A Benchmark for Audio Reasoning Capabilities of Multimodal Large Language Models

Iwona Christop[1], Mateusz Czyżnikiewicz[2], Paweł Skórzewski[1], Łukasz Bondaruk[2], Jakub Kubiak[2], Marcin Lewandowski[2], Marek Kubis[1]

[1]Adam Mickiewicz University, ul. Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
[2]Samsung R&D Institute Poland, Plac Europejski 1, 00-844 Warszawa, Poland

Abstract

The present benchmarks for testing the audio modality of multimodal large language models concentrate on testing various audio tasks such as speaker diarization or gender identification in isolation. Whether a multimodal model can answer the questions that require reasoning skills to combine audio tasks of different categories, cannot be verified with their use. To address this issue, we propose Audio Reasoning Tasks (ART), a new benchmark for assessing the ability of multimodal models to solve problems that require reasoning over audio signal.

1 Introduction

Multimodal large language models (MLLMs) expand upon conventional language models by providing capabilities to process visual and audio data. While there exist elaborate benchmarks that require advanced reasoning skills to solve problems grounded in vision (Lee et al., 2024), the prevalent paradigm for evaluating auditory modality of MLLMs relies on testing selected capabilities in isolation. Unfortunately, this approach does not guarantee that problems that require reasoning across different categories of audio tasks are solved with satisfactory performance, even if superhuman performance of the model is reported for separate tasks such as speech recognition, acoustic scene classification or audio captioning. This issue is of particular concern considering that a common practice for building MLLMs involves using textual and audio components pre-trained separately to develop the final model (e.g., Grattafiori et al., 2024; Fixie.ai, 2025).

To address deficiencies of current approaches proposed for evaluating auditory modality of MLLMs, we introduce Audio Reasoning Tasks (ART), a new benchmark that comprises a set of tasks designed to assess MMLMs' ability to solve problems that require combining diverse skills in understanding audio signals with the ability to reason over their combination. The benchmark is designed to be comprehensible by a person without hearing impairment. The evaluation results reported in Section 4 show that although the proposed tasks can be easily solved by a human, they present a challenge for multimodal large language models.

2 Related Work

Benchmarks aimed at evaluating the audio capabilities of language models have been developed for several years. However, early approaches, such as ASR-GLUE (Feng et al., 2022), had limited capacity for assessing reasoning across different audio tasks. ASR-GLUE adapts GLUE (Wang et al., 2018), a benchmark for evaluating natural language understanding models, by converting its textual input into speech. Specifically, five NLU tasks from GLUE were selected, and their textual input was converted to audio by recording human speakers under various acoustic conditions. While valuable for evaluating NLU model robustness to ASR errors, this approach does not assess the audio reasoning capabilities of the models. Although ASR-GLUE provides a valuable starting point for audio modality evaluation, it should be noted that textual LLMs are evaluated using more comprehensive benchmarks that include complex reasoning tasks (Gao et al., 2023; Liang et al., 2023). Additionally, ASR-GLUE focuses solely on natural language understanding, omitting other audio-oriented tasks.

Wang et al. (2024a) proposed AudioBench, a benchmark designed for evaluating Audio LLMs. It includes 26 datasets, both new and existing, and focuses on three main areas: speech understanding, audio scene understanding, and audio classification. However, these categories are evaluated in isolation, which means that it cannot be a compre-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 953–983

hensive evaluation of audio reasoning. The benchmark uses a system formed from Whisper (Radford et al., 2023) and Llama-3 (Grattafiori et al., 2024) as a baseline. With the exception of automatic speech recognition tasks, evaluated with word error rate, the evaluation process primarily relies on the LLM-as-a-judge paradigm (Zheng et al., 2023) for performance assessment, raising potential bias concerns. Although AudioBench aims to evaluate key aspects of general-purpose Audio LLMs, it mainly focuses on the accuracy of individual audio tasks.

AIR-Bench, introduced by Yang et al. (2024), evaluates Audio LLMs in two dimensions: *foundation* and *chat*. The *foundation* component consists of 19 single-choice question tasks related to speech, sound, and music. The *chat* component features open-ended questions covering the same categories. The authors also create a category of mixed problems by combining speech-based tasks with sound and music tasks. The benchmark exclusively uses a closed-source LLM (GPT-4, OpenAI et al., 2024) as a judge in the evaluation process, without considering alternative evaluation methods. A pipeline consisting of GPT-4 combined with Whisper ASR (Radford et al., 2023) also serves as the baseline system. The study does not address the potential impact of using the same LLM as a judge and as an evaluated model, a known issue in the literature (Liu et al., 2023, Liu et al., 2024). Similar to AudioBench, AIR-Bench focuses on evaluating the accuracy of individual tasks rather than reasoning about the entire audio input, therefore it cannot be classified as a proper audio-reasoning benchmark.

SALMon, presented by Maimon et al. (2024), concentrates on features such as background noise, emotion, speaker identity, and room impulse response. It evaluates both intra-recording consistency and alignment with spoken text. The task involves comparing the model's likelihood assignments to two samples, one of which is more plausible. For consistency evaluation, one sample includes a feature shift mid-recording (e.g. background noise or speaker voice). The two samples being compared can differ in terms of background sound, e.g., the same conversation about opening a bank account might be accompanied by street noise in one instance, while in another, it could feature a more fitting background sound, such as that of a bank environment. While humans easily identify the more coherent sample, models often struggle, highlighting an interesting, previously underexplored area. However, unlike our work, the SALMon benchmark focuses on evaluating a single aspect of the model rather than comprehensively assessing whether the model can perform inference based on the source recording.

Audio modality performance results released by commercial providers of multimodal LLMs are rather limited. GPT-4o reports only automatic speech recognition and audio translation performance[1] using WER and BLEU metrics, with CoVoST-2 (Wang et al., 2020) as the evaluation dataset. Gemini 1.5 (Georgiev et al., 2024) follows a similar approach, using internal datasets alongside MLS (Pratap et al., 2020), FLEURS (Conneau et al., 2023), and CoVoST-2. Llama 3.1 (Grattafiori et al., 2024), a prominent open-weight multimodal model, also reports speech recognition and audio translation results on MLS, LibriSpeech (Panayotov et al., 2015), VoxPopuli (Wang et al., 2021), FLEURS, and CoVoST-2.

A common characteristic of the audio modality benchmarks discussed above is their emphasis on tasks that assess fundamental audio processing abilities, such as audio classification or sound event detection, in isolation. However, they tend to lack comprehensive, end-to-end evaluation of audio-based inference and reasoning, particularly within a cross-modal context, which is still an underexplored area. This situation has started to change only recently. An example could be the newly introduced MMAU benchmark (Sakshi et al., 2024), designed to assess multimodal audio understanding models on tasks requiring expert knowledge and complex reasoning. It comprises 10 000 audio clips paired with human-annotated questions and answers related to speech, environmental sounds, and music. The benchmark questions focus on information extraction and reasoning. While MMAU covers 27 different skills, none of the tasks proposed in our benchmark has a counterpart in MMAU.[2] Furthermore, compared to the benchmark presented in this paper, MMAU has several limitations. Specifically, a number of tasks rely on specialized musical knowledge, such as harmony, chord progression, or melody structure, which can hinder comprehensive error analysis by the broader expert community. A more significant issue is the separation of modalities: the questions about the recordings are in text form and are not an integral part of the

[1] https://openai.com/index/hello-gpt-4o

[2] A comparison of ART and MMAU tasks is provided in Appendix A.

audio input. In contrast, in our benchmark, the questions were converted to speech and integrated seamlessly with the input audio. Furthermore, in MMAU, questions are presented as closed-choice tests, with additional answers (distractors) generated using a closed-source language model (GPT-4). Taking into consideration that LLMs tend to recognize their own output (cf. Panickssery et al., 2024), this can affect the evaluation process in an unpredictable way. By comparison, *Yes/No* variant of our benchmark eliminates the risk of potential bias in evaluating GPT-4-based models due to the construction of the answer set.

3 Tasks

The preparation of the dataset is a multi-stage procedure. First, we formulate a set of rules for the tasks to be included in the dataset. We then survey a group of domain experts to recommend candidate tasks that align with these rules. Next, we eliminate any candidate tasks that cannot be reliably evaluated. Finally, we operationalize the proposed tasks by developing a set of templates that are instantiated with sound samples and synthesized speech in the last stage. The whole process is outlined in Figure 1.

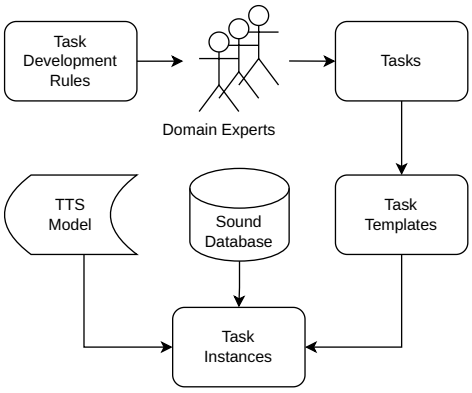

Figure 1: Task preparation process.

3.1 Task Selection

As the performance of MLLMs in tasks that address specific audio problems such as speaker diarization or gender identification can already be tested with the use of targeted datasets and the benchmarks that aim to evaluate AudioLLMs incorporate task-specific datasets (e.g. Wang et al., 2024a; Yang et al., 2024), we refrained from testing the qualities of MLLMs in isolation. Instead, we focused on tasks that combine different sound phenomena to assess the capability of MLLMs to reason over diverse input signals. Thus, to be included in the ART dataset, the task has to obey the following rule:

> **Rule 1:** The task should not be solvable by an LLM that consumes the output of a single specialized module that approaches a specific task and ignores all other sound phenomena presented in the audio signal.

This rule excludes tasks that can be solved by an LLM acting on speech transcription. Furthermore, it eliminates tasks that can be solved using an audio captioning model followed by a question answering system, effectively excluding the majority of *Chat* tasks proposed by Yang et al. (2024).

To simplify error analysis for independent researchers who decide to use our benchmark, we also assumed that the proposed tasks should not rely on the superhuman performance of the model or the competence of highly skilled individuals, such as musicians or sound engineers. Thus, the second rule that the tasks have to obey is:

> **Rule 2:** The task should be approachable by a person without professional training.

To gather tasks that obey the aforementioned rules, we surveyed a group of experienced speech and natural language processing engineers that included the authors of this paper. Initially, we collected 25 candidate tasks from 7 participants in total. In our effort to create a benchmark that produces results easily verifiable by an unskilled person, we decided to exclude from our candidate set all tasks that could lead to different outcomes due to individual variability. Thus, we rejected tasks that involved emotion classification or subjective assessment of sound quality. Furthermore, since we intended to provide an option to evaluate the model without depending on another LLM to serve as a judge, we eliminated tasks that cannot be framed as *Yes/No* questions. As a result, our benchmark is composed of nine tasks presented in Table 1.

3.2 Task Templates

We relied on the use of templates to generate the tasks. The rationale for adopting this approach was twofold. First, we wanted to control the size of the benchmark. Templates allowed us to easily increase the diversity of questions by expanding

Table 1: Descriptions and examples of questions for each task included in the ART benchmark.

Task name	Description	Example
Audio Arithmetics	Performing simple arithmetic reasoning with regard to the sounds heard.	Are there as many bell rings as there are cat meows?
Audio Transformation Detection	Recognizing whether one recording is a transformed version of the other.	Is the first recording a sped up version of the second recording?
Cross-Recording Language Identification	Comparison of the languages spoken in the recordings.	Is Budapest the capital of the country this speaker comes from?
Cross-Recording Speaker Identification	Comparison of the speakers in the recordings.	Is the same person heard speaking on both recordings?
Selective Text Inference	Inference based on some uttered content which is selected on the basis of the properties of some of the speakers.	Is green the answer to the question asked by a man?
Sound Reasoning	Reasoning based on the recognized sound.	Is the animal that makes the following sound bigger than a horse?
Speech Features Comparison	Comparison of two recordings regarding speech features present in them.	Is the second recording the same text but read with a Scottish accent?
Text and Sound Reasoning	Questions that require both sound features and text understanding to be answered.	Is the person talking about the following sound?
Text and Temporal Localization Reasoning	Questions that require both noises from localization (surroundings) and text understanding to be answered.	Does the speaker describe the acoustic scene that they are in?

sets of possible slot values while maintaining the preferred size of the dataset. Second, as with any dataset released to the public, there is an ongoing risk of training data contamination for models that will be released in the future. Having a set of templates that can be populated with fresh sound samples from undisclosed sources will allow developers of future models to re-run the evaluation procedure while mitigating the risk of obtaining over-optimistic results due to data contamination. For each task, we developed a set of question templates containing empty slots filled with appropriate values. The values could be single words, whole sentences, or special tags, later to be replaced with sounds. During template creation, values were randomly chosen from previously prepared sets in a manner that allowed us to automatically generate a target answer to the question. An example of a template is shown in Figure 2.[3]

3.3 Task Instances

Based on this information, speech samples for voice-cloning were randomly chosen from previously prepared sets with required characteristics. This approach also helped to enhance the diversity of the dataset.

