What do MLLMs hear? Examining reasoning with text and sound components in Multimodal Large Language Models

Enis Berk Çoban[1]
ecoban@gradcenter.cuny.edu

Michael I Mandel[1]
mim@mr-pc.org

Johanna Devaney[1,2]
jdevaney@gc.cuny.edu

[1] The Graduate Center, CUNY   [2] Brooklyn College, CUNY

Abstract

Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities, notably in connecting ideas and adhering to logical rules to solve problems. These models have evolved to accommodate various data modalities, including sound and images, known as multimodal LLMs (MLLMs), which are capable of describing images or sound recordings. Previous work has demonstrated that when the LLM component in MLLMs is frozen, the audio or visual encoder serves to caption the sound or image input facilitating text-based reasoning with the LLM component. We are interested in using the LLM's reasoning capabilities in order to facilitate classification. In this paper, we demonstrate through a captioning/classification experiment that an audio MLLM cannot fully leverage its LLM's text-based reasoning when generating audio captions. We also consider how this may be due to MLLMs separately representing auditory and textual information such that it severs the reasoning pathway from the LLM to the audio encoder.

1 Introduction

Humans can learn from descriptions of events or ideas and can recognize them afterwards, even if they are observing such an event for the first time. Can the reasoning abilities in large language models (LLMs) enable them to achieve a similar goal? It has recently been shown that LLMs trained on internet-scale data have zero and few-shot capabilities [Brown et al., 2020, Kojima et al., 2022], demonstrating that they can solve tasks for which they were not specifically trained. For example, LLMs trained to predict the next chunks of text in a given text can also perform other natural language tasks, such as summarizing a given text. More complex tasks that LLMs cannot solve from scratch can be solved by in-context learning, where question and answer pairs are provided in the prompt, which act like training data. Another way to leverage in-context learning is by providing related information in the prompt to help the model generate solutions based on the given information. LLMs can identify and connect related ideas in the given text, and draw conclusions regarding these facts or ideas. Reasoning abilities allow LLMs to make connections between related concepts and provide responses by collating different information present in their training data [Wei et al., 2022], although there is a debate around whether these abilities are emergent in larger-scale models [Schaeffer et al., 2023]. While reasoning abilities are present in LLMs, they are limited, both due to catastrophic forgetting, which causes reasoning to disappear [De Lange et al., 2021], and hallucinations, which leads to fallacy generation [Tonmoy et al., 2024].

Multimodal large language models (MLLMs), where the output of image or audio encoders are tokenized and input into an LLM, exhibit some of the reasoning capabilities found in non-multimodal LLMs [Wang et al., 2024]. We are interested in leveraging their reasoning ability to classify
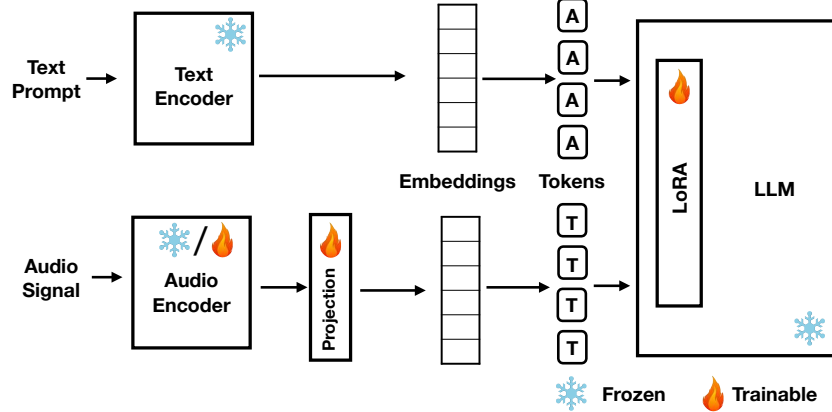
Figure 1: Generic audio MLLM architecture, specific components may vary with specific models. The snowflake represents components that are typically frozen and the flame represents those that are typically trained (or can be in the case of the '/').

