, almost 'shhh' sound that is very harsh and unnatural, sounding like a digital artifact rather than a natural sound; it has a 'crackling' or 'sputtering' quality, especially noticeable in the higher frequencies, and the sound is not smooth but has a 'gritty' texture ... Finally, the primary artifact is a harsh, high-frequency sibilant noise that sounds like crackling or sputtering, which is a common artifact introduced by neural vocoders during the reconstruction process, particularly affecting the high-frequency components of the audio signal, and it's often described as 'stuttering' or 'aliasing' noise. </think> The audio clip exhibits a prominent **harsh, high-frequency sibilant noise** that sounds like crackling or sputtering ... This artifact occurs when the vocoder struggles to accurately synthesize the fine details of the original speech signal's high-frequency components.
Turn-2 Audio	**Turn-2 Audio**
`Speech/LibriSpeech/ambient_noise_level_3.wav`	`Sound/FSD50K/hifigan_level_3.wav`
Turn-2 Text Prompt	**Turn-2 Text Prompt**
Now listen to the second audio example. In both clips, ambient noise is present. Which clip exhibits higher severity ambient noise contamination? Please choose the best answer from the following options: A. The first clip. B. The second clip. Please reply using only A or B.	Continue by listening to the second sound event recording. Both clips are affected by vocoder reconstruction errors. Which sounds less distorted due to vocoder reconstruction? Please choose the best answer from the following options: A. The first clip. B. The second clip. Your answer must be A or B only.
Turn-2 Output	**Turn-2 Output**
<think> Got it, let's tackle the problem. First, the user wants me to compare two audio clips to determine which has higher severity ambient noise contamination, with the first clip already analyzed as having a low-frequency hum or drone. Then, the second audio clip contains the same speech but features significantly more prominent ambient noise: a loud, continuous, and complex background sound resembling a busy public space like a train station or airport terminal, with distinct mechanical noises and indistinct chatter that partially obscures the speech. Next, comparing the two, the first clip's noise is a subtle low-frequency hum while the second clip's noise is loud, multi-layered, and actively interferes with speech clarity, making it substantially more severe. Finally, the second clip exhibits higher severity ambient noise contamination, so the answer is B. </think> **Filtered Answer: B**	<think> Got it, let's tackle the problem. First, the user wants me to compare two audio clips, both processed with a neural vocoder and exhibiting reconstruction artifacts, and determine which one sounds less distorted. Then, analyzing the first clip: it is a recording of a sibilant sound (like "sh" or "s"), and the distortion is very prominent as a harsh, crackling, and stuttering noise that is almost as loud as the intended speech sound itself, making it very unnatural and difficult to understand the underlying phoneme. ... , so it sounds more like a natural sibilant with some background noise rather than a heavily corrupted signal. Then, comparing the two clips: the first clip has severe, dominant artifacts that make it sound very unnatural and distorted, while the second clip has noticeable artifacts but they are less severe and less dominant, so the second clip is less distorted. Finally, the answer is B. </think> **Filtered Answer: B**
Ground Truth: B	**Ground Truth: A**

Table 7: Examples of correct and incorrect responses from Qwen3-Omni-Thinking on the DSC task.