To generate instances of each task, we needed three types of audio recordings – questions, utterances and sounds.

To unify audio prompts, we synthesized all questions using the same sample for voice cloning. As the goal was to generate intelligible speech, the choice was limited to these samples from the LJ Speech dataset (Ito and Johnson, 2017), for which the Whisper medium model (Radford et al., 2023) obtained WER of 0. The chosen samples were used to generate synthetic speech, which was again transcribed to find samples that yielded a WER of 0. The remaining samples were evaluated by human reviewers, and the most natural-sounding sample was selected as the prompt for question synthesis.

Five of the tasks required additional utterances spoken by different speakers. The samples for voice cloning were selected from the GLOBE dataset (Wang et al., 2024b). As with selecting the prompt for the questions, the samples were transcribed using the Whisper medium model, used as

[3]Task templates are described in detail in Appendix B.

Table 2: Summary statistics of the ART benchmark.

Task	# Samples	# Templates	# Speakers	# Utterances	# Sounds	Total length
AA	1000	6	N/A	N/A	5	3h 46m 10s
ATD	1000	4	N/A	N/A	4	3h 53m 30s
CRLI	1000	6	12	12	N/A	3h 32m 1s
CRSI	1000	4	4	8	N/A	3h 46m 17s
STI	1000	4	4	36	N/A	3h 9m 47s
SR	1000	15	N/A	N/A	20	3h 8m 15s
SFC	1000	4	10	3	N/A	2h 37m 7s
TSR	1000	8	4	16	17	3h 25m 24s
TTLR	1000	4	4	15	8	3h 47s
Total	9000	55	22	86	25	30h 19m 18s

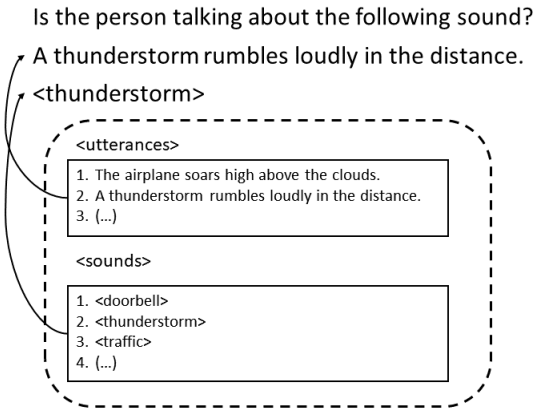

Figure 2: Example of a template. Both the sentence and the sound are chosen randomly from predefined lists; the target answer can be inferred from these values. Proper speakers for voice cloning for each part of the template are selected.

prompts for voice cloning, and transcribed again. Thus, after human evaluation, 10 samples (one per speaker) were selected to synthesize utterances.

One of the tasks, Cross-Recording Language Identification, required utterances in languages other than English. For this purpose, we used the VoxPopuli dataset (Wang et al., 2021). We selected four languages – Estonian, Finnish, Hungarian and Polish. For each language, we identified speakers for which WER obtained with the Whisper medium model was equal to 0. We then narrowed our selection to recordings with a duration between 2 and 5 seconds. This process resulted in the selection of three speakers (one utterance each) for each language.

The last type of audio recordings, sounds, included short sounds of e.g. an animal, tunes, and background noises. All sounds were manually selected from samples available under the Creative Commons 0 license in the Freesound dataset (Font et al., 2013), resulting in 13 short sounds, four short tunes, and eight background sounds. The short sounds were manually trimmed as necessary to include a single sound event per sample.

For speech synthesis, we utilized a pipeline based on Voicebox (Le et al., 2023), a text-to-speech model that demonstrates zero-shot capabilities in reconstructing audio segments from textual inputs and speech prompts. The architecture adapts the transformer model (Vaswani et al., 2017), with modifications including the use of rotary positional embedding (Su et al., 2023) instead of ALiBi self-attention bias (Press et al., 2022). To predict token durations, we employ a DurationPredictor model, similar to Voicebox but smaller in size. We also utilize CTC-based forced alignment to discover token durations in an unsupervised manner. A HiFi-GAN vocoder (Kong et al., 2020) is used to map audio features to speech, consisting of a fully convolutional generator and two discriminators. The input tokens are represented as phonetic labels, and mel-scale spectrograms with 80 channels are used as audio features.

After obtaining all necessary samples of questions, utterances, and sounds, the audio recordings were automatically merged according to the prepared template configurations. All recordings were normalized to -20 dBFS. The duration of silence between the question and the utterance or sound was manually adjusted to ensure that audio prompts sound natural and that the gaps between recordings are sufficient to distinguish them. The short sounds and tunes concatenated with a question were truncated to the maximum duration of 5 seconds, and a fade-out was applied to eliminate sudden vol-

Table 3: Results of model evaluation on the ART benchmark using *Yes/No* approach.

Model	% Relevant	Absolute Accuracy	Relative Accuracy
Whisper + Llama	99.91	0.5404	0.5408
Whisper + Qwen	**99.94**	**0.5621**	**0.5625**
Audio Flamingo 3	**100.00**	**0.5473**	**0.5473**
GAMA	42.56	0.2155	0.5064
Qwen-Audio-Chat	64.09	0.3312	0.5168
Qwen2-Audio (zero-shot)	85.41	0.4431	0.5188
Qwen2-Audio (one-shot, same template)	87.30	0.4710	0.5395
Qwen2-Audio (one-shot, different template)	68.08	0.3526	0.5179
Ultravox v0.4.1	89.19	0.4682	0.5250
Ultravox v0.6	99.74	0.5309	0.5323

ume changes. In the case of instances that required an overlay of background sounds, the underlying audio was attenuated by 20 dB to ensure speech intelligibility, and both fade-in and fade-out were applied.

The consolidation of templates resulted in 9 000 samples that constitute the final dataset, which amounts to over 30 hours of audio, as shown in Table 2. The prepared dataset is fully balanced across labels and tasks. Of the 9 000 total samples, 4 500 have the expected answer of *Yes* and 4 500 have the expected answer of *No*. This balance is maintained within each of the nine tasks, with 1 000 samples evenly split between 500 *Yes* and 500 *No* instances. Furthermore, the balance was maintained at the level of task templates wherever possible. For instance, in the **Audio Transformation Detection** task, template **ATD_0** has 56 *Yes* and 56 *No* samples; template **ATD_1** has 56/56; template **ATD_2** has 112/112; and template **ATD_3** has 276/276. This ensures a uniform distribution of labels across tasks.

4 Experiments

4.1 Setup

We conducted a series of experiments evaluating multimodal models on prepared tasks. Two different approaches were adopted:

- *Yes/No* – where the model was instructed to answer only *Yes* or *No*,
- *Descriptive* – where the form of the answer was not specified and the model was allowed to give a descriptive answer.

As all the tasks were designed to have either an affirmative or negative answer, we focused on *Yes/No* approach in this section. However, to gain better understanding of the reasons why the models fail to accomplish the tasks, we also studied the open-ended responses yielded by *Descriptive* approach with summary scores reported in Table 5 and the detailed analysis given in Appendix E.

To ensure the reliability of the results, the responses obtained using *Yes/No* approach were automatically evaluated. If the answer was *Yes* or *No*, it was marked as relevant, and irrelevant otherwise. The responses marked as relevant were further classified as correct or incorrect. As LLM-as-a-judge is a commonly used evaluation method, we decided to use it in the *Descriptive* approach. For this purpose, we used two models – Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and Qwen3-32B (Yang et al., 2025). In the prompt, the model was instructed that it would receive a question, an expected answer and the received answer. Its task was to determine whether the generated answer was relevant or not, and to state whether it was correct or why it was labeled as irrelevant. The full prompt used for this evaluation is shown in Appendix D, along with the inference parameters.

4.2 *Yes/No* approach

The results of evaluation using *Yes/No* approach are shown in Table 3. The relative accuracy is defined as the accuracy calculated only on relevant responses. In this case, the models were instructed to answer only *Yes* or *No*. The inference on each of the models was run five times and the results were averaged to assess whether the models exhibit superiority over random guessing.

Two cascaded systems where evaluated, both using Whisper Large v3 (Radford et al., 2023) to obtain transcriptions of the audio prompts. First

Table 4: Absolute accuracy per task on the ART benchmark using *Yes/No* approach.

Model	AA	ATD	CRLI	CRSI	STI	SR	SFC	TSR	TTLR
Whisper + Llama	0.505	0.483	0.458	**0.505**	**0.665**	0.525	**0.557**	0.643	0.521
Whisper + Qwen	**0.510**	**0.504**	**0.551**	0.501	0.625	**0.654**	0.532	**0.653**	**0.528**
Audio Flamingo 3	**0.516**	0.492	0.569	**0.517**	0.554	**0.700**	0.494	0.566	**0.517**
GAMA	0.036	0.376	0.276	0.349	0.338	0.050	0.299	0.068	0.147
Qwen-Audio-Chat	0.498	**0.521**	0.008	0.123	0.201	0.549	0.181	0.525	0.375
Qwen2-Audio (zero-shot)	0.488	0.517	0.501	0.212	0.532	0.432	0.524	0.358	0.425
Qwen2-Audio (one-shot, same template)	0.493	0.282	0.517	0.454	0.534	0.517	**0.527**	0.543	0.373
Qwen2-Audio (one-shot, different template)	0.441	0.245	0.464	0.340	0.477	0.176	0.467	0.392	0.172
Ultravox v0.4.1	0.475	0.288	0.516	0.489	0.478	0.438	0.507	0.511	0.512
Ultravox v0.6	0.480	0.501	**0.579**	0.504	**0.559**	0.526	0.521	**0.590**	0.513
Average	0.436	0.413	0.430	0.386	**0.490**	0.430	0.457	0.476	0.396

system used Llama-3.3-70B-Instruct (Grattafiori et al., 2024) to answer the questions, and the second one utilized Qwen3-32B (Yang et al., 2025). Both cascaded systems generated almost 100% relevant answers, achieving 54.04% and 56.21% absolute accuracy, respectively.

Considering MLLMs, Audio Flamingo 3 (Goel et al., 2025) is the only model that returned 100% relevant answers. It also outperformed the other models with the accuracy of 54.73%. Ultravox-v0.6-Llama3.3-70B (Fixie.ai, 2025) obtained the second highest number of relevant responses. It managed to achieve absolute accuracy of 53.09%. The previous version of this model, Ultravox-v0.4.1-Llama3.1-8B, generated 10% less relevant responses, achieving less than 50% absolute accuracy in all five runs. Qwen2-Audio-7B-Instruct (Chu et al., 2023) generated only 85.41% relevant responses in the zero-shot approach. As the authors of this model suggest using the one-shot approach, two additional experiments were performed. In the first experiment, the model was given a sample from the same template as an example. This resulted in less than 2% more relevant answers and less than 3% greater accuracy. In the second experiment, the model was given as an example a sample from the same task but a different template. This approach resulted in a 17% decrease in the number of relevant responses and a 9% decrease in accuracy compared to zero-shot approach. The lowest results in the Qwen family of models were achieved by Qwen-Audio-Chat (Chu et al., 2023). It generated only 64.09% relevant responses, achieving average absolute accuracy of 33.12%. The GAMA model (Ghosh et al., 2024) reached only 21.55% accuracy, which was the lowest in the evaluation using *Yes/No* approach.

The results per task achieved by the models for the *Yes/No* approach are shown in Table 4. Based on these, the **Cross Recording Speaker Identification** proved to be the hardest task – the average absolute accuracy is 38.63% and none of the models achieved accuracy higher than 50% in any of the five runs. Of these, the best average absolute accuracy was achieved for the **Selective Text Inference** task.

4.3 *Descriptive* approach

Table 5 shows the results of the experiments performed with the *Descriptive* approach and both Llama and Qwen3 as a judge.

According to both judges, none of the models achieved satisfactory results. Only the cascaded system using Qwen3 achieved absolute accuracy higher than 50%, but only when it evaluated itself. The agreement between judges exceeded 74% in all cases, and the highest value reached was 97.3%. However, this case involved the GAMA model, which had an accuracy of less than 2.5%.

Among the MLLMs, Qwen2-Audio in the one-shot approach using the same template as an example, achieved the best results – Llama 3.3 and Qwen3 judged it achieved 47.35% and 49.96% accuracy, respectively. This experiment also resulted in the highest fraction of relevant answers.