low-resource classes using only descriptions of the given classes for both zero- and few-shot learning. For the latter, MLLMs could use text-based reasoning to not only learn from a small number of labeled samples but also generalize better, such that the presence of unrelated data in samples, like background noise, would not impact classification. Previous demonstrations of MLLM's reasoning abilities have been reasoning *about* the image or audio input [e.g., Gong et al., 2023b]. In order to leverage the reasoning to learn to classify unseen audio from textual descriptions, the MLLM needs to co-reason *with* the multimodal content. Research on vision MLLMs, however, has demonstrated that they do not possess these co-reasoning capabilities, that the models depend on the input modalities to control output as if they are simple flags turning on and off a specific task [Qi et al., 2023]. In this paper, we analyze audio MLLMs capabilities in order to more fully understand their co-reasoning abilities and limitations, and to discuss possible solutions to this problem.

2   Reasoning in multimodal large language models

The earliest MLLMs had encoders that output embeddings representing each input modality, which were then combined using different strategies, such as simple combinations, before being fed into another language model [Wu et al., 2023]. Initial models used different encoders for each modality, depending on what was most commonly used for such input (e.g., CNNs for images and RNNs for text [You et al., 2016]). The subsequent development and success of transformers on different modalities led to a uniform encoder approach, with similar attention architectures and sizes, that is commonly used in current MLLMs [Radford et al., 2021]. This has facilitated the development of MLLMs that can incorporate images and/or audio with the LLM core text-based functionality.

2.1   Visual reasoning in MLLMs

Much of the work on reasoning in vision MLLMs builds on the Visual Question Answering (VQA) [Antol et al., 2015], a captioning task that arguably requires a model to demonstrate complex reasoning capabilities. Modifications to the original VQA approach include using different metrics that better reflect real-world visual concepts [Kervadec et al., 2021] and visualization approaches that allow for finer-grained investigation of vision MLLMs' reasoning capabilities [Jaunet et al., 2021]. Recent evaluations on visual reasoning have shown that the representation of visual information in MLLMs is a bag of words, rather than an ordered representation, as evidenced by the inability of the models to answer any questions related to the order of the objects in the images [Yuksekgonul et al., 2022]. Vision MLLMs also lack spatial reasoning capabilities when queried about the left-right location of objects in an image [Kamath et al., 2023]. Another problem with vision MLLMs is their lack of understanding of relationships between objects, such as when they incorrectly assign human actions to animals, and vice-versa [Thrush et al., 2022]. Research into visual reasoning capabilities in vision MLLMs is facilitated by the development of benchmarks, which have demonstrated some of their

ongoing reasoning and hallucination issues [Fu et al., 2023, Fan et al., 2024]. These include vision MLLMs performing worse than LLMs on instruction following [Zeng et al., 2023] and vision MLLMs exhibiting a lack of semantic grounding, which limits their ability to leverage textual relationships present in the LLM in the visual modality [Lu et al., 2024].

[Kervadec et al., 2019, Song et al., 2023] to facilitate reasoning on more fine-grained aspects of the data. While generating more samples is a sufficient solution for specific abilities, it requires having matching data samples from each modality. As a result, it becomes even harder to come up with sufficient training data for data-hungry deep-learning models. A well-aligned model would be able to use abilities gained in one modality in another without requiring any extra data. For example, if a model can converse about modes of transportation, it should be able to show similar reasoning capabilities given an image of a car.

In-context learning through prompts is used to facilitate and analyze reasoning in MLLMs [Zhao et al., 2023]. Reasoning abilities typically improve when a model is forced to take specific steps through its prompt [Kojima et al., 2022, Yao et al., 2022] and prompt probing (including prompts related to visual, textual, and outside information) has been key in the understanding of MLLMs visual reasoning limitations [Qi et al., 2023]. Such probing has shown that in MLLMs, the use of non-linguistic prompts can increase the risk of catastrophic forgetting [Wang et al., 2023], reducing any reasoning capabilities that they do exhibit. It also underlines the importance of incorporating outside knowledge in the testing paradigm when assessing reasoning [Marino et al., 2019] as a way to assess the contribution of how much text-based reasoning concepts in the language model are being leveraged in the MLLM. Typically multi-step training is employed to mitigate catastrophic forgetting, the second modality is aligned to the frozen LLM and then fine-tuned with LLM low-rank adaption (LoRA) [Alayrac et al., 2022, Ye et al., 2023].