Correct Response	Wrong Response
Turn-1 Audio	**Turn-1 Audio**
`Sound/FSD50K/lowpass_level_3.flac`	`Music/FMA-small/mp3_level_1.flac`
Turn-1 Text Prompt	**Turn-1 Text Prompt**
Please listen to the first recording carefully and pay attention to the perceived effect of the low-pass filtering artifact.	Listen to the first audio example carefully and pay attention to the MP3 compression artifact.
Turn-1 Output	**Turn-1 Output**
The recording consists of heavy, rhythmic breathing sounds. The low-pass filtering is severe and highly intrusive, removing most high-frequency information and making the audio sound extremely dull, dark, and muffled. The artifact dominates the recording and substantially reduces naturalness and clarity.	The audio is an electronic music track with noticeable MP3 compression artifacts. The model describes a swirly, watery, and slightly metallic phasing sound in the high frequencies, especially on synthesizers and hi-hats, along with muffling, reduced crispness, and smeared transients.
Turn-2 Audio	**Turn-2 Audio**
`Sound/FSD50K/lowpass_level_1.flac`	`Music/FMA-small/mp3_level_3.flac`
Turn-2 Text Prompt	**Turn-2 Text Prompt**
Now listen carefully to the second recording and pay attention to the low-pass filtering artifact.	Please listen to the second clip carefully. It is compressed using MP3.
Turn-2 Output	**Turn-2 Output**
The second recording contains the same heavy, rhythmic breathing sounds, but the low-pass filtering is absent or much less noticeable. The high-frequency hiss, friction, and airiness of the breaths are restored, making the audio clearer and more natural than the first clip.	The second clip has much more severe MP3 compression artifacts than the first. The model describes an extreme underwater or swirling effect, severe high-frequency roll-off, transient smearing, mushiness, and metallic chirping or ringing, indicating that the original musical information is heavily damaged.
Turn-3 Audio	**Turn-3 Audio**
`Sound/FSD50K/lowpass_level_2.flac`	`Music/FMA-small/mp3_level_2.flac`
Turn-3 Text Prompt	**Turn-3 Text Prompt**
Continue by listening to the third music clip. Order the three clips from more distorted by MP3 compression to less distorted by MP3 compression. Choose the best answer from A, B, C, or D. A. the third clip > the second clip > the first clip B. the first clip > the second clip > the third clip C. the second clip > the first clip > the third clip D. the third clip > the first clip > the second clip Output only a single letter: A, B, C, or D.	Next, listen to the third sound event recording. Rank the three clips based on low-pass filtering from more severe degradation to less severe degradation. Choose the best answer from A, B, C, or D. A. the third clip > the first clip > the second clip B. the first clip > the second clip > the third clip C. the first clip > the third clip > the second clip D. the second clip > the first clip > the third clip Your answer must be A, B, C, or D only.
Turn-3 Output: C	**Turn-3 Output: C**
Ground Truth: C	**Ground Truth: D**

Table 8: Examples of correct and incorrect responses from Gemini 3.1 Pro on the DSR task.

	{audio_type}
1	**Speech:** speech clip.
2	**Music:** music clip.
3	**Sound:** sound event recording.

	{degradation_label}
1	Ambient noise.
2	Low-pass filtering.
3	Band-pass filtering.
4	Reverberation.
5	Echo.
6	MP3 compression.
7	Neural vocoder artifacts.
8	DSP denoising artifacts.
9	Packet loss.

	Round 1 Text Prompt
1	Please listen carefully to the first {audio_type} and treat it as the reference recording. Pay attention to its clarity, timbre, and overall acoustic characteristics.
2	Listen carefully to the first {audio_type} and use it as the clean reference. Focus on its overall quality, background condition, and spectral characteristics.
3	Please listen to the first {audio_type} and treat it as the clean reference recording. Focus on its naturalness and overall sound quality.
4	Carefully listen to the first {audio_type} and use it as the baseline for comparison. Pay attention to the audio quality and acoustic details.
5	The first {audio_type} is the reference example. Please listen to it carefully and note its clarity, tonal balance, and background condition.
6	Please use the first {audio_type} as the clean reference. Listen carefully to its overall fidelity, quality, and acoustic texture.
7	Listen closely to the first recording and treat it as the reference signal. Pay attention to its naturalness, clarity, and background environment.
8	Please listen to the first clip as the reference audio. Focus on its timbre, sharpness, and overall listening quality.
9	Use the first {audio_type} as the baseline. Listen carefully to its tonal characteristics, perceptual quality, and background cleanliness.
10	Please listen to the first recording carefully and consider it the reference condition. Pay attention to its overall audio quality and naturalness.
11	Treat the first {audio_type} as the original reference. Listen carefully and note its clarity, spectral balance, and ambient condition.
12	Please listen carefully to the first sample and use it as the comparison reference. Focus on its clean acoustic characteristics and overall quality.
13	The first clip serves as the reference recording. Listen closely to its timbre, fidelity, and acoustic quality.
14	Listen to the first audio example and treat it as the clean baseline. Pay attention to its natural quality and acoustic details.
15	Please use the first clip as the reference audio. Listen carefully to its clarity, quality, and overall perceptual characteristics.