It is worth noting that Llama 3.3 recognized more relevant answers when judging itself, Qwen3, and both Ultravox v0.4.1 and Ultravox v0.6. On the other hand, Qwen3 was significantly more indulgent when evaluating the MLLMs from the Qwen

Table 5: Results of the model evaluation on the ART benchmark using *Descriptive* approach.

Model	Llama 3.3		Qwen3		
	Relevant	**Accuracy**	**Relevant**	**Accuracy**	**Agreement**
Whisper + Llama	62.62%	0.3535	62.17%	0.4834	74.07%
Whisper + Qwen	**89.68%**	**0.4882**	**87.97%**	**0.5697**	**76.09%**
Audio Flamingo 3	74.07%	0.3536	76.99%	0.3524	82.10%
GAMA	1.31%	0.0089	2.40%	0.0154	**97.30%**
Qwen-Audio-Chat	14.50%	0.0778	19.89%	0.0548	86.04%
Qwen2-Audio (zero-shot)	92.26%	0.4490	94.94%	0.4683	82.43%
Qwen2-Audio (one-shot, same template)	**93.94%**	**0.4735**	**96.36%**	**0.4996**	83.31%
Qwen2-Audio (one-shot, different template)	93.16%	0.4495	95.28%	0.4834	81.30%
Ultravox v0.4.1	60.20%	0.2398	50.59%	0.1692	76.19%
Ultravox v0.6	79.42%	0.3952	77.69%	0.4425	77.14%

family of models.

An in-depth analysis of the results per task in the *Descriptive* approach is available in Appendix E. In case of cascaded systems, Qwen3 assessed its own performance significantly better across all tasks except for **Speech Features Comparison**. The best average accuracy was obtained on the **Speech Features Comparison** task when judged by Llama 3.3, and on **Text and Sound Reasoning** judged by Qwen3. However, the results on the remaining tasks are significantly understated due to the accuracy obtained by Qwen-Audio-Chat (5.48-7.78%) and GAMA (0.89-1.54%).

4.4 Error analysis

Error analysis for both approaches revealed distinct yet overlapping failure patterns based on human evaluation. Errors in the *Yes/No* approach were primarily driven by failures in audio understanding. Specifically, models did not recognize the presence of speech or sound. Additionally, there was systematic task confusion, which led to transcription or speaker recognition instead of question answering.

In contrast, the *Descriptive* approach exhibited a broader and more heterogeneous set of errors. In addition to frequently failing to recognize the question, the models often produced random, speculative, or language-inconsistent responses and showed stronger biases toward transcription and speaker identification. These behaviors suggest less stable response control in case of the *Descriptive* approach.

Overall, the results suggest that both approaches are susceptible to task misinterpretation. However, in the *Yes/No* approach errors are concentrated around audio perception and task confusion. In contrast, the *Descriptive* approach results in more diverse and less predictable failure behaviors. A quantitative analysis is provided in Appendix G.

5 Benchmark Validation

Although the task collection process described in Section 3 relies on human expertise and clearly defined rules, the task instantiation procedure depends on templates and synthesized speech which may potentially result in a benchmark that is either too complex to be comprehensible by humans or too simple for models to solve.

To address the first issue, we designated ART-H, a subset of the dataset that consists of 24 samples per task resulting in 216 samples in total, that enables manual verification of the evaluation results in under one hour. We presented ART-H prompts to human evaluators, who answered the questions with either *Yes* or *No*. In this way, we achieved a human baseline of 92.90%, thus confirming the suitability of the prepared tasks. Three tasks – **Audio Arithmetics**, **Selective Text Inference**, and **Text and Temporal Localization Reasoning** – turned out to be the easiest. The worst results were obtained on **Speech Features Comparison** and **Cross-Recording Language Identification**. The reason may be that these tasks focus on distinguishing accents or languages, which can be challenging for those unfamiliar with a particular dialect. Detailed results of the human evaluation are provided in Appendix C. Furthermore, to verify if the choice of particular samples to be included in ART-H impacts the results in a meaningful way,

we evaluated the models under study with respect to randomly sampled 216-element subsets of ART and demonstrated that this procedure results in a standard deviation of less than 0.035 in terms of absolute accuracy (cf. Appendix H).

The reliance on synthesized samples raises concerns about the impact of the audio quality on the benchmark results. If the benchmark encompasses audio samples of poor quality, the models could underperform due to artefacts in data. Taking into consideration that we use a state-of-the-art TTS model and perform human evaluation of the ART-H subset, this is not the case. On the other hand, high-quality synthesis can potentially lead to overoptimistic results in speech-related tasks. However, the dependence on synthesized speech ensures that any mistakes observed in the models' performance are due to reasoning errors rather than poor or ambiguous input. If a model demonstrates weak performance on clean audio samples, it will likely perform worse in noisy conditions. The results presented in Section 4 show that none of the models attained satisfactory performance with regard to the synthesized data. Therefore, we believe that the use of audio data that exhibit more demanding acoustic conditions can be postponed.

6 Conclusion

In this paper, we proposed Audio Reasoning Tasks (ART), a new benchmark for assessing the performance of MLLMs. Contrary to the existing benchmarks for testing the audio modality of MLLMs that aim at evaluating various audio capabilities in isolation, our dataset encompasses tasks that, to be solved, require combining a diverse set of skills grounded in the audio domain. The benchmark was designed to be solvable by an unskilled person without hearing impairment. The experiments that were conducted showed that while the benchmark is easily comprehensible by humans, it still poses a challenge for the state-of-the-art open-weight models that we investigated.

7 Limitations

While our benchmark can be employed to determine if the model under study is capable of reasoning over an audio signal, the proposed set of tasks cannot be considered complete. Therefore, a successful completion of this test cannot guarantee that the model's audio reasoning skills are at the human level.

Acknowledgments

This research was partially funded by the *SPEAKCAIR: SPEech-Aware Conversational AI Research* project, a cooperation between Adam Mickiewicz University and Samsung Electronics Poland.

References

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: FEW-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Lingyun Feng, Jianwei Yu, Yan Wang, Songxiang Liu, Deng Cai, and Haitao Zheng. 2022. ASR-Robust Natural Language Understanding on ASR-GLUE dataset. In *Proc. Interspeech 2022*, pages 1101–1105.

Fixie.ai. 2025. Ultravox v0.6. Https://huggingface.co/fixie-ai/ultravox-v0_6-llama-3_3-70b.

Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 411–412, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Petko Georgiev et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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. *Preprint*, arXiv:2406.11768.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034. Curran Associates, Inc.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy S Liang. 2024. VHELM: A holistic evaluation of vision language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 140632–140666. Curran Associates, Inc.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. *Preprint*, arXiv:2311.09766.

Gallil Maimon, Amit Roth, and Yossi Adi. 2024. A suite for acoustic language model evaluation. *arXiv preprint arXiv:2409.07437*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Proc. Interspeech 2020*, pages 2757–2761.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *Preprint*, arXiv:2108.12409.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

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

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. RoFormer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Pino. 2020. CovoST 2 and massively multilingual speech-to-text translation. *Preprint*, arXiv:2007.10310.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024b. GLOBE: A high-quality English corpus with global accents for zero-shot speaker adaptive text-to-speech. *Preprint*, arXiv:2406.14875.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A ART vs. MMAU tasks

A.1 Audio Arithmetics

The closest task to **Audio Arithmetics** is **Temporal Event Reasoning** which includes questions about how many times a given sound occurred. However, there is only one kind of sound present in the audio recording, while in **Audio Arithmetics** the sounds occur next to each and the model is required to count both of them and compare, or reason about the evenness of the number of occurrences. There are two different similar tasks – **Counting**, which requires counting how many speakers are involved in a conversation, and **Phonological Sequence Decoding**, which involves counting how many times a given word appears in the audio, but none of them is the exact match to **Audio Arithmetics**.

A.2 Audio Transformation Detection

There is no task similar to **Audio Transformation Detection**. While **Temporal Event Reasoning** includes question about identifying the shortest or longest sound, it does not involve comparison of two sounds and detection of modification applied on the audio.

A.3 Cross-Recording Language Identification

There is no task similar to **Cross-Recording Language Identification** – none of them involves recognition of the language spoken. While **Counting** requires counting speakers involved in a conversation, no reasoning over their identity is included.

A.4 Cross-Recording Speaker Identification

While **Counting** involves identification of the number of speakers, it does not require comparison of their identities. There is no other task in MMAU that would involve determining whether the same speaker is being heard on both recordings or if the same speakers are involved in two different dialogues, as in **Cross-Recording Speaker Identification**.

A.5 Selective Text Inference

The closest tasks to **Selective Text Inference** are: **Key highlight Extraction** and **Conversational Fact Retrieval**, which involve questions about what speaker 1 or 2 said, and **Event-Based Knowledge Retrieval**, which requires question answering based on the given utterance. However, none of this tasks involves reasoning based on comparison of speaker characteristics. In **Selective Text Inference**, in addition to question answering, each template includes identification of speaker's gender or recognition of the utterance's subject.

A.6 Sound Reasoning

Sound-Based Event Recognition requires identifying of an event based on audio. **Ambient Sound Interpretation** includes two sounds – one of them is mentioned in the question, and the second one is to be identified. **Eco-Acoustic Knowledge** is based on recognition of the environment and the question often suggests the answer, e.g. waves can be heard and the question asks what natural disaster can be inferred that typically results in significant loss of life and property due to subsequent flooding particularly in coastal regions. **Event-Based Sound Reasoning** asks what caused a given sound, e.g. howling – which again suggests the answer, as a howl is often associated with wolves. While all of these tasks involve some kind of reasoning over sounds, none of them is the same as **Sound Reasoning**, which requires not only sound recognition, but also actual reasoning. For example, the task is to recognize two animals based on their sounds and determining whether both of them are mammals.

A.7 Speech Features Comparison

There is no task similar to **Speech Features Comparison** as none of them considers speakers' characteristics such as accent, age or gender. While **Counting** involves counting how many speakers can be heard, no reasoning over their characteristics is required.

A.8 Text and Sound Reasoning

There is no task that could be a match for **Text and Sound Reasoning**. While some tasks require reasoning over sounds, as described for **Sound Reasoning** task, none of them involves reasoning based on both sounds and utterances. For example, the **Text and Sound Reasoning** task requires recognition of two animals based on a sound and an utterance, and determining whether both of them are smaller than an elephant.

A.9 Text and Temporal Localization Reasoning

The most similar task is **Acoustic Scene Reasoning**, which includes identification of localization where a given sound can be heard. While **Text and Temporal Localization Reasoning** also asks about

the time and place of the recording, it involves reasoning over both speech and background sounds and comparison of the acoustic environment of two speakers.

B Task Templates

B.1 Audio Arithmetics

Let S be the predefined set of sounds:

$$S = \{\text{bell ring}, \text{cat meow}, \text{dog bark}, \text{elephant trumpet}, \text{horse neigh}\}$$

Template AA_0 requires comparing the number of occurrences of sounds s_A and s_B. To do this, the model must count how many times sound s_A and sound s_B occurred, and then determine if the number is equal.

$$\text{Are there as many } s_A \text{ as there are } s_B? + (n_A \times s_A) + (n_B \times s_B) \quad \text{(AA\_0)}$$

where $s_A, s_B \in S$, $s_A \neq s_B$; $n_A, n_B \in \langle 1, 4 \rangle$. In this template, the answer is *Yes* if $n_A = n_B$.

Similarly, in AA_1, the model must determine whether sound s_A occurs twice as much as sound s_B.

$$\text{Are there twice as many } s_A \text{ as there are } s_B? + (n_A \times s_A) + (n_B \times s_B) \quad \text{(AA\_1)}$$

where $s_A, s_B \in S$, $s_A \neq s_B$; $n_A, n_B \in \langle 1, 4 \rangle$. The answer is *Yes* if $n_A = 2n_B$.

Templates AA_2 and AA_3 involve determining whether sound s_A occurred an even or an odd number of times. In both cases, sound s_B can occur next to s_A, which requires correct sound recognition. In template AA_2 the answer is *Yes* if $n_A \mod 2 = 0$, and in AA_3 if $n_A \mod 2 \neq 0$.

$$\text{Does the } s_A \text{ an even number of times?} + (n_A \times s_A) + (n_B \times s_B) \quad \text{(AA\_2)}$$

$$\text{Does the } s_A \text{ an odd number of times?} + (n_A \times s_A) + (n_B \times s_B) \quad \text{(AA\_3)}$$

where $s_A, s_B \in S$, $s_A \neq s_B$; $n_A \in \langle 1, 4 \rangle$; $n_B \in \langle 0, 4 \rangle$.