2.2 Audio reasoning in MLLMs

Audio MLLMs are a more recent development than vision MLLMs [Deshmukh et al., 2023, Silva et al., 2023, Gong et al., 2023b, Tang et al., 2023, 2024, e.g.,]. Arguably, the first audio MLLM was Pengi [Deshmukh et al., 2023], which frames all audio tasks as a text generation task in order to leverage the causal capabilities of the integrated LLM on the input audio prompt. It uses the hierarchical audio transformer HTS-AT [Chen et al., 2022] for its audio encoder, CLIP [Radford et al., 2021] for its text encoder, and applies contrastive pre-training with CLAP [Elizalde et al., 2023]. While Deshmukh et al. [2023] evaluate Pengi on a range of open-ended tasks, such as audio captioning and audio question/answering, and closed tasks related to classification and retrieval, they do not evaluate Pengi's reasoning capabilities. Some slightly more recent audio MLLMs that explore the concept of reasoning are Listen, Think, Understand (LTU) [Gong et al., 2023b] and SALMONN [Tang et al., 2023]. Figure 1 provides a generalized overview of the architecture of these models.

LTU uses Audio Spectrogram Transformer (AST) [Gong et al., 2021] for its audio encoder and the LLaMA text encoder [Touvron et al., 2023] along with CAV-MAE [Gong et al., 2022] for contrastive pre-training. Like Pengi, LTU treats all audio as input for automatic audio captioning (AAC), from which it leverages the reasoning abilities of LLaMA to reason *about* the audio captions. They freeze AST and use Low-rank Adaptation (LoRA) [Hu et al., 2021] to force the model to condition on the audio captions and not rely simply on the language model, which helps to minimize hallucinations. In evaluating LTU's reasoning capabilities, Gong et al. [2023b] are concerned with the model's ability to "think", which the authors argue is demonstrated by tasks where the model explains an audio caption, and "understand", which they further argue is demonstrated by tasks where the model has to infer further action. Gong et al. [2023a] also released LTU-AS, which integrates a speech model.

SALMONN uses two audio encoders one from a speech model and one from a generic audio model, which gives it automatic speech recognition (ASR) capabilities, like LTU-AS. SALMONN feeds the output of Whisper's speech encoder [Radford et al., 2023] and BEAT's audio encoder [Chen et al., 2023a] for generic audio sound into Q-Former query transformer [Li et al., 2023] to generate audio tokens for input into Vicuna [Chiang et al., 2023]. Tang et al. [2023] specifically evaluate the effect of fine-tuning on the reasoning tasks. Since most of their data consists of ASR and AAC instructions, the model tends to ignore the prompt and respond with one of these answers. To address this, activation tuning is applied, which is lowering the scaling factor of the LoRA method. As with Pengi and LTU, the reasoning tasks in SALMONN are performed on the text generated from the audio samples, either captions or transcribed speech. Evaluation of SALMONN revealed a similar

phenomenon to those observed in visual MLLMs, where the model forgets some of the text-based commonsense knowledge available in the LLM.

Evaluations of reasoning in audio MLLMs, primarily with LTU and SALMONN, have examined and proposed ways to address the observation made in visual MLLM literature that MLLMs exhibit issues with instruction following compared to LLMs. The lack of semantic grounding, however, has not been evaluated and is the focus of this paper.

3 Experiment 1: In-context audio classification

Audio MLLMs have demonstrated competence, if not state-of-the-art performance, on in-context classification tasks. These models, when provided with a minimal set of examples or a succinct description of the anticipated output, are capable of discerning the underlying structure and accurately completing the given input. In multimodal systems, in-context learning extends to interactions with image or audio inputs, broadening the model's understanding to other modalities. This enables the model to comprehend examples that reference the subject of interest in an image or audio context. As a result, even when provided solely with visual or auditory descriptions of expected labels in-context, the model can effectively undertake classification tasks across different modalities, suggesting a form of reasoning. This capability proves particularly advantageous for resource-constrained classes, such as rare bird species that have detailed written reports on their calls and songs [e.g., Hannon et al., 2020]. Audio MLLMs, leveraging their in-context learning proficiency, could effectively harness this written data to enhance classification performance in low-resource scenarios.