	Round 2 Text Prompt
1	Now listen to the second {audio_type}. Compared with the first audio clip, what is the most prominent audible degradation or artifact in the second clip? Please choose the correct answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.
2	Please listen to the second {audio_type}. Relative to the first reference clip, which degradation best describes the main artifact you hear in the second clip? Please choose one of the following options: A. {A}, B. {B}, C. {C}, D. {D}.
3	Now compare the second clip against the first reference clip. What is the dominant perceptual degradation present in the second {audio_type}? Please choose the best answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.
4	After listening to both clips, identify the primary artifact introduced in the second clip compared with the first one. Please select one of the following answers: A. {A}, B. {B}, C. {C}, D. {D}.
5	Using the first clip as a clean reference, determine which audible degradation is most evident in the second {audio_type}. Please choose the best answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.
6	Please compare the second audio with the first reference audio. Which option best matches the most noticeable degradation in the second clip? Please choose one of the following options: A. {A}, B. {B}, C. {C}, D. {D}.
7	Listen to the second clip and compare it with the first clip. What type of artifact is most clearly present in the second audio? Please select one of the following options: A. {A}, B. {B}, C. {C}, D. {D}.
8	Compared with the first reference recording, what is the main audible impairment in the second recording? Please choose the best answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.
9	Based on your comparison between the first and second clips, which degradation category best describes the second clip? Please select one of the following options: A. {A}, B. {B}, C. {C}, D. {D}.
10	The second clip contains a noticeable artifact relative to the first clip. What is the most appropriate label for that degradation? Please choose one of the following options: A. {A}, B. {B}, C. {C}, D. {D}.
11	Please identify the dominant degradation in the second audio when compared with the first reference audio. Please select the correct answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.
12	Using the first clip as the baseline, what is the main perceptual artifact present in the second clip? Please choose one of the following answers: A. {A}, B. {B}, C. {C}, D. {D}.
13	Now evaluate the second audio relative to the first. Which of the following best describes the most prominent degradation you hear? Please choose the best answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.
14	Compare the second clip to the first reference clip and determine the primary audible distortion in the second clip. Please select one of the following options: A. {A}, B. {B}, C. {C}, D. {D}.
15	Please judge the second clip against the first clean reference. What is the principal degradation or artifact in the second clip? Please choose the best answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.

	Answer Constraints
1	Answer with only A, B, C, or D.
2	Respond with only A, B, C, or D.
3	Your answer must be A, B, C, or D only.
4	Please reply using only A, B, C, or D.
5	Output only a single letter: A, B, C, or D.
6	Return only A, B, C, or D as your answer.

Table 9: DTI prompt templates. For each DTI question, we randomly sample one Round 1 prompt, one Round 2 prompt, and one answer constraint. The placeholder {audio_type} is replaced with the domain-specific phrase.