Template AA_4 requires comparison if sound s_A occurred more times than sound s_B. Then, the answer is *Yes* if $n_A > n_B$. Similarly, in template AA_5 the answer is *Yes* if sound s_A occurred less times than sound s_B, i.e. $n_A < n_B$.

$$\text{Are there more } s_A \text{ than } s_B? + (n_A \times s_A) + (n_B \times s_B) \quad \text{(AA\_4)}$$

$$\text{Are there less } s_A \text{ than } s_B? + (n_A \times s_A) + (n_B \times s_B) \quad \text{(AA\_5)}$$

where $s_A, s_B \in S$, $s_A \neq s_B$; $n_A, n_B \in \langle 1, 4 \rangle$.

B.2 Audio Transformation Detection

Let S be the predefined set of sounds and T the predefined set of transformations:

$$S = \{\text{circus}, \text{drum}, \text{elevator}, \text{piano}\}$$
$$T = \{\text{higer}, \text{louder}, \text{lower}, \text{quieter}, \text{reversed}, \text{slowed down}, \text{sped up}, \text{truncated}\}$$

then $t(s)$ means that transformation $t \in T$ is applied to sound $s \in S$.

Template ATD_0 involves determining whether first sound is a modified version of the second recording.

$$\text{Is the first recording a } t \text{ version of the second recording?} + t_A(s_A) + s_B \quad \text{(ATD\_0.1)}$$

where $s_A, s_B \in S$; $t, t_A \in T$. The answer is *Yes* if $s_A = s_B \wedge t = t_A$.

To include additional affirmative answers, a second version of this template, ATD_0.2, was designed. In this case, transformation t_B is applied to sound s_B. The answer is *Yes* if t_B is the inverse transformation of t, i.e. $t_B = t^{-1}$, and $s_A = s_B$.

$$\text{Is the first recording a } t \text{ version of the second recording?} + s_A + t_B(s_B) \quad \text{(ATD\_0.2)}$$

where $s_A, s_B \in S$; $t, t_B \in T \setminus \{\text{reversed}, \text{truncated}\}$.

Similarly, two versions of template ATD_1 were designed. ATD_1.1 requires determining whether sound s_B is a transformed version of s_A. The answer is *Yes* if $s_A = s_B \wedge t = t_B$.

Is the second recording a t first recording?
$+ s_A + t_B(s_B)$ (ATD_1.1)

where $s_A, s_B \in S$; $t, t_B \in T$.

In ATD_1.2, transformation t_A is applied to s_A, and the answer is affirmative if $s_A = s_B \wedge t_A = t^{-1}$.

Is the second recording a t first recording?
$+ t_A(s_A) + s_B$ (ATD_1.2)

where $s_A, s_B \in S$; $t, t_A \in T \setminus \{\text{reversed}, \text{truncated}\}$.

Template ATD_2 requires determining if one of the recordings is the same as the other one but with transformation t applied.

Is one of the recordings a t version of the other one? $+ t_A(s_A) + t_B(s_B)$ (ATD_2)

where $s_A, s_B \in S$; $t_A, t_B \in T \cup \{\text{none}\}$; $\neg(t_A = \text{none} \wedge t_B = \text{none})$. In this case, the answer is *Yes* if $s_A = s_B \wedge \left(t \in \{t_A, t_B, t_A^{-1}, t_B^{-1}\}\right)$.

In case of template ATD_3, transformations are applied to both sounds, and the model must determine whether the sounds are the same but the transformations differ. Then, the answer is affirmative if $s_A = s_B \wedge t_A \neq t_B$.

Are these two recordings the same with different transformations? $+ t_A(s_A) + t_B(s_B)$ (ATD_3)

where $s_A, s_B \in S$; $t_A, t_B \in T$.

B.3 Cross-Recording Language Identification

Let L be the predefined set of languages and S the predefined set of speakers:

$$L = \{\text{Estonian}, \text{Finnish}, \text{Hungarian}, \text{Polish}\}$$
$$S = \{l_i \mid l \in L,\ i \in \{1, 2, 3\}\}$$

then

$$\text{language}(l_i) := l \text{ for all } l \in L,\ i \in \{1, 2, 3\}$$

Template CRLI_0 requires determining if all of the three people in the recording speak the same language.

Do all the people in the recording speak the same language? $+ s_A + s_B + s_C$ (CRLI_0)

where $s_A, s_B, s_C \in S$; $s_A \neq s_B \neq s_C$. The answer is *Yes* if $\text{language}(s_A) = \text{language}(s_B) = \text{language}(s_C)$.

Template CRLI_1 is similar, but consists of only two recordings. In this case, the answer is *Yes* if $\text{language}(s_A) = \text{language}(s_B)$.

Can two different languages be heard on the recording? $+ s_A + s_B$ (CRLI_1)

where $s_A, s_B \in S$; $s_A \neq s_B$.

Template CRLI_2 is the opposite of CRLI_0, i.e. the model is must assess whether the speakers use three different languages.

Can three different languages be heard on the recording? $+ s_A + s_B + s_C$ (CRLI_2)

where $s_A, s_B, s_C \in S$ and $s_A \neq s_B \neq s_C$. For CRLI_2, the answer is *Yes* if $\text{language}(s_A) \neq \text{language}(s_B) \neq \text{language}(s_C)$.

Template CRLI_3 requires recognition of the speaker's country of origin and determining if the given city is it's capital. Let C be the predefined set of cities:

$$C = \{\text{Budapest}, \text{Helsinki}, \text{Tallinn}, \text{Warsaw}\}$$

then $\text{is\_capital}(c, s)$ implies that city $c \in C$ is the capital city of the speaker $s \in S$ country of origin.

Is c the capital of the country this speaker comes from? $+ s$ (CRLI_3)

where $c \in C$, $s \in S$. In this case, the answer is affirmative if $\text{is\_capital}(c, s)$.

Template CRLI_4 involves assessing whether both speakers s_A and s_B are from the same country. Since each language in set L is an official language in one country, obtaining the correct answer comes

down to comparing the languages spoken by the speakers, i.e. the answer is *Yes* if $\text{language}(s_A) = \text{language}(s_B)$.

Are both speakers from the same country? $+ s_A + s_B$ (CRLI_4)

where $s_A, s_B \in S$, $s_A \neq s_B$.

In template CRLI_5, the model must verify whether the given country borders the speaker's country of origin. Let N be the predefined set of countries:

$$N = \{\text{Belarus}, \text{Estonia}, \text{Finland}, \text{Latvia}, \text{Norway}, \text{Slovakia}, \text{Ukraine}\}$$

then $\text{borders}(n, s)$ implies that country $n \in N$ borders speaker $s \in S$ country of origin.

Does n border the speaker's country of origin? $+ s$ (CRLI_5)

where $s \in S$, $n \in N$. The answer to CRLI_5 is *Yes* if $\text{borders}(n, s)$.

B.4 Cross-Recording Speaker Identification

Let S be the predefined set of speakers:

$$S = \{s_1, s_2, s_3, s_4\}$$

and U the predefined set of utterances:

- "The soft glow of the morning sun filtered through the curtains."
- "The rain tapped softly against the window."
- "The birds chirped a gentle melody."
- "It was a perfect moment to pause and breathe deeply."

then $s(u)$ means that utterance $u \in U$ is spoken by speaker $s \in S$.

Templates CRSI_0 and CRSI_1 require determining whether both utterances u_A and u_B are spoken by the same person, i.e. the answer is affirmative if $s_A = s_B$.

Are both recordings spoken by the same person? $+ s_A(u_A) + s_B(u_B)$ (CRSI_0)

Is the same person heard speaking on both recordings? $+ s_A(u_A) + s_B(u_B)$ (CRSI_1)

where $s_A, s_B \in S$; $u_A, u_B \in U$; $u_A \neq u_B$.

Template CRSI_2 involves assessing whether the same people are speaking in both dialogues d_A and d_B.

Let D be the predefined set of dialogues:

- "Did you remember to water the plants today?"
  "Oh no, I completely forgot!"
- "Did you hear about the new café downtown?"
  "No, what's it like?"
- "Are we meeting at 7 tonight?"
  "I thought it was 7:30?"
- "Did you finish the group project?"
  "Almost, I just need to finalize the slides."

then $d(s_A, s_B)$ means that the first utterance in dialogue $d \in D$ is spoken by the speaker $s_A \in S$, and the second by the speaker $s_B \in S$.

Are the same people speaking in both recordings? $+ d_A(s_{A_0}, s_{A_1}) + d_B(s_{B_0}, s_{B_1})$ (CRSI_2)

where

$$s_{A_0}, s_{A_1}, s_{B_0}, s_{B_1} \in S$$
$$d_A, d_B \in D, \quad d_A \neq d_B$$
$$s_{A_0} \neq s_{A_1} \wedge s_{B_0} \neq s_{B_1}$$

The answer for CRSI_2 is *Yes* if $(s_{A_0}, s_{A_1}) = (s_{B_0}, s_{B_1}) \vee (s_{A_0}, s_{A_1}) = (s_{B_1}, s_{B_0})$.

Similarly, template CRSI_3 requires verifying if the same people are speaking in both dialogues, but in this case the order of speakers is also taken into account.

Do the people on both recordings speak in the same order? $+ d_A(s_{A_0}, s_{A_1}) + d_B(s_{B_0}, s_{B_1})$ (CRSI_3)

where

$$s_{A_0}, s_{A_1}, s_{B_0}, s_{B_1} \in S$$
$$d_A, d_B \in D, \quad d_A \neq d_B$$
$$s_{A_0} \neq s_{A_1} \wedge s_{B_0} \neq s_{B_1}$$

The answer for CRSI_3 is *Yes* if $s_{A_0} = s_{B_0} \wedge s_{A_1} = s_{B_1}$.

B.5 Selective Text Inference

Let G and S be the predefined sets of genders and speakers, respectively:

$$G = \{\text{woman}, \text{man}\}$$
$$S = \{s_1, s_2, s_3, s_4\}$$

and U_0 the predefined set of utterances:

- "He walked into the barber shop and requested a traditional shave and a haircut."
- "I saw her yesterday, she is pregnant."
- "This guy is playing computer games with his dog."
- "He adjusted his tie before heading to the office."
- "Person leaving the building was wearing a dress so she appeared to be a woman."
- "He tossed his keys onto the counter after walking into the house."

then:

- $s(u)$ means that utterance $u \in U_i, \quad i \in \{0, \ldots, 3\}$ is spoken by speaker $s \in S$
- $\text{gender}(s) \in G$ is a function that returns the gender of the speaker $s \in S$
- $\text{talking\_about}(u, g)$ means that the utterance $u \in U_i, \quad i \in \{0, \ldots, 3\}$ is talking about gender $g \in G$

Template STI_0 requires determining whether the speaker s is talking about a person of the opposite gender, i.e. the answer is affirmative if $\neg\text{talking\_about}(u, \text{gender}(s))$.

Is the person the speaker is talking about of the opposite gender? $+ s(u)$ (STI_0)

where $s \in S$, $u \in U_0$.

Let B be the predefined set of subjects:

$$B = \{\text{health}, \text{entertainment}, \text{environment}, \text{politics}, \text{sports}, \text{technology}\}$$

and U_1 the predefined set of utterances:

- "The constitution sets the rules for a country's government."
- "The Earth's environment is our responsibility to protect."
- "Theme parks and amusement parks offer thrilling rides and attractions."
- "Tennis players compete in Grand Slam tournaments."
- "Cloud computing has made it easier to store and access data remotely."
- "A good night's sleep is crucial for physical and mental recovery."
- "Conserving water is essential for our ecosystem."
- "Eating a balanced diet is essential for well-being."
- "Reading books is a great way to escape reality."

then $\text{subject}(u) \in B$ is a function that returns the subject of the utterance $u \in U_1$.

Template STI_1 involves assessing whether a speaker of the given gender is talking about the given subject. This requires the model to recognize gender of both speaker s_A and s_B, identify if either of them is $g \in G$, and if so, verify if they are talking about $b \in B$.

Is the g in the recording talking about b?
$+ s_A(u_A) + s_B(u_B)$ (STI_1)

where

$$g \in G, \quad b \in B$$
$$s_A, s_B \in S, \quad s_A \neq s_B$$
$$u_A, u_B \in U_1, \quad u_A \neq u_B$$
$$\text{subject}(u_A) \neq \text{subject}(u_B)$$
$$\neg((\text{gender}(s_A) = \text{gender}(s_B) = g) \wedge (\text{subject}(u_A) = \text{subject}(u_B) = b))$$

The answer to template STI_1 is *Yes* if

$$(\text{gender}(s_A) = g \wedge \text{subject}(u_A) = b) \\ \vee (\text{gender}(s_B) = g \wedge \text{subject}(u_B) = b)$$

Let A be the predefined set of answers:

$$A = \{\text{Ag}, \text{Alaska}, \text{Antarctica}, \text{blue}, \text{Denali}, \\ \text{Jupiter}, \text{ostrich}, \text{Saturn}, \text{Shakespeare}, \\ \text{Warsaw}, \text{whale}\}$$

and U_2 the predefined set of questions:

- "What is the largest desert in the world?"
- "What is the largest living species of bird?"
- "What is the highest mountain peak in North America?"
- "What is the largest mammal on Earth?"
- "What planet has the most extensive rings?"
- "What is the largest state in the United States?"
- "What is the chemical symbol for silver?"
- "Who wrote Romeo and Juliet?"
- "What is the capital of Poland?"
- "What is the largest planet in our solar system?"
- "What color is the sky?"

then $\text{answer}(u) \in A$ is a function that returns the answer to the question $u \in U_2$.