3.1 Methodology

We used the pre-trained LTU model as our base model in these experiments. The LTU model was trained using existing audio datasets, their labels, descriptions, transcripts, and metadata as available. As described in Section 2.2, LTU generates text captions for an input sound. These captions can be utilized for classification by calculating the similarity between them and the expected labels. This similarity is then interpreted as a confidence score for each label. We adopted the methodology outlined in the LTU paper for this task, employing OpenAI's text-embedding-ada-002 model to obtain text embeddings for both the labels and model outputs. For each sample, we computed the cosine similarity between the label embeddings and the model's output or caption. These similarity scores were then used as confidence indicators for the corresponding label of the given data point.

The advantage of relying on the similarity between text embeddings is that it enables the matching of related captions with the correct label, even if the given label does not appear in the caption. However, a potential disadvantage is the loss of semantic information, which could lead to incorrect label matching due to the diverse information contained within the caption. In order to quantify this, we conducted the same experiments repeatedly, with the variance being 0.7%, a figure negligible in comparison to the changes in results—up to 20%—observed between different experiments, as demonstrated in Tables 2 and 1.

We ran this experiment on EDANSA [Çoban et al., 2022], a bioacoustics audio dataset collected in Alaska, which comprises 10,782 10-second samples, totaling 27 hours of audio. This multi-label dataset contains 28 distinct labels, organized in a hierarchical structure. We selected 12 labels that represent main events and have more than 400 samples in total, namely: Anthrophony, Aircraft, Biophony, Songbird, Bird, Grouse, Insect, Waterfowl, Geophony, Rainfall, Wind, and Silence. Since the LTU model does not use EDANSA, nor any other ecological soundscapes, in its training set, it is an ideal choice for testing LTU's out-of-distribution performance. Furthermore, labels such as Geophony and Biophony are not common in LTU's training set, allowing for better evaluation of the effect of fine-tuning.

While LTU has been shown to underperform compared to supervised approaches for classification tasks [Gong et al., 2023b], we are interested in its in-context classification capabilities using knowledge from the text domain. Therefore, we tested the original LTU model on EDANSA labels to establish a baseline performance. In this experiment, we employed the most effective prompt from the original LTU classification experiments: 'write an audio caption describing the sound.' We noted that the model often failed to output expected labels such as biophony due to a lack of examples in

the training. Consequently, we devised alternative labels (Appendix B) that are synonymous and incorporated these into our experiments.

However, even if using similarity measures aids in classification, LTU's open-ended captions may not always match expected labels, even after fine-tuning. This is a limitation compared to supervised models with fixed output labels, as the model might not use exact labels or might miss silent or secondary sound events. To address this, we included expected labels in the prompts, such as "Write an audio caption describing the sound. Could the sound be {list of comma separated labels}?".

Subsequently, we fine-tuned the LTU model on the EDANSA dataset using the suggested train, test, and validation split from the original paper. We employed LoRA to fine-tune the language model, and for the audio encoder, we experimented with full fine-tuning and only training the projection layer. With LoRA, we used a learning rate of 16, as suggested by the original LTU paper [Gong et al., 2023b], and also experimented with values of 2, 4, and 8. We opted for lower values based on research indicating that a lower learning rate enhances performance by preventing catastrophic forgetting [Tang et al., 2023]. We used a batch size of 256, a LoRA alpha of 1, and a LoRA dropout of 0.05. We applied early stopping if validation did not improve after 10 epochs and used the checkpoint with the lowest validation score.

For these experiments, we primarily used the 13B version of the LTU model. The in-context learning experiments in this Section 3 were run with both 7B and 13B models, but we found the performance of the 13B model to be superior, and thus its performance is reported for all experiments.

To conduct these experiments, we utilized two NVIDIA A40 GPUs, each with 48GB of memory. Each fine-tuning run took approximately 3 hours, and we conducted a total of 46 runs. This amounted to 11.5 days of GPU time in total.