	{audio_type}
1	**Speech:** speech clip.
2	**Music:** music clip.
3	**Sound:** sound event recording.
	{degradation_label}
1	Ambient noise.
2	Low-pass filtering.
3	Band-pass filtering.
4	Reverberation.
5	Echo.
6	MP3 compression.
7	Neural vocoder artifacts.
8	DSP denoising artifacts.
9	Packet loss.
	{be_verb}
1	has
2	includes
3	contains
4	exhibits
5	presents
6	is affected by
	Round 1 Text Prompt
1	Please listen carefully to the first {audio_type}. It {be_verb} {degradation_label}.
2	Listen carefully to the first {audio_type}. It {be_verb} {degradation_label}.
3	Please listen to the first {audio_type} and pay attention to the {degradation_label}.
4	Carefully listen to the first {audio_type}. It {be_verb} {degradation_label}.
5	The first {audio_type} will serve as the first comparison sample. It {be_verb} {degradation_label}. Please listen carefully.
6	Please use the first {audio_type} as the first comparison clip and listen closely to the {degradation_label}.
7	Listen closely to the first recording and note the perceptual effect of the {degradation_label}.
8	Please listen to the first clip carefully. It {be_verb} {degradation_label}.
9	Use the first {audio_type} as the first comparison sample and listen carefully to the {degradation_label}.
10	Please listen to the first recording carefully and pay attention to the perceived effect of the {degradation_label}.
11	Treat the first {audio_type} as the first comparison sample and listen carefully to the {degradation_label}.
12	Please listen carefully to the first sample and note the perceptual characteristics of the {degradation_label}.
13	The first clip serves as the first comparison recording. It {be_verb} {degradation_label}. Listen closely.
14	Listen to the first audio example carefully and pay attention to the {degradation_label}.
15	Please use the first clip as the first comparison audio and listen carefully to the {degradation_label}.
	Round 2 Introduction Prompt
1	Now listen to the second {audio_type}.
2	Please now listen carefully to the second {audio_type}.
3	Next, listen to the second {audio_type}.
4	Now listen to the second comparison {audio_type}.
5	Please listen to the second clip carefully.
6	Continue by listening to the second {audio_type}.
7	Now listen carefully to the second recording.
8	Please proceed to the second {audio_type} and listen carefully.
9	Next, listen carefully to the second comparison clip.
10	Now listen to the second audio example.
	Round 2 Option Format
1	Please choose the best answer from the following options: A. The first clip; B. The second clip.
	Answer Constraints
1	Answer with only A or B.
2	Respond with only A or B.
3	Your answer must be A or B only.
4	Please reply using only A or B.
5	Output only a single letter: A or B.
6	Return only A or B as your answer.

Table 10: DSC general prompt templates. For each instance, {audio_type} and {degradation_label} are determined by the corresponding audio, while the remaining prompt components are sampled per question.