In template STI_2, the model must verify if $a \in A$ is an answer to the question asked by a speaker of the given gender. In this case, it must correctly recognize the gender of both speaker s_A and s_B, and if either of them is $g \in G$, answer their question and compare the response to a.

Is a a correct answer to the question asked by a g? $+ s_A(u_A) + s_B(u_B)$ (STI_2)

where

$$s_A, s_B \in S, \quad s_A \neq s_B \\ u_A, u_B \in U_1, \quad u_A \neq u_B \\ \text{subject}(u_A) \neq \text{subject}(u_B) \\ \neg((\text{gender}(s_A) = \text{gender}(s_B) = g) \\ \wedge (\text{answer}(u_A) = \text{answer}(u_B) = a))$$

The answer is *Yes* if

$$(\text{gender}(s_A) = g \wedge \text{answer}(u_A) = a) \\ \vee (\text{gender}(s_B) = g \wedge \text{answer}(u_B) = a)$$

Let U_3 be the predefined set of utterances:

- "Marie Curie was the first woman to win a Nobel Prize, awarded for her work in physics."
- "Ava enjoyed practicing yoga in the morning."
- "Leonardo da Vinci was a polymath and one of the most influential artists of the Renaissance."
- "Charles Darwin was a British naturalist who proposed the theory of evolution through natural selection."
- "Jackson was excited to go on his summer vacation."
- "Olivia loved to spend her free time reading books."
- "Benjamin Franklin was a leading figure in the American Enlightenment, known for his scientific and literary contribution."
- "Cinderella's fairy godmother helped her get ready for the ball."
- "Rapunzel's long hair was used as a ladder by a prince to reach her tower."
- "Rosa Parks was a civil rights activist who sparked the Montgomery Bus Boycott in 1955."

then $\text{historical}(u)$ means that utterance $u \in U_3$ is about a famous historical figure.

Template STI_3 involves verifying whether a speaker of the given gender is talking about a famous historical figure. That means that the model must recognize the gender of both speakers s_A

and s_B, and if either of them is $g \in G$, verify who is the person they are talking about.

Is the person the g is talking about a famous historical figure? $+ s_A(u_A) + s_B(u_B)$ (STI_3)

where

$$s_A, s_B \in S, \quad s_A \neq s_B$$
$$u_A, u_B \in U_3, \quad u_A \neq u_B$$
$$\text{historical}(u_A) \Rightarrow \neg\text{historical}(u_B)$$

In this case, the answer is *Yes* if

$$(\text{gender}(s_A) = g \wedge \text{historical}(u_A)) \vee (\text{gender}(s_B) = g \wedge \text{historical}(u_B))$$

B.6 Sound Reasoning

Let S be the predefined set of sounds:

$$S = \{\text{airplane}, \text{bell}, \text{bird}, \text{bird}_2, \text{cat}, \text{coffee shop}, \text{concert}, \text{dog}, \text{elephant}, \text{football game}, \text{horse}, \text{kids}, \text{motorcycle}, \text{opera}, \text{rain}, \text{sea}, \text{sheep}, \text{thunderstorm}, \text{traffic}, \text{train station}\}$$

then:

- $\text{animal}(s)$ is a function that implies that $s \in S$ is a sound made by an animal
- $\text{bigger}(a_1, a_2)$ is a function that implies that a_1 is bigger than a_2, where $\text{animal}(a_1) \wedge \text{animal}(a_2)$

Template SR_0 requires determining if the animal that makes the sound s is bigger or smaller than the one mentioned in the question.

Is the animal that makes the following sound z than a ? $+ s$ (SR_0)

where

$$a \in \{\text{bird}, \text{cat}, \text{dog}, \text{elephant}, \text{horse}, \text{sheep}\}$$
$$z \in \{\text{bigger}, \text{smaller}\}$$
$$s \in S, \quad \text{animal}(s) \wedge s \neq a$$
$$a = \text{bird} \rightarrow \neg(s = \text{bird}_2)$$
$$a = \text{cat} \rightarrow \neg(s = \text{dog})$$
$$a = \text{dog} \rightarrow \neg(s = \text{cat})$$

The answer to template SR_0 is

$$\text{Yes} \Leftrightarrow (z = \text{bigger} \wedge \text{bigger}(s, a)) \vee (z = \text{smaller} \wedge \text{bigger}(a, s))$$

Template SR_1 involves assessing whether the sounds heard indicate that the weather is nice.

Do the following sounds indicate that the weather is d ? $+ s$ (SR_1)

where

$$d \in \{\text{bad}, \text{nice}\}$$
$$s \in \{\text{bird}, \text{bird}_2, \text{rain}, \text{thunderstorm}\} \subseteq S$$

Let $\text{is\_nice}(s)$ be a function that implies that sound s indicates that the weather is nice. Then, the answer to SR_1 is affirmative if $(d = \text{bad} \wedge \neg\text{is\_nice}(s)) \vee (d = \text{nice} \wedge \text{is\_nice}(s))$.

In template SR_2, the model is required to recognize an animal based on the given sound and determine whether one can ride such animal.

Can you ride an animal that makes this sound? $+ s$ (SR_2)

where $s \in \{\text{bird}, \text{bird}_2, \text{cat}, \text{dog}, \text{elephant}, \text{horse}, \text{sheep}\} \subseteq S$. Let $\text{is\_rideable}(s)$ be a function that implies that an animal making sound s is rideable. Then, the answer is *Yes* if $\text{is\_rideable}(s)$.

Template SR_3 involves assessing whether a thing making the given sound can fly.

Can a thing that makes the following sound fly? $+ s$ (SR_3)

where $s \in \{\text{airplane}, \text{bell}, \text{bird}, \text{bird}_2, \text{cat}, \text{dog}, \text{elephant}, \text{horse}, \text{motorcycle}, \text{sheep}\} \subseteq S$. Let $\text{can\_fly}(s)$ be a function that implies that a thing making the sound s can fly. Then, the answer to SR_3 is *Yes* if $\text{can\_fly}(s)$.

Template SR_4 is similar to SR_1, as it requires the model to decide whether the sounds indicate that one should leave the house.

Should I leave the house now? $+ s$ (SR_4)

where $s \in \{\text{bird}, \text{bird}_2, \text{rain}, \text{thunderstorm}\} \subseteq S$. The answer is affirmative if $\text{is\_nice}(s)$.

For template SR_5, the model must determine whether the given sound can be heard at the given place. Let P be the predefined set of places:

$$P = \{\text{at the airport, at the café, at the concert, at the opera, at the school, at the stable, at the stadium, at the train station, at the zoo, by the sea, in the forest, in the street}\}$$

then $\text{heard\_at}(s, p)$ implies that sound $s \in S$ can be heard at place $p \in P$.

Can the following sound be heard p ? $+ s$ (SR_5)

where $s \in \{\text{airplane, bird, coffee shop, concert, elephant, football game, horse, kids, opera, sea, traffic, train station}\} \subseteq S$.

Template SR_6 requires determining if the vehicle that makes the given sound is bigger or smaller than the one mentioned in the question.

Is the vehicle making the following sound z than v? $+ s$ (SR_6)

where

$$z \in \{\text{bigger}, \text{smaller}\}$$
$$v \in \{\text{airplane}, \text{car}, \text{motorcycle}\}$$
$$s \in \{\text{airplane}, \text{motorcycle}\} \subseteq S, \quad s \neq v$$

Let $\text{bigger}(v_1, v_2)$ be the function that implies that v_1 is bigger than v_2. Then, the answer to SR_6 is *Yes* if $(z = \text{bigger} \wedge \text{bigger}(s, v)) \vee (z = \text{smaller} \wedge \text{bigger}(v, s))$.

In template SR_7 the model must recognize an animal based on the sound and determine whether it is a mammal. Let $\text{class}(s) \in \{\text{aves}, \text{mammals}\}$ be a function that returns a biological class to which belongs the animal making the sound $s \in \{s' \in S \mid \text{animal}(s)\}$.

Is the animal that makes the following sound a mammal? $+ s$ (SR_7)

where $s \in \{s' \in S \mid \text{animal}(s')\}$. The answer is affirmative if $\text{class}(s) = \text{mammals}$.

Template SR_8 requires similar reasoning, but there are two sounds to recognize.

Are both animals that make the following sounds mammals? $+ s_A + s_B$ (SR_8)

where $s_A, s_B \in \{s' \in S \mid \text{animal}(s')\}$, $s_A \neq s_B$. Then, the answer is *Yes* if $\text{class}(s_A) = \text{mammals} \wedge \text{class}(s_B) = \text{mammals}$.

Similarly, template SR_9 involves verifying whether two animals are from the same biological class based on the sounds they make.

Are both animals that make the following sounds from the same biological class? $+ s_A + s_B$ (SR_9)

where $s_A, s_B \in \{s' \in S \mid \text{animal}(s')\}$, $s_A \neq s_B$. The answer is *Yes* if $\text{class}(s_A) = \text{class}(s_B)$.

Template SR_10 is requires recognition of two animals based on the sounds they make and determining whether both of them are bigger or smaller than the given animal.

Are both animals that make the following sounds z than a? $+ s_A + s_B$ (SR_10)

where

$$z \in \{\text{bigger}, \text{smaller}\}$$
$$a \in \{\text{bird}, \text{cat}, \text{dog}, \text{elephant}, \text{horse}, \text{sheep}\}$$
$$s_A, s_B \in \{s' \in S \mid \text{animal}(s')\}$$
$$s_A \neq s_B \neq a$$

The answer is *Yes* if

$$(z = \text{bigger} \wedge \text{bigger}(s_A, a) \wedge \text{bigger}(s_B, a)) \vee (z = \text{smaller} \wedge \text{bigger}(a, s_A) \wedge \text{bigger}(a, s_B))$$

In template SR_11 the model must assess whether both animals making the given sounds are rideable.

Can you ride both animals that make these sounds? $+ s_A + s_B$ (SR_11)

where $s_A, s_B \in \{s' \in S \mid \text{animal}(s')\}$, $s_A \neq s_B$. The answer is affirmative if $\text{is\_rideable}(s_A) \wedge \text{is\_rideable}(s_B)$.

Similarly, template SR_12 requires assessing if both things making the given sounds can fly.

Can both things that make the following sounds fly? $+ s_A + s_B$ (SR_12)

where $s_A, s_B \in \{\text{airplane, bell, bird, bird}_2\text{, cat, dog, elephant, horse, motorcycle, sheep}\} \subseteq S$, $s_A \neq s_B$. The answer to SR_12 is *Yes* if $\text{can\_fly}(s_A) \wedge \text{can\_fly}(s_B)$.

Templates SR_13 and SR_14 involve verifying whether given sounds can be heard indoors or outdoors.

Can this sound be usually heard w? $+ s$ (SR_13)

Can both these sounds be usually heard w? $+ s_A + s_B$ (SR_14)

where

$$w \in \{\text{indoors}, \text{outdoors}\}$$
$$s, s_A, s_B \in S, \quad s_A \neq s_B$$

The answer to SR_13 is affirmative if $\text{heard\_at}(s, w)$, and to SR_14 if $\text{heard\_at}(s_A, w) \wedge \text{heard\_at}(s_B, w)$.

B.7 Speech Features Comparison

Let S, C, G, and A be the predefined sets of speakers, accents, genders, and ages, respectively:

$$S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$$
$$C = \{\text{Australian}, \text{English}, \text{Scottish}\}$$
$$G = \{\text{female}, \text{male}\}$$
$$A = \{\text{fifties}, \text{twenties}\}$$

and U the predefined set of utterances:

- "The coffee smells amazing this morning."
- "I can't believe how bright the moon is tonight."
- "The birds chirped a gentle melody."

then:

- $s(u)$ means that utterance $u \in U$ is spoken by speaker $s \in S$
- $\text{accent}(s) \in C$ is a function that returns the accent of speaker $s \in S$
- $\text{gender}(s) \in G$ is a function that returns the gender of speaker $s \in S$
- $\text{age}(s) \in A$ is a function that returns the age of speaker $s \in S$

Template SFC_0 requires determining whether the second utterance is the same as the first one, but read with the given accent.