3.1.1   In-Context Learning with Grouse Call Descriptions

Of all the labels we examined, Grouse proved to be the most formidable challenge in terms of classification. This is primarily due to the limited number of samples available for this label, coupled with the complex and diverse array of calls that the Grouse is known to produce. The grouse featured in the EDANSA dataset, several species of ptarmigan, is renowned for its wide range of sounds, including hissing, chirping, and peeping. However, it is most notably recognized for the drumming sounds produced by the male [Hannon et al., 2020]. As a result of this complexity and diversity, supervised models encounter significant difficulty when classifying the Grouse label. These models typically require a larger sample size when dealing with a class that exhibits a high degree of diversity within its samples. Given the known descriptions of these specific calls and the limited number of samples, we hypothesized that the LTU model would effectively leverage the written information and outperform the supervised model.

To test this hypothesis, we incorporated descriptions of the grouse's calls into the prompt, asking the model to identify these sounds within the given audio. It's important to note that we only provided this additional information during the test and validation phases. The prompt for this experiment reads as follows: "Provide labels for the audio file. Could it be a Grouse? Here is detailed information on Grouse call types: ...(3) Krrow is a medium-length (50-300 ms) call that rises quickly and falls slowly in frequency: In males, it sounds like 'bugow'; in females, like 'meow'; it is typically given during aggressive disputes ...Please list the labels for the audio file." A comprehensive version of all call types is provided in the supplemental materials in Appendix A.

3.2   Results

The results of our audio classification experiments are encapsulated in Table 1. As anticipated, the LTU model's performance in classifying samples from the EDANSA dataset falls short when compared to the supervised model. The supervised approach achieves a mean AUC score across labels of 0.97, whereas the LTU model lags considerably behind with a mean AUC of 0.69. Upon training the audio projection layer ("Partially Fine-tuned"), which serves as the intermediary between the LLM and the audio encoder, the performance experiences a marginal increase of 0.03 points. This layer is instrumental in aligning audio embeddings with input tokens, and the modest improvement suggests that the LLM may necessitate additional training to effectively recognize out-of-distribution sounds. When we proceed to fully fine-tune both the audio encoder and the LLM ("Fully Fine-tuned"),

Table 1: LTU classification performance versus a supervised model. 'Vanilla' is un-fine-tuned LTU, 'Partially Fine-tuned' is with tuning the audio projection, 'Fully Fine-tuned' is tuning the audio encoder and applying LoRA training. Mean AUC and F1 are computed across all EDANSA classes.

Experiment	Mean AUC	Mean F1
Supervised	0.97	0.90
Vanilla	0.69	0.54
Partially Fine-tuned	0.72	0.59
Fully Fine-tuned	0.81	0.68

Table 2: In-context prompting results, compared to a supervised model and fully fine-tuning (fine-tuning the audio encoder and applying LoRA training to the LLM). 'Label in Prompt' refers to explicitly listing all class labels in the prompt and 'Grouse Call in Prompt' refers to providing acoustic information about the Grouse call in the prompt.

Experiment	Mean		Grouse		Insect	
	AUC	F1	AUC	F1	AUC	F1
Supervised	0.97	0.90	0.97	0.80	0.97	0.96
Fully Fine-Tuned	0.81	0.68	0.82	0.41	0.94	0.95
Labels in Prompt	0.81	0.68	0.79	0.35	0.93	0.81
Grouse Call in Prompt	0.81	0.68	0.76	0.46	0.93	0.82

the performance is enhanced by 0.09. This improvement suggests that the LLM is acquiring new audio concepts and refining the mapping from audio to text for novel classes.

Table 2 compares the performance of different prompt strategies with the fully fine-tuned LTU model. Interestingly, supplying potential labels in the prompt to steer the model towards outputting labels of interest does not enhance the average performance. While some labels exhibit a slight improvement, others deteriorate, indicating that the model does not accord substantial attention to the information presented in the prompt.

Our final experiment with prompt modifications, which involved incorporating descriptions of Grouse calls into the prompt, also failed to meaningfully change the results. Although the AUC score for Grouse classification decreased by 0.06, the F-1 score rose by 0.05 to 0.46. Note, however, this score is still significantly lower than the performance achieved by the supervised model. The F-1 score was selected by calculating the F-1 for all threshold values with 0.001 increments on the validation set and choosing the threshold that yielded the optimal F-1 score for the reported test set results.