ID	Comparison Phrases: {cmp} (Higher Severity)	Comparison Phrases: {cmp} (Lower Severity)
1	higher severity	lower severity
2	greater intensity	lower intensity
3	higher degree	lower degree
4	more severe degradation	less severe degradation
5	higher artifact level	lower artifact level
	Round 2 Text Prompt (Ambient Noise)	
1	Both clips contain ambient noise. Which clip contains more audible ambient noise?	Both clips contain ambient noise. Which clip contains less audible ambient noise?
2	The two clips are both affected by ambient noise. Which clip is more affected by ambient noise?	The two clips are both affected by ambient noise. Which clip is less affected by ambient noise?
3	Each clip is mixed with ambient noise. Which clip has {cmp} of ambient noise?	Each clip is mixed with ambient noise. Which clip has {cmp} of ambient noise?
4	In both clips, ambient noise is present. Which clip exhibits {cmp} ambient noise contamination?	In both clips, ambient noise is present. Which clip exhibits {cmp} ambient noise contamination?
	Low-pass Filtering	
1	Both clips undergo low-pass filtering. Which clip is more band-limited?	Both clips undergo low-pass filtering. Which clip is less band-limited?
2	The two clips are both low-pass filtered. Which clip shows a greater loss of high-frequency content?	The two clips are both low-pass filtered. Which clip preserves more high-frequency content?
3	Each clip is processed with low-pass filtering. Which clip has {cmp} low-pass filtering?	Each clip is processed with low-pass filtering. Which clip has {cmp} low-pass filtering?
4	In both clips, high-frequency content is attenuated by low-pass filtering. Which clip exhibits {cmp} attenuation of high-frequency content?	In both clips, high-frequency content is attenuated by low-pass filtering. Which clip exhibits {cmp} attenuation of high-frequency content?
	Band-pass Filtering	
1	Both clips undergo band-pass filtering. Which clip is more band-limited?	Both clips undergo band-pass filtering. Which clip is less band-limited?
2	The two clips are both band-pass filtered. Which clip shows more restricted frequency content?	The two clips are both band-pass filtered. Which clip preserves a broader frequency range?
3	Each clip is processed with band-pass filtering. Which clip has {cmp} band-pass filtering?	Each clip is processed with band-pass filtering. Which clip has {cmp} band-pass filtering?
4	In both clips, the spectrum is constrained by band-pass filtering. Which clip exhibits {cmp} spectral restriction?	In both clips, the spectrum is constrained by band-pass filtering. Which clip exhibits {cmp} spectral restriction?
	Reverberation	
1	Both clips are affected by reverberation. Which clip sounds more reverberant?	Both clips are affected by reverberation. Which clip sounds less reverberant?
2	The two clips both exhibit reverberation. Which clip has a more noticeable reverberation effect?	The two clips both exhibit reverberation. Which clip has a less noticeable reverberation effect?
3	Each clip contains reverberation. Which clip has {cmp} reverberation?	Each clip contains reverberation. Which clip has {cmp} reverberation?
4	In both clips, reverberant energy is present. Which clip exhibits {cmp} reverberant tail?	In both clips, reverberant energy is present. Which clip exhibits {cmp} reverberant tail?
	Echo	
1	Both clips are mixed with echo. Which clip has more prominent echo?	Both clips are mixed with echo. Which clip has less prominent echo?
2	The two clips both include delayed repetitions. Which clip contains more noticeable delayed repetitions?	The two clips both include delayed repetitions. Which clip contains fewer noticeable delayed repetitions?
3	Each clip contains an echo effect. Which clip has {cmp} echo effect?	Each clip contains an echo effect. Which clip has {cmp} echo effect?
4	In both clips, echo is present. Which clip exhibits {cmp} audibility of the delayed signal?	In both clips, echo is present. Which clip exhibits {cmp} audibility of the delayed signal?
	MP3 Compression	
1	Both clips are compressed using MP3. Which clip has more compression artifacts?	Both clips are compressed using MP3. Which clip has fewer compression artifacts?
2	The two clips are both affected by MP3 compression. Which clip sounds more distorted due to compression?	The two clips are both affected by MP3 compression. Which clip sounds less distorted due to compression?
3	Each clip contains MP3 compression artifacts. Which clip has {cmp} MP3 compression artifacts?	Each clip contains MP3 compression artifacts. Which clip has {cmp} MP3 compression artifacts?
4	In both clips, codec artifacts are present. Which clip exhibits {cmp} codec distortion?	In both clips, codec artifacts are present. Which clip exhibits {cmp} codec distortion?
	Neural Vocoder Artifacts	
1	Both clips contain neural vocoder reconstruction artifacts. Which clip has more vocoder reconstruction artifacts?	Both clips contain neural vocoder reconstruction artifacts. Which clip has fewer vocoder reconstruction artifacts?
2	The two clips are both affected by vocoder reconstruction errors. Which clip sounds more distorted due to vocoder reconstruction?	The two clips are both affected by vocoder reconstruction errors. Which clip sounds less distorted due to vocoder reconstruction?
3	Each clip includes artifacts from neural vocoder reconstruction. Which clip has {cmp} vocoder reconstruction artifacts?	Each clip includes artifacts from neural vocoder reconstruction. Which clip has {cmp} vocoder reconstruction artifacts?
4	In both clips, neural vocoder reconstruction artifacts are present. Which clip exhibits {cmp} vocoder reconstruction distortion?	In both clips, neural vocoder reconstruction artifacts are present. Which clip exhibits {cmp} vocoder reconstruction distortion?
	DSP Denoising Artifacts	
1	Both clips contain DSP denoising artifacts. Which clip has more oversuppression artifacts?	Both clips contain DSP denoising artifacts. Which clip has fewer oversuppression artifacts?
2	The two clips both exhibit denoising-related distortion. Which clip is more degraded due to oversuppression?	The two clips both exhibit denoising-related distortion. Which clip is less degraded due to oversuppression?
3	Each clip is affected by DSP denoising artifacts. Which clip has {cmp} oversuppression?	Each clip is affected by DSP denoising artifacts. Which clip has {cmp} oversuppression?
4	In both clips, artifacts from denoising are present. Which clip exhibits {cmp} oversuppression caused by denoising algorithm processing?	In both clips, artifacts from denoising are present. Which clip exhibits {cmp} oversuppression caused by denoising algorithm processing?
	Packet Loss	
1	Both clips are affected by packet loss. Which clip has more audible glitches or dropouts?	Both clips are affected by packet loss. Which clip has fewer audible glitches or dropouts?
2	The two clips are both degraded by packet loss. Which clip contains more dropouts or missing segments?	The two clips are both degraded by packet loss. Which clip contains fewer dropouts or missing segments?
3	Each clip contains packet loss artifacts. Which clip has {cmp} packet loss artifacts, such as glitches or dropouts?	Each clip contains packet loss artifacts. Which clip has {cmp} packet loss artifacts, such as glitches or dropouts?
4	In both clips, packet loss introduces temporal discontinuities. Which clip exhibits {cmp} temporal discontinuities caused by packet loss?	In both clips, packet loss introduces temporal discontinuities. Which clip exhibits {cmp} temporal discontinuities caused by packet loss?