Is the second recording the same text but read with c accent? $+ s_A(u_A) + s_B(u_B)$ (SFC_0)

where

$$c \in C$$
$$s_A, s_B \in S, \quad s_A \neq s_B$$
$$u_A, u_B \in U$$
$$\text{accent}(s_A) = c \Rightarrow \neg(\text{accent}(s_B) = c)$$

The answer is affirmative if $u_A = u_B \wedge \text{accent}(s_B) = c$.

Template SFC_1 involves verifying whether the second utterance is the same as the first one, but read with the given gender's voice.

Is the second recording the same text as the first recording but spoken by a g voice? $+ s_A(u_A) + s_B(u_B)$ (SFC_1)

where

$$g \in G$$
$$s_A, s_B \in S, \quad s_A \neq s_B$$
$$u_A, u_B \in U$$
$$\text{gender}(s_A) = g \Rightarrow \neg(\text{gender}(s_B) = g)$$

The answer is *Yes* if $u_A = u_B \wedge \text{gender}(s_B) = g$.

In template SFC_2, the model must recognize the gender of both speakers and verify if it is the same.

Are both speakers the same gender? $+ s_A(u_A)$ $+ s_B(u_B)$ (SFC_2)

where $s_A, s_B \in S$, $s_A \neq s_B$; $u_A, u_B \in U$, $u_A \neq u_B$. Then, the answer is *Yes* if $\text{gender}(s_A) = \text{gender}(s_B)$.

Similarly, template SFC_3 requires recognition of the age of both speakers and verifying if it the same.

Are both speakers the same age? $+ s_A(u_A)$ $+ s_B(u_B)$ (SFC_3)

where $s_A, s_B \in S$, $s_A \neq s_B$; $u_A, u_B \in U$, $u_A \neq u_B$. Then, the answer is *Yes* if $\text{age}(s_A) = \text{age}(s_B)$.

B.8 Text and Sound Reasoning

Let S_P, S_A, S_O be the predefined sets of speakers, sounds of animals, and other sounds respectively:

$$S_P = \{s_1, s_2, s_3, s_4\}$$
$$S_A = \{\text{bird}, \text{bird}_2, \text{cat}, \text{dog}, \text{elephant}, \text{horse}, \text{sheep}\}$$
$$S_O = \{\text{airplane}, \text{bell}, \text{kids}, \text{motorcycle}, \text{opera}, \text{rain}, \text{thunderstorm}, \text{sea}, \text{concert}, \text{traffic}\}$$

U_A the predefined set of utterances about animals:

- "The bird sings a cheerful melody at dawn."
- "The cat purs softly on the windowsill."
- "The dog wagged its tail happily."
- "The elephant drinks water with its long trunk."
- "A horse gallops across the open field."
- "The sheep grazed peacefully in the green meadow."

and U_O the predefined set of other utterances:

- "The airplane soars high above the clouds."
- "The bell rings, echoing through the hallway."
- "The kids laughed loudly as they ran around playing tag."
- "A motorcycle is a fast vehicle."
- "Opera singing tells dramatic stories."
- "Rain taps gently against the glass."
- "A thunderstorm rumbles loudly in the distance."
- "The waves crash against the shore."
- "A concert is a live performance of music."
- "The street was bustling with honking cars."

then:

- $s_p(u)$ means that utterance $u \in U_A \cup U_O$ is spoken by speaker $s_p \in S$
- $\text{about}(u)$ is a function that returns the main topic of the utterance $u \in U_A \cup U_O$

Template TSR_0 requires determining whether the spoken utterance is about the following sound.

Is the person talking about the following sound? $+ s_p(u_o) + s_o$ (TSR_0)

where $s_p \in S_P$, $u_o \in U_O$, and $s_o \in S_O$. The answer is *Yes* if $\text{about}(u_o) = s_o$.

Similarly, template TSR_1 involves verifying whether the spoken utterance is about an animal making the following sound.

Is the person talking about an animal making the following sound? $+ s_p(u_a) + s_A$ (TSR_1)

where $s_p \in S_P$, $u_a \in U_A$, and $s_a \in S_A$. The answer is *Yes* if $\text{about}(u_a) = s_a$.

In template TSR_2, the model is required to recognize two animals based on an utterance and a sound, and then assess whether the second is smaller or bigger. Let $\text{bigger}(a_1, a_2)$ be a function that implies that animal a_1 is bigger than a_2.

Is the animal the person is talking about z than the one making the sound? $+ s_p(u_a)$ $+ s_a$ (TSR_2)

where

$$
\begin{aligned}
&z \in \{\text{bigger}, \text{smaller}\} \\
&s_p \in S_P \\
&u_a \in U_A, \quad s_a \in S_A, \quad \text{about}(u_a) \neq s_a \\
&s \neq \begin{cases} \text{bird}_2, & \text{if about}(u_a) = \text{bird}, \\ \text{bird}, & \text{if about}(u_a) = \text{bird}_2, \\ \text{dog}, & \text{if about}(u_a) = \text{cat}, \\ \text{cat}, & \text{if about}(u_a) = \text{dog}. \end{cases}
\end{aligned}
$$

The answer is *Yes* if

$$
\begin{aligned}
&(z = \text{bigger} \wedge \text{bigger}(\text{about}(u_A), s_A)) \\
&\vee (z = \text{smaller} \wedge \text{bigger}(s_A, \text{about}(u_A)))
\end{aligned}
$$

Template TSR_3 requires recognition of animals based on utterance and sound, and determining whether both of them are mammals. Let $\text{class}(a) \in \{\text{aves}, \text{mammals}\}$ be a function that returns a biological class to which an animal a belongs.

Are both the animal the person is talking about and the one making the sound mammals? $+ s_p(u_a) + s_a$ (TSR_3)

where

$$
\begin{aligned}
&s_p \in S_P \\
&u_a \in U_A, \quad s_a \in S_A, \quad \text{about}(u_a) \neq s_a \\
&\text{about}(u_a) = \text{bird} \Rightarrow s_a \neq \text{bird}_2
\end{aligned}
$$

The answer is affirmative if $\text{class}(\text{about}(u_a)) = \text{mammals} \wedge \text{class}(s_a) = \text{mammals}$.

Similarly, template TSR_4 involves determining whether both animals are from the same biological class.

Are both the animal the person is talking about and the one making the sound from the same biological class? $+ s_p(u_a) + s_a$ (TSR_4)

where

$$
\begin{aligned}
&s_p \in S_P \\
&u_a \in U_A, \quad s_a \in S_A, \quad \text{about}(u_a) \neq s_a \\
&\text{about}(u_a) = \text{bird} \Rightarrow s_a \neq \text{bird}_2
\end{aligned}
$$

The answer is affirmative if $\text{class}(\text{about}(u_a)) = \text{class}(s_a)$.

In template TSR_5, the model must recognize animals based on utterance and sound, and verify if one can ride both of them. Let $\text{is\_rideable}(a)$ be a function that implies that animal a is rideable.

Can you ride both the animal the person is talking about and the one making the sound? $+ s_p(u_a) + s_a$ (TSR_5)

where

$$
\begin{aligned}
&s_p \in S_P \\
&u_a \in U_A, \quad s_a \in S_A, \quad \text{about}(u_a) \neq s_a \\
&\text{about}(u_a) = \text{bird} \Rightarrow s_a \neq \text{bird}_2
\end{aligned}
$$

The answer is affirmative if $\text{is\_rideable}(\text{about}(u_a)) \wedge \text{is\_rideable}(s_a)$.

Template TSR_6 involves recognition of two objects or animals, and determining whether both of them can fly. Let $\text{can\_fly}(o)$ be a function that object or animal o can fly.

Can both the thing the person is talking about and the one making the sound fly? $+ s_p(u) + s$ (TSR_6)

where

$$
\begin{aligned}
&s_p \in S_P \\
&u \in U' \subseteq U_A \cup U_O \\
&\text{about}(u) \neq s \\
&\text{about}(u) = \text{bird} \Rightarrow s \neq \text{bird}_2 \\
&s \in \left\{ \begin{array}{l} \text{bird}, \text{bird}_2, \text{cat}, \text{dog}, \\ \text{elephant}, \text{horse}, \text{sheep}, \\ \text{airplane}, \text{bell}, \text{motorcycle} \end{array} \right\} \subseteq S_A \cup S_O
\end{aligned}
$$

and U' consists of:

- "The bird sings a cheerful melody at dawn."
- "The cat purs softly on the windowsill."
- "The dog wagged its tail happily."
- "The elephant drinks water with its long trunk."

- "A horse gallops across the open field."
- "The sheep grazed peacefully in the green meadow."
- "The airplane soars high above the clouds."
- "The bell rings, echoing through the hallway."
- "A motorcycle is a fast vehicle."

The answer is affirmative if $\text{can\_fly}(\text{about}(u)) \wedge \text{can\_fly}(s)$.

Template TSR_7 requires assessing whether both objects, events or animals recognized based on utterance and sound, can be heard indoors or outdoors. Let $\text{heard\_at}(s, p)$ be a function that implies that object, event or animal s can be heard at place p.

Can both the thing the person is talking about and the one making the sound be usually heard p ? $+ s_p(u) + s$ (TSR_7)

where

$$
\begin{aligned}
& p \in \{\text{indoors}, \text{outdoors}\} \\
& s_p \in S_P \\
& u \in U_A \cup U_O \\
& s \in S_A \cup S_O \\
& \text{about}(u) \neq s \\
& \text{about}(u) = \text{bird} \Rightarrow s \neq \text{bird}_2
\end{aligned}
$$

The answer to TSR_7 is *Yes* if $\text{heard\_at}(s, p) \wedge \text{heard\_at}(\text{about}(u), p)$.

B.9 Text and Temporal Localization Reasoning

Let Sp, P, Sn be the predefined sets of speakers, places, and sounds, respectively:

$$
\begin{aligned}
Sp &= \{s_1, s_2, s_3, s_4\} \\
P &= \{\text{at the school}, \text{at the café}, \\
&\qquad \text{at the train station}, \text{at the concert}\} \\
Sn &= \{\text{kids}, \text{coffee shop}, \text{train station}, \text{concert}, \\
&\qquad \text{football game}, \text{traffic}, \text{sea}, \text{Christmas}\}
\end{aligned}
$$

and D_0 the predefined set of dialogues:

- "Did you understand what she meant by that last part?"
  "Not really. I'll need to check my notes later."
- "Do you want to sit by the windows or over here?"
  "Let's sit here. It's quieter."
- "Did you see that lighting effect during the last part?"
  "Yeah, it was incredible - it matched the rhythm perfectly!"
- "How long do we have to wait now?"
  "Just a few more minutes, I think."

then:

- $d(s_A, s_B)$ means that the first utterance in dialogue $d \in D_0 \cup D_2 \cup D_3$ is spoken by the speaker $s_A \in Sp$, and the second by the speaker $s_B \in Sp$
- $\text{place}(x) \in P$ is a function that returns a place where $x \in Sn \cup D_0 \cup D_2 \cup D_3 \cup U$ can be heard
- $\frac{d(s_A, s_B)}{n}$ means that the background sound $n \in Sn$ is added to the dialogue $d(s_A, s_B)$; $d \in D_0 \cup D_2 \cup D_3$; $s_A, s_B \in Sp$

Template TTLR_0 requires recognizing whether a conversation takes place in the given spot based on dialogue and background sounds.

Does the following conversation take place p ? $+ \dfrac{d(s_A, s_B)}{n}$ (TTLR_0)

where

$$
\begin{aligned}
& p \in P \\
& d \in D_0 \\
& s_A, s_B \in Sp, \quad s_A \neq s_B \\
& n \in \left\{ \begin{matrix} \text{kids}, \text{coffee shop}, \\ \text{concert}, \text{train station} \end{matrix} \right\} \subseteq Sn
\end{aligned}
$$

The answer is affirmative if $p = \text{place}(d) = \text{place}(n)$.

Let U be the predefined set of utterances:

- "The energy here is unreal, and the players are giving it their all!"
- "This street is packed with honking cars and bustling crowds!"

- "The waves crash against the shore, and the salty breeze is so refreshing!"

Template TTLR_1 involves assessing whether a speaker is describing the acoustic scene that they are in.

Does the speaker describe the acoustic scene that they are in? $+ \dfrac{s(u)}{n}$ (TTLR_1)

where

$$s \in Sp, \quad u \in U$$
$$n \in \{\text{football game}, \text{traffic}, \text{sea}\} \subseteq Sn$$

The answer is *Yes* if $\text{place}(u) = \text{place}(n)$.

Let D_2 be the predefined set of dialogues for template TTLR_2:

- "Did you remember to pack sunscreen?"
  "Yeah, and I grabbed some snacks for the trip too."
- "Did you get everything on your list?"
  "Almost, but I still need to wrap a few things."

then $\text{time}(x) \in \{\text{Christmas}, \text{summer}\}$ is a function that returns a time when $x \in Sn \cup D_2$ can be heard.