3.3 Discussion

We might expect that given the extensive training of the audio encoder, it would summarize the general properties of any audio input first as embeddings and then audio tokens. However, our results reveal a more complex interaction. The performance improvement observed when transitioning from partial to full fine-tuning suggests that the LLM is not merely processing the general audio properties encapsulated in the embeddings. Instead, it appears to focus on specific audio tokens that correspond to the related audio class, indicating that the LLM is mapping specific sound events to their associated words. This exposes a limitation. Ideally, the LLM should leverage the general properties in the embeddings to be able to reason with audio. However, its focus on specific audio tokens suggests it is not fully utilizing the information in the embeddings, potentially limiting its generalization capabilities. These findings underscore the challenges inherent in leveraging LLMs for in-context audio classification tasks, particularly when dealing with out-of-distribution sounds. They also highlight the potential benefits and limitations of various fine-tuning strategies and prompt modifications. While the LLM's ability to map specific sound events to associated words is beneficial, it is also a limitation in generalization to a wider range of audio inputs. The minimal impact of text prompts further suggests a lack of direct linkage between audio and text representations within the LLM.

Table 3: Prompts used in experiment on concept representations in an audio MLLM for the two categories considered: similarity (synonyms) and hierarchy (hypernym). Slightly different prompts were crafted within each category for text- and audio-based queries.

Similarity (synonym)	*Text*	P1: Is {concept} similar to {synonym}?
	Audio	P2: Is the sound of the object in this audio signal similar to {synonym}?
Hierarchy (hypernym)	*Text*	P3: Is {concept} a type of {hypernym}?
	Audio	P4: Is the sound of the object in this audio signal a type of {hypernym}?

4   Experiment 2: Examining concept representations in an audio MLLM

Building upon the findings of Experiment 1, our second experiment aims to delve deeper into the reasoning capabilities of MLLMs. Specifically, we are interested in discerning whether these models utilize information from the audio modality in their textual reasoning processes, or if they primarily map this information to individual keywords. We designed an input and expected output that necessitates the activation of reasoning abilities, such that we could modify the input to uncover what triggers these reasoning capabilities.

While the interaction between concepts and their textual representations can manifest in myriad relations, one of the most structured and extensively studied is the semantic relation between words. This is meticulously catalogued in lexical databases such as WordNet [Miller, 1995]. WordNet organizes words into sets of synonyms called synsets, and records a variety of relations among these sets or their members, including synonyms, antonyms, hypernyms and hyponyms. This rich network of semantically related words and concepts provides a structured framework that can be used to understand and analyze the complex interrelationships between different concepts and their textual representations. Synonyms have previously been used to to evaluate vision MLLMs [e.g., Zohar et al., 2023], however the use of hypernyms is less common. Hypernyms have been used more widely in evaluating LLMs without a multimodal component [e.g., Shani et al., 2023] but are typically only mentioned in passing in evaluations of visual MLLMs [e.g., Chen et al., 2023b].

We anticipated that when an LLM is augmented with additional modalities, such as audio, its capacity to answer questions pertaining to the relationships between concepts should extend to their corresponding representations within the new modality. This expectation is grounded in the understanding that the semantic relationships modeled by LLMs, such as hypernyms, are not confined to textual representations but can be extrapolated to other modalities as well.

4.1   Methodology

LLMs model semantic relationships and can answer questions requiring reasoning, such as hypernyms and synonyms. Hypernyms represent a type of semantic relationship where one term serves as a broader category encompassing a set of other terms. For instance, 'fruit' is a hypernym for 'apple' and 'orange'. Synonyms represent another type of semantic relationship where two different terms share a similar meaning. We use this property of the hypernym and synonym relationships to construct the text prompt P3 from Table 3. For synonyms, we pose the question "Is concept similar to synonym?" which maps to the text prompt P1 from Table 3. We expect LLMs to be able to answer these questions correctly if they understand the relationship between these concepts. This is because the ability to correctly identify and use hypernym and synonym relationships is a key aspect of understanding semantic relationships in language, and is a strong indicator of a system's ability to reason and generate meaningful responses in natural language processing tasks. The hypernym and synonym approaches can also be applied to the audio modality. If the model has learned the semantic relationship between the concepts from textual data, it should be able to transfer this understanding to audio data. Given the sound of a concept, the model should be able to identify it as a type of hypernym or as similar to a synonym, demonstrating its ability to reason. To test this, we pose a question using an audio file that represents a concept. The question is framed as "Is the sound of the object in this audio signal a type of hypernym?" for hypernyms, and "Is the sound of the object in this audio signal similar to synonym?" for synonyms. These question formats correspond to the audio prompts P4 and P2 from Table 3, respectively. For example, if we have an audio file of a songbird's chirping, the question would be "Is the sound of the object in this audio signal a type of bird?", to