Table 11: DSC Round 2 comparison phrases and degradation-specific text prompt templates.

	{audio_type}
1	**Speech:** speech clip.
2	**Music:** music clip.
3	**Sound:** sound event recording.

	{degradation_label}
1	Ambient noise.
2	Low-pass filtering.
3	Band-pass filtering.
4	Reverberation.
5	Echo.
6	MP3 compression.
7	Neural vocoder artifacts.
8	DSP denoising artifacts.
9	Packet loss.

	{be_verb}
1	has
2	includes
3	contains
4	exhibits
5	presents
6	is affected by

	Round 1 Text Prompt
1	Please listen carefully to the first {audio_type}. It {be_verb} {degradation_label}.
2	Listen carefully to the first {audio_type}. It {be_verb} {degradation_label}.
3	Please listen to the first {audio_type} and pay attention to the {degradation_label}.
4	Carefully listen to the first {audio_type} sample. It {be_verb} {degradation_label}.
5	The first {audio_type} will serve as the first comparison sample. It {be_verb} {degradation_label}. Please listen carefully.
6	Please use the first {audio_type} as the first comparison clip and listen closely to the {degradation_label}.
7	Listen closely to the first recording and note the perceptual effect of the {degradation_label}.
8	Please listen to the first clip carefully. It {be_verb} {degradation_label}.
9	Use the first {audio_type} as the first comparison sample and listen carefully to the {degradation_label}.
10	Please listen to the first recording carefully and pay attention to the perceived effect of the {degradation_label}.
11	Treat the first {audio_type} as the first comparison sample and listen carefully to the {degradation_label}.
12	Please listen carefully to the first sample and note the perceptual characteristics of the {degradation_label}.
13	The first clip serves as the first comparison recording. It {be_verb} {degradation_label}. Listen closely.
14	Listen to the first audio example carefully and pay attention to the {degradation_label}.
15	Please use the first clip as the first comparison audio and listen carefully to the {degradation_label}.