In template TTLR_2, the model is required to recognize the time of the year based on dialogue and background sound, and verify if it matches the given time.

Based on the sounds and conversation, is it t now? $+ \dfrac{d(s_A, s_B)}{n}$ (TTLR_2)

where

$$t \in \{\text{Christmas}, \text{summer}\}$$
$$d \in D_2$$
$$s_A, s_B \in Sp, \quad s_A \neq s_B$$
$$n \in \{\text{Christmas}, \text{sea}\} \subseteq Sn$$

In this case, the answer is *Yes* if $\text{time}(d) = \text{time}(n) = t$.

Let D_3 be the predefined set of dialogues for TTLR_3:

- "Hey, are you still coming over later?"
  "Yeah, I'll be there around six - traffic's a bit slow right now."
- "Do you want to grab something to eat later?"
  "Sure, how about that new place downtown?"
- "Have you seen my phone anywhere?"
  "I think you left it on the kitchen counter."
- "Did you remember to water the plants today?"
  "Oh no, I completely forgot!"
- "Did you hear about the new café downtown?"
  "No, what's it like?"
- "Did you finish the group project?"
  "Almost, I just need to finalize the slides."

Template TTLR_3 involves verifying if the acoustic environment of two speakers is the same.

Is the acoustic environment of both people the same? $+ d\left(\dfrac{s_A}{n_A}, \dfrac{s_B}{n_B}\right)$ (TTLR_3)

where

$$d \in D_3$$
$$s_A, s_B \in Sp, \quad s_A \neq s_B$$
$$n_A, n_B \in Sn$$

For TTLR_3, the answer is affirmative if $n_A = n_B$.

C Human Evaluation Results

Table 6 presents the accuracy per task achieved by three evaluators (**A**, **B**, and **C**) on the ART-H subset of the benchmark.

Each evaluator was presented with 216 prompts, to which they responded with a *Yes* or *No* answer. Overall, the tasks did not pose a challenge to any of the evaluators. The highest accuracy, 98.6%, was achieved for **Audio Arithmetics**, **Selective Text Inference**, and **Text and Temporal Localization Reasoning**. Tasks requiring recognition of accents or languages (i.e., **Cross-Recording Language Identification** and **Speech Features Comparison**) resulted in the lowest accuracies: 80.6% and 73.6%, respectively.

All evaluators achieved a total accuracy of at least 90%, resulting in a human baseline of 92.9%.

Listing 1: System prompt used for the LLM-as-a-judge approach.

```
You will receive a question, an expected answer (which can always be reduced to "
    Yes" or "No"), and an actual answer that was given. Your task is to check
    wheter the given answer is relevant to the question. If it is relevant,
    respond with "Relevant" on the first line, followed by another line stating
    whether the given answer is correct or incorrect compared to the expected
    answer. If it is irrelevant, respond with "Irrelevant" on the first line,
    followed by an explanation on the next line.
Example Input:
Question: Are there as many dog barks as there are bell rings?
Expected Answer: No
Given Answer: There are more dog barks than bell rings.
Example Output:
Relevant
Incorrect
Example Input:
Question: Is one of the recordings a quieter version of the other one?
Expected Answer: Yes
Given Answer: Yes, the second recording is a shortened version of the first one.
Example Output:
Irrelevant
The answer does not address the question about the quieter audio.
```

Table 6: Accuracy of the human evaluation of the ART benchmark.

	A	B	C	Average
AA	1.000	0.958	1.000	0.986
ATD	1.000	0.958	0.917	0.958
CRLI	0.958	0.625	0.833	0.806
CRSI	1.000	0.917	1.000	0.972
STI	1.000	0.958	1.000	0.986
SR	1.000	0.958	0.958	0.972
SFC	0.750	0.750	0.708	0.736
TSR	0.958	1.000	0.917	0.958
TTLR	1.000	1.000	0.958	0.986
Total	0.963	0.903	0.921	0.929

D Experimental Setup

Table 7 presents the parameters used during inference. Each model was prompted using the default settings suggested by the authors.

The system prompt used for the evaluation with LLM as a judge is shown in Listing 1. The same prompt was used for both Llama 3.3 and Qwen3.

Listing 2: General text prompt used for the inference in *Yes/No* experiments.

```
Answer the question from the audio.
Answer only "Yes" or "No".
```

Listing 3: General text prompt used for the inference in *Descriptive* experiments.

```
Answer the question from the audio.
```

The general text prompt used for the experiments with *Yes/No* approach is shown in Listing 2, and with *Descriptive* approach in Listing 3

The GAMA model required modification of the general prompt to improve the quality of the generated results. The prompts used in this case are shown in Listings 4 and 5.

Listing 4: Text prompt used for the inference on GAMA model in *Yes/No* experiments.

```
You will receive an audio sample
containing a spoken question. Your
task is to provide a concise and
accurate answer. Answer only "Yes" or
"No".
```

Listing 5: Text prompt used for the inference on GAMA model in *Descriptive* experiments.

```
You will receive an audio sample
containing a spoken question. Your
task is to provide a concise and
accurate answer. Answer only the
question contained in the audio
without adding unnecessary details.
```

E Analysis of the results of experiments with the *Descriptive* approach

Table 8 shows the accuracy of each model on each of the nine tasks as evaluated by Llama 3.3 as a judge.

Overall, the **Speech Features Comparison** and **Text and Sound Reasoning** tasks have the highest average accuracy scores, suggesting that these tasks are easier for the models. In contrast, **Audio Transformation Detection**, **Cross-Recording**

Table 7: Inference parameters.

Model	Parameters		
Whisper Large v3	max_length: 448 return_timestamps: false		
Llama 3.3	max_model_len: 2048 presence_penalty: 0.0 frequency_penalty: 0.0	repetition_penalty: 1.0 temperature: 0.6 top_p: 0.9	top_k: 0 min_p: 0.0
Qwen3	enable_thinking: false max_model_len: 2048 presence_penalty: 0.0	frequency_penalty: 0.0 repetition_penalty: 1.0 temperature: 0.6	top_p: 0.95 top_k: 20 min_p: 0.0
Audio Flamingo 3	max_length: 20 length_penalty: 1.0	repetition_penalty: 1.0 temperature: 1.0	top_k: 50 top_p: 1.0
GAMA	max_new_tokens: 400 do_sample: true	temperature: 0.1 top_p: 0.95	top_k: 500
Qwen-Audio-Chat	max_new_tokens: 512 do_sample: true	top_p: 0.5 top_k: 0	
Qwen2-Audio	max_model_len: 2048 presence_penalty: 0.0 frequency_penalty: 0.0	repetition_penalty: 1.0 temperature: 0.7 top_p: 0.5	top_k: 20 min_p: 0.0
Ultravox v0.4.1	max_model_len: 2048 presence_penalty: 0.0 frequency_penalty: 0.0	repetition_penalty: 1.0 temperature: 1.0 top_p: 1.0	top_k: 0 min_p: 0.0
Ultravox v0.6	max_model_len: 2048 presence_penalty: 0.0 frequency_penalty: 0.0	repetition_penalty: 1.0 temperature: 1.0 top_p: 1.0	top_k: 0 min_p: 0.0

Table 8: Absolute accuracy per task on the ART benchmark using *Descriptive* approach and Llama 3.3 as a judge.

Model	AA	ATD	CRLI	CRSI	STI	SR	SFC	TSR	TTLR
Whisper + Llama	0.158	0.157	0.126	0.275	**0.503**	0.240	0.603	0.591	**0.522**
Whisper + Qwen	**0.451**	**0.390**	**0.268**	**0.394**	0.446	**0.609**	**0.671**	**0.646**	0.516
Audio Flamingo 3	0.046	0.015	0.042	0.026	0.031	0.061	0.053	0.034	0.043
GAMA	0.000	0.000	0.000	0.000	0.000	0.020	0.006	0.051	0.003
Qwen-Audio-Chat	0.052	0.011	0.054	0.000	0.003	0.136	0.261	0.149	0.034
Qwen2-Audio (zero-shot)	0.439	0.382	0.496	0.385	**0.501**	0.453	0.481	0.460	0.443
Qwen2-Audio (one-shot, same template)	0.425	**0.397**	**0.513**	**0.397**	0.480	**0.560**	0.492	0.525	**0.472**
Qwen2-Audio (one-shot, different template)	**0.446**	0.301	0.494	0.374	0.482	0.466	0.500	**0.559**	0.426
Ultravox v0.4.1	0.158	0.022	0.319	0.220	0.281	0.126	0.436	0.346	0.250
Ultravox v0.6	0.405	0.352	0.286	0.176	0.372	0.394	**0.566**	0.540	0.462
Average	0.281	0.224	0.284	0.247	0.341	0.334	**0.446**	**0.430**	0.347

Table 9: Absolute accuracy per task on the ART benchmark using *Descriptive* approach and Qwen3 as a judge.

Model	AA	ATD	CRLI	CRSI	STI	SR	SFC	TSR	TTLR
Whisper + Llama	0.125	0.008	0.128	0.491	**0.559**	0.233	**0.639**	0.724	0.676
Whisper + Qwen	**0.660**	**0.512**	**0.280**	**0.513**	0.466	**0.645**	0.626	**0.728**	**0.697**
Audio Flamingo 3	0.054	0.018	0.042	0.026	0.033	0.057	0.049	0.032	0.042
GAMA	0.000	0.000	0.000	0.000	0.000	0.004	0.087	0.045	0.003
Qwen-Audio-Chat	0.065	0.028	0.037	0.000	0.010	0.088	0.112	0.125	0.028
Qwen2-Audio (zero-shot)	0.542	**0.353**	**0.604**	0.373	**0.502**	0.492	0.490	0.411	0.448
Qwen2-Audio (one-shot, same template)	0.541	0.344	0.522	0.426	0.474	**0.579**	0.527	0.517	0.566
Qwen2-Audio (one-shot, different template)	**0.554**	0.307	0.544	**0.463**	0.491	0.506	0.477	0.532	0.477
Ultravox v0.4.1	0.100	0.003	0.197	0.186	0.205	0.064	0.290	0.228	0.250
Ultravox v0.6	0.521	0.340	0.260	0.207	0.427	0.399	**0.587**	**0.615**	**0.626**
Average	0.345	0.211	0.286	0.296	0.348	0.334	**0.426**	**0.436**	0.419

Speaker Identification, and **Audio Arithmetics** have the lowest average accuracy scores, indicating greater ambiguity or difficulty.

Performance varies considerably across tasks at the model level. Several models demonstrate task-specific strengths, achieving high accuracy on certain tasks but performing poorly on others. This indicates limited generalization across task types.

Table 9 shows the accuracy obtained by each model as evaluated by Qwen3 as a judge.

Once again **Speech Features Comparison** and **Text and Sound Reasoning** have the highest average accuracies, with **Text and Temporal Localization Reasoning** close behind. In contrast, **Audio Transformation Detection** and **Cross-Recording Language Identification** have the lowest average accuracy scores.

Overall, evaluation with Qwen3 as a judge reinforces the presence of systematic differences in task difficulty. These results largely mirror the trends observed with Llama 3.3 as a judge, while also highlighting differences in absolute accuracy levels across tasks that depend on the method of evaluation.

F Human-LLM Alignment

The level of agreement among a human judge and two LLM-based judges was evaluated on the ART-H subset of the benchmark. All of the systems were evaluated using the *Descriptive* approach. Table 10 shows the results, which measure agreement as the percentage of aligned judge decisions.

Overall, the agreement between the two LLM judges was consistently higher than the agreement between either LLM judge and the human evaluator. The Llama 3.3-Qwen3 alignment ranged from 68.06% to 98.61%, whereas the Human-Llama 3.3 agreement ranged from 47.22% to 99.54%, and the Human-Qwen3 agreement ranged from 54.17% to 98.15%. These results suggest that the two LLM judges have more similar evaluation criteria than the human judge.

There were significant model-level differences. The GAMA model exhibited near-perfect agreement across all judge pairings ($\geq$ 98%), indicating strong consistency and minimal ambiguity in evaluation outcomes. However, it should be noted that the human evaluator deemed all of this model's responses irrelevant. In contrast, the Ultravox v0.4.1 model showed substantially lower agreement with the human judge (47.22% with Llama 3.3 and 54.17% with Qwen3), despite moderate agreement between the two LLM judges (68.06%). This difference stems from the fact that this model tended to provide speculative answers based on its knowledge of the real world.

Qwen2-Audio, Ultravox v0.6 and both cascaded systems exhibited intermediate agreement levels, where the alignment between LLM judges remained relatively high (73.15-84.72%), while human-LLM agreement was more variable. Notably, Audio Flamingo 3 and Qwen-Audio-Chat achieved comparatively strong human-LLM alignment ($\geq$ 75%), suggesting closer alignment between automated and human evaluation in these cases.

In summary, the results suggest that LLM judges are consistent within themselves but not aligned with human judgments. While some models

Table 10: Judgment alignment between human, Llama 3.3, and Qwen3.

Model	Human-Llama 3.3	Human-Qwen3	Llama 3.3-Qwen3
Whisper + Llama	65.28%	60.65%	73.15%
Whisper + Qwen	72.22%	61.11%	79.63%
Audio Flamingo 3	78.70%	75.00%	75.93%
GAMA	99.54%	98.15%	98.61%
Qwen-Audio-Chat	87.50%	80.09%	85.19%
Qwen2-Audio (zero-shot)	69.44%	62.04%	82.87%
Qwen2-Audio (one-shot, same template)	80.56%	71.76%	84.72%
Qwen2-Audio (one-shot, different template)	72.22%	64.35%	80.09%
Ultravox v0.4.1	47.22%	54.17%	68.06%
Ultravox v0.6	72.69%	61.57%	78.70%

demonstrate strong agreement, others reveal substantial divergence. This underscores the importance of model-specific analysis and careful interpretation of LLM-based evaluation outcomes.

G Error Analysis

G.1 *Yes/No* approach

Table 11 presents an error analysis for the *Yes/No* approach based on human evaluation conducted on the ART-H subset of the benchmark.

The most common errors across models were related to failures in understanding or processing the audio input. Several models often failed to recognize speech or sound altogether. For example, GAMA model exhibites this behavior in 47.83% of erroneous cases, while Ultravox v0.4.1 exhibited this behavior in an even higher percentage of cases, at 66.67%. In Qwen2-Audio with zero-shot approach, failure to recognize speech or sound accounted for half of the observed errors. These results suggest limitations in audio perception or interpretation for these models.

Another common error was providing responses that were unrelated to the task. These responses constituted 50% of errors for Qwen2-Audio with zero-shot approach, 27.59% with one-shot approach and example from the same template, 24.05% with one shot-approach and example from a different template, and 33.33% for Ultravox v0.4.1 model. This pattern suggests that when the models failed to interpret the input properly, they often relied on unconstrained generation rather than explicitly signaling uncertainty.

Some models exhibited systematic, yet task-inappropriate, behaviors. For example, Qwen-Audio-Chat consistently returned a transcription of the audio instead of answering the question, accounting for 100% of its errors. This indicates a significant bias toward speech-to-text functionality that potentially overrides the intended question-answering objective. Similarly, Qwen2-Audio with one-shot approach, frequently produced responses identifying the speaker (44.83% and 35.44%), indicating confusion between speaker recognition and the intended task.

Table 11: Error analysis of the model responses for *Yes/No* approach based on human evaluation.

GAMA	
Did not recognize the question	51.30%
Did not recognize speech or sound	47.83%
Random	0.87%
Qwen-Audio-Chat	
Transcription	100.00%
Qwen2-Audio (zero-shot)	
Did not recognize speech or sound	50.00%
Random	50.00%
Qwen2-Audio (one-shot, same template)	
Speaker recognition	44.83%
Random	27.59%
Responded in different language	10.34%
Did not recognize speech or sound	10.34%
Cannot help	6.90%
Qwen2-Audio (one-shot, different template)	
Speaker recognition	35.44%
Random	24.05%
Responded in different language	21.52%
Did not recognize speech or sound	16.46%
Returned timestamps	2.53%
Ultravox v0.4.1	
Did not recognize speech or sound	66.67%
Random	33.33%

Language-related errors were also observed.

Qwen2-Audio with one-shot approach responded in a different language 10.34% and 21.52% of the time. This highlights inconsistencies in language control under failure conditions. Other notable errors were less frequent but still significant, including an explicit refusal or inability to help (6.90% for the same template) and returning timestamps instead of answers (2.53% for a different template).

Overall, the error analysis reveals that failures in *Yes/No* approach are primarily due to misinterpretation or non-recognition of the audio input and systematic task confusion, in which the models default to transcription, speaker identification, or an unrelated generation.

Audio Flamingo 3, Ultravox v0.6, and the cascaded systems were excluded from this analysis, as they produced nearly 100% relevant answers across all 9 000 samples of the ART benchmark.

G.2 *Descriptive* approach

Tables 12 and 13 provide an error analysis for the *Descriptive* approach, which is based on human evaluation conducted on the ART-H subset of the benchmark.

Table 12: Error analysis of the cascaded systems' responses for *Descriptive* approach based on human evaluation.

Whisper + Llama	
Cannot help	44.65%
Asked for transcription	30.19%
Did not recognize the question	15.09%
Responded in different language	9.43%
Speculative answer	0.63%
Whisper + Qwen	
Did not recognize the question	49.32%
Cannot help	30.14%
Responded in different language	20.55%

A common error for the *Descriptive* approach was failing to recognize the intent of the question. This behavior was particularly common among Whisper + Qwen (49.32%) and Ultravox v0.6 (45.36%). In many cases, this lack of recognition resulted in fallback behaviors rather than an outright refusal to complete the task.

Another common error was models providing transcriptions instead of answering the question. This was particularly evident with Audio Flamingo 3 and Qwen-Audio-Chat, which returned transcriptions 60.82% and 78.65% of the time, respectively.

Table 13: Error analysis of the AudioLLMs' responses for *Descriptive* approach based on human evaluation.

Audio Flamingo 3	
Transcription	60.82%
Random	21.65%
Did not recognize the question	17.53%
GAMA	
Random	71.76
Did not recognize the question	28.24
Qwen-Audio-Chat	
Transcription	78.65%
Did not recognize the question	12.35%
Random	8.99%
Qwen2-Audio (zero-shot)	
Responded in different language	61.11%
Random	17.13%
Speaker recognition	10.65%
Political matters	6.48%
Cannot help	2.78%
Returned timestamps	1.39%
Did not recognize speech or sound	0.46%
Qwen2-Audio (one-shot, same template)	
Speaker recognition	41.63%
Random	35.41%
Responded in different language	8.13%
Did not recognize speech or sound	7.18%
Cannot help	4.31%
Returned timestamps	3.35%
Qwen2-Audio (one-shot, different template)	
Random	37.74%
Speaker recognition	36.32%
Responded in different language	10.38%
Did not recognize speech or sound	7.08%
Returned timestamps	4.72%
Cannot help	3.77%
Ultravox v0.4.1	
Speculative answer	48.19%
Described the prompt	30.57%
Respondend in different language	10.88%
Cannot help	9.33%
Random	1.04%
Ultravox v0.6	
Did not recognize the question	45.36%
Cannot help	35.05%
Speculative answer	13.40%
Transcription	3.09%
Responded in different language	2.06%
Random	1.03%

Table 14: Results of the random sampling analysis using *Yes/No* approach.

Model	Relevant			Absolute Accuracy		
	mean	min	max	mean	min	max
Whisper + Llama	99.92 ± 0.19	98.89	100.00	0.5410 ± 0.0307	0.4704	0.6185
Whisper + Qwen	99.92 ± 0.10	99.63	100.00	0.5619 ± 0.0248	0.5102	0.6194
Audio Flamingo 3	100.00 ± 0.00	100.00	100.00	0.5446 ± 0.0301	0.4537	0.6204
GAMA	42.26 ± 1.62	38.15	46.30	0.2144 ± 0.0137	0.1843	0.2546
Qwen-Audio-Chat	64.03 ± 1.54	60.19	67.22	0.3315 ± 0.0213	0.2750	0.3898
Qwen2-Audio (zero-shot)	85.06 ± 1.92	80.28	90.93	0.4419 ± 0.0239	0.3694	0.4972
Qwen2-Audio (one-shot, same template)	87.27 ± 1.56	83.70	90.74	0.4704 ± 0.0263	0.3991	0.5241
Qwen2-Audio (one-shot, different template)	67.93 ± 2.26	63.33	72.59	0.3485 ± 0.0229	0.2880	0.4000
Ultravox v0.4.1	89.17 ± 1.19	85.93	92.31	0.4707 ± 0.0207	0.4278	0.5194
Ultravox v0.6	99.82 ± 0.25	98.98	100.00	0.5315 ± 0.0237	0.4722	0.5935

This suggests a persistent bias toward speech-to-text behavior.

Several models exhibited a substantial proportion of random or incoherent responses, most notably GAMA (71.76%) and Qwen2-Audio with one-shot approach (35.41% and 37.74%). These responses suggest that models often generate unconstrained outputs when uncertain rather than signaling uncertainty or failure.

Qwen2-Audio often confused the given task with speaker recognition. Language control failures were observed across multiple models, including Qwen2-Audio, Whisper + Qwen, and Ultravox v0.4.1. Notable but less common errors included speculative answers given by both Ultravox models, answering the prompt with a description instead of a direct response, and providing an unrelated safety-style refusal. For example, Qwen2-Audio with zero-shot approach stated that it could not discuss political manners, despite the task being unrelated.

Overall, the observed error patterns for the *Descriptive* approach suggest that failures are primarily caused by a misunderstanding of the task objective, strong transcription and speaker identification biases, and an increased rate of speculative or random generation. Compared to the *Yes/No* approach, these findings indicate that the *Descriptive* approach includes a broader diversity of error categories, highlighting the need for improved task grounding.

H ART-H Robustness

Since ART benchmark consists of 9 000 samples, which equals to over 30 hours of audio, we designed ART-H. It is a subset of 24 samples per task, where half of the instances has *Yes* answers and the rest has *No* answers.

To ensure that the models' performance is stable across samples, a random sampling analysis was conducted. For each approach, 216 outputs were selected randomly 100 times, computing the number of relevant answers and the accuracy for each sample. The results are shown in Tables 14–16.

The resulting accuracies closely match the values achieved on the full set of 9 000 outputs, demonstrating that the models' performance is consistent even on randomly selected subsets of outputs.

Table 15: Results of the random sampling analysis using *Descriptive* approach and Llama 3.3 as a judge.

	Relevant			Absolute Accuracy		
Model	**mean**	**min**	**max**	**mean**	**min**	**max**
Whisper + Llama	62.35 ± 2.37	56.02	67.76	0.3529 ± 0.0265	0.2824	0.4120
Whisper + Qwen	89.69 ± 1.58	85.45	93.43	0.4871 ± 0.0321	0.4000	0.5748
Audio Flamingo 3	74.21 ± 2.92	65.26	80.93	0.3562 ± 0.0286	0.2817	0.4507
GAMA	1.28 ± 0.74	0.00	3.70	0.0084 ± 0.0061	0.0000	0.0278
Qwen-Audio-Chat	14.51 ± 1.94	8.80	18.06	0.0766 ± 0.0160	0.0370	0.1157
Qwen2-Audio (zero-shot)	92.35 ± 1.59	88.32	95.77	0.4476 ± 0.0267	0.3632	0.5209
Qwen2-Audio (one-shot, same template)	94.01 ± 1.79	89.10	97.64	0.4720 ± 0.0249	0.4140	0.5592
Qwen2-Audio (one-shot, different template)	93.18 ± 1.69	87.68	97.66	0.4463 ± 0.0283	0.3814	0.5093
Ultravox v0.4.1	60.15 ± 3.17	52.09	68.84	0.2403 ± 0.0210	0.1831	0.3023
Ultravox v0.6	79.41 ± 2.40	73.49	85.51	0.3965 ± 0.0299	0.3224	0.4744

Table 16: Results of the random sampling analysis using *Descriptive* approach and Qwen3 as a judge.

	Relevant			Absolute Accuracy		
Model	**mean**	**min**	**max**	**mean**	**min**	**max**
Whisper + Llama	62.25 ± 2.06	56.94	67.59	0.3974 ± 0.0284	0.3380	0.4630
Whisper + Qwen	87.90 ± 1.63	83.80	91.67	0.5691 ± 0.0320	0.4907	0.6435
Audio Flamingo 3	77.10 ± 2.36	70.37	81.94	0.3555 ± 0.0269	0.2870	0.4398
GAMA	2.20 ± 0.90	0.46	5.09	0.0133 ± 0.0071	0.0000	0.0324
Qwen-Audio-Chat	19.92 ± 2.45	12.04	26.85	0.0529 ± 0.0153	0.0139	0.0972
Qwen2-Audio (zero-shot)	94.95 ± 1.28	92.59	98.15	0.4667 ± 0.0263	0.4120	0.5370
Qwen2-Audio (one-shot, same template)	96.43 ± 1.36	91.67	99.07	0.4993 ± 0.0257	0.4583	0.5787
Qwen2-Audio (one-shot, different template)	95.32 ± 1.28	91.67	98.15	0.4819 ± 0.0278	0.4167	0.5417
Ultravox v0.4.1	50.80 ± 3.00	43.98	57.41	0.1707 ± 0.0226	0.1250	0.2222
Ultravox v0.6	77.55 ± 2.73	71.76	83.80	0.4449 ± 0.0298	0.3657	0.5231