	Round 2 Text Prompt
1	Now listen to the second {audio_type}. It {be_verb} {degradation_label}.
2	Please now listen carefully to the second {audio_type}. It {be_verb} {degradation_label}.
3	Next, listen to the second {audio_type} and pay attention to the {degradation_label}.
4	Now listen to the second comparison {audio_type}. It {be_verb} {degradation_label}.
5	Please listen to the second clip carefully. It {be_verb} {degradation_label}.
6	Continue by listening to the second {audio_type} and note the perceptual effect of the {degradation_label}.
7	Now listen carefully to the second recording and pay attention to the {degradation_label}.
8	Please proceed to the second {audio_type} and listen carefully to the {degradation_label}.
9	Next, listen carefully to the second comparison clip. It {be_verb} {degradation_label}.
10	Now listen to the second audio example and pay attention to the {degradation_label}.
11	Please listen to the second sample carefully and note the perceptual characteristics of the {degradation_label}.
12	The second clip serves as the next comparison recording. It {be_verb} {degradation_label}. Listen closely.
13	Use the second {audio_type} as the next comparison sample and listen carefully to the {degradation_label}.
14	Please listen to the second recording carefully and note the perceived effect of the {degradation_label}.
15	Now listen to the second clip and remember the perceptual effect of the {degradation_label} for later comparison.

	Round 3 Introduction Prompt
1	Now listen to the third {audio_type}.
2	Please now listen carefully to the third {audio_type}.
3	Next, listen to the third {audio_type}.
4	Now listen to the third comparison {audio_type}.
5	Please listen to the third clip carefully.
6	Continue by listening to the third {audio_type}.
7	Now listen carefully to the third recording.
8	Please proceed to the third {audio_type} and listen carefully.
9	Next, listen carefully to the third comparison clip.
10	Now listen to the third audio example.

	Round 3 Option Format
1	Please choose the best answer from the following options: A. {A}, B. {B}, C. {C}, D. {D}.

	Answer Constraints
1	Answer with only A, B, C, or D.
2	Respond with only A, B, C, or D.
3	Your answer must be A, B, C, or D only.
4	Please reply using only A, B, C, or D.
5	Output only a single letter: A, B, C, or D.
6	Return only A, B, C, or D as your answer.

Table 12: DSR general prompt templates. For each DSR instance, {audio_type} and {degradation_label} are determined by the corresponding audio, while the remaining prompt components are sampled per question.

ID	Ranking Directions: {low_word} → {high_word}	Ranking Directions: {high_word} → {low_word}
1	lower severity → higher severity	higher severity → lower severity
2	lower intensity → greater intensity	greater intensity → lower intensity
3	lower degree → higher degree	higher degree → lower degree
4	less severe degradation → less severe degradation	more severe degradation → less severe degradation
5	lower artifact level → higher artifact level	higher artifact level → lower artifact level
	Round 3 Text Prompt (Ambient Noise)	
1	All three clips contain ambient noise. Rank them from {low_word} to {high_word}.	All three clips contain ambient noise. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on ambient noise from {low_word} to {high_word}.	Rank the three clips based on ambient noise from {high_word} to {low_word}.
3	Order the three clips from less audible ambient noise to more audible ambient noise.	Order the three clips from more audible ambient noise to less audible ambient noise.
4	Determine which ordering reflects ambient noise becoming progressively more noticeable.	Determine which ordering reflects ambient noise becoming progressively less noticeable.
	Low-pass Filtering	
1	All three clips are processed with low-pass filtering. Rank them from {low_word} to {high_word}.	All three clips are processed with low-pass filtering. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on low-pass filtering from {low_word} to {high_word}.	Rank the three clips based on low-pass filtering from {high_word} to {low_word}.
3	Order the three clips from less muffled to more muffled.	Order the three clips from more muffled to less muffled.
4	Determine which ordering reflects increasing loss of high-frequency content.	Determine which ordering reflects decreasing loss of high-frequency content.
	Band-pass Filtering	
1	All three clips are processed with band-pass filtering. Rank them from {low_word} to {high_word} in terms of band-limited severity.	All three clips are processed with band-pass filtering. Rank them from {high_word} to {low_word} in terms of band-limited severity.
2	All three clips are affected by band-pass filtering. Rank the three clips in terms of band-limited severity from {low_word} to {high_word}.	All three clips are affected by band-pass filtering. Rank the three clips in terms of band-limited severity from {high_word} to {low_word}.
3	All three clips contain artifacts caused by band-pass filtering. Order the three clips from less band-limited to more band-limited.	All three clips contain artifacts caused by band-pass filtering. Order the three clips from more band-limited to less band-limited.
4	All three clips exhibit band-limiting effects due to band-pass filtering. Determine which ordering reflects increasing band-limiting.	All three clips exhibit band-limiting effects due to band-pass filtering. Determine which ordering reflects decreasing band-limiting.
	Reverberation	
1	All three clips are affected by reverberation. Rank them from {low_word} to {high_word}.	All three clips are affected by reverberation. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on reverberation from {low_word} to {high_word}.	Rank the three clips based on reverberation from {high_word} to {low_word}.
3	Order the three clips from less reverberant to more reverberant.	Order the three clips from more reverberant to less reverberant.
4	Determine which ordering reflects increasingly strong reverberant tails.	Determine which ordering reflects progressively weaker reverberant tails.
	Echo	
1	All three clips are mixed with echo. Rank them from {low_word} to {high_word}.	All three clips are mixed with echo. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on echo from {low_word} to {high_word}.	Rank the three clips based on echo from {high_word} to {low_word}.
3	Order the three clips from less prominent echo to more prominent echo.	Order the three clips from more prominent echo to less prominent echo.
4	Determine which ordering reflects increasingly noticeable delayed repetitions.	Determine which ordering reflects progressively weaker delayed repetitions.
	MP3 Compression	
1	All three clips are compressed using MP3. Rank them from {low_word} to {high_word}.	All three clips are compressed using MP3. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on MP3 compression artifacts from {low_word} to {high_word}.	Rank the three clips based on MP3 compression artifacts from {high_word} to {low_word}.
3	Order the three clips from less distorted by MP3 compression to more distorted by MP3 compression.	Order the three clips from more distorted by MP3 compression to less distorted by MP3 compression.
4	Determine which ordering reflects increasingly audible codec artifacts.	Determine which ordering reflects decreasing audibility of codec artifacts.
	Neural Vocoder Artifacts	
1	All three clips contain neural vocoder artifacts. Rank them from {low_word} to {high_word}.	All three clips contain neural vocoder artifacts. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on neural vocoder artifacts from {low_word} to {high_word}.	Rank the three clips based on neural vocoder artifacts from {high_word} to {low_word}.
3	Order the three clips from fewer vocoder artifacts to more vocoder artifacts.	Order the three clips from more vocoder artifacts to fewer vocoder artifacts.
4	Determine which ordering reflects increasingly noticeable vocoder reconstruction distortion.	Determine which ordering reflects decreasing vocoder reconstruction distortion.
	DSP Denoising Artifacts	
1	All three clips contain DSP denoising artifacts. Rank them from {low_word} oversuppression to {high_word} oversuppression.	All three clips contain DSP denoising artifacts. Rank them from {high_word} oversuppression to {low_word} oversuppression.
2	Rank the three clips based on DSP denoising artifacts from {low_word} oversuppression to {high_word} oversuppression.	Rank the three clips based on DSP denoising artifacts from {high_word} oversuppression to {low_word} oversuppression.
3	Order the three clips from less oversuppression to more oversuppression.	Order the three clips from more oversuppression to less oversuppression.
4	Determine which ordering reflects increasing oversuppression artifacts caused by denoising algorithm processing.	Determine which ordering reflects decreasing oversuppression artifacts caused by denoising algorithm processing.
	Packet Loss	
1	All three clips are affected by packet loss. Rank them from {low_word} to {high_word}.	All three clips are affected by packet loss. Rank them from {high_word} to {low_word}.
2	Rank the three clips based on packet loss from {low_word} to {high_word}.	Rank the three clips based on packet loss from {high_word} to {low_word}.
3	Order the three clips from fewer audible glitches and dropouts to more audible glitches and dropouts.	Order the three clips from more audible glitches and dropouts to fewer audible glitches and dropouts.
4	Determine which ordering reflects increasingly severe temporal discontinuities.	Determine which ordering reflects decreasing temporal discontinuities.

Table 13: DSR Round 3 ranking phrases and degradation-specific text prompt templates.