which the answer should be yes. We also test a condition where a silent audio file is provided with the text prompt to assess whether the mere presence of audio changes the MLLM's reasoning processes.

We have constructed a concise benchmark that comprises 12 concept words. Each word is associated with up to 4 hypernyms, 4 synonyms, and 4 unrelated terms, culminating in a total of 111 word relationships (see Appendix C for a full list). We expect the model to respond favorably to all prompts involving concept hypernyms and synonyms, while showing a negative response to pairs of concepts and unrelated terms. For each word, we employ 4 audio files, each repeated 4 times, resulting in 16 queries. Consequently, we have 16 outputs per word pairs. To interpret these outputs, we employ regular expressions to discern whether the model's response aligns with the question, i.e., whether it is responding positively or negatively. A positive response, or an affirmation, is classified as a 'yes' response. We then compute the 'yes' rate, which is defined as the percentage of 'yes' responses out of the total 16 responses for each word pair. Notably, a 'no' response to an unrelated term is correct in our context and is thus flipped to 'yes' for consistency in the 'yes' rate, ensuring it accurately represents the model's correct identifications of both related and unrelated terms. Our samples are carefully selected from the evaluation set of AudioSet. We specifically opt for samples that contain only the target label, deliberately excluding any other sound events. A full list of the audio files we used is available in Appendix D. We use the same compute resources described for experiment 1, however we only run models in inference mode, adding up to less than 30 hours of GPU time.

4.2 Results

Figure 2 presents the results of our experiments on concept representations in an audio MLLM. In the similarity category, the text-only approach and the text with silent audio approach both showed perfect performance. This indicates that the model was able to perfectly identify synonyms in a text-only context and when paired with silent audio. However, when actual audio data was introduced, there was a slight decrease in performance. The AudioSet condition had a median yes rate of 1.0 but with a slightly wider IQR of 0.031, indicating a small amount of variability in the responses. The EDANSA condition showed a further decrease in performance, with a median yes rate of 1.0 but an even wider IQR of 0.093, suggesting greater variability in the model's responses.

In the hierarchy category, the text-only approach and the text with silent audio approach again showed perfect performance. This demonstrates the model's proficiency in identifying hypernyms in a text-only context and when paired with silent audio. However, the introduction of actual audio data resulted in a significant decrease in performance. The AudioSet condition had a median yes rate of 28.1% and an IQR of 0.234, indicating a substantial amount of variability in the responses. The performance further deteriorated with the EDANSA condition, which had a median yes rate of 0.0 and an IQR of 0.265625.

4.3 Discussion

Our empirical findings show a distinct performance disparity in the Large Language Model (LLM) when tasked with answering questions pertaining to text-to-text versus audio-to-text relationships. In the similarity task, the model adeptly reasons about tokens, irrespective of their origin from audio captions or text, demonstrating its proficiency in identifying related words. However, a slight performance edge is observed in text-based tasks, indicating the model's stronger affinity for its native modality. In contrast, the hierarchy task presents a more complex challenge. While the LLM effectively leverages its reasoning capabilities with text, its performance wanes when presented with sound, suggesting a lack of connections between audio and textual concepts. This limitation is echoed in the findings from Experiment 1, where the model excelled in tasks involving one-to-one mapping, akin to the similarity task, but struggled with tasks requiring a broader understanding of relationships between concepts, as in the hierarchy task. We also observed that the model's performance is worse with EDANSA samples compared to AudioSet samples, indicating potential difficulties in processing and understanding out-of-distribution sound data.

5 Limitations

The experiments in this paper are exploratory and limited in scope. Most notably, the Grouse-specific components of Experiment 1 used only the prompts described in Appendix A. Although this is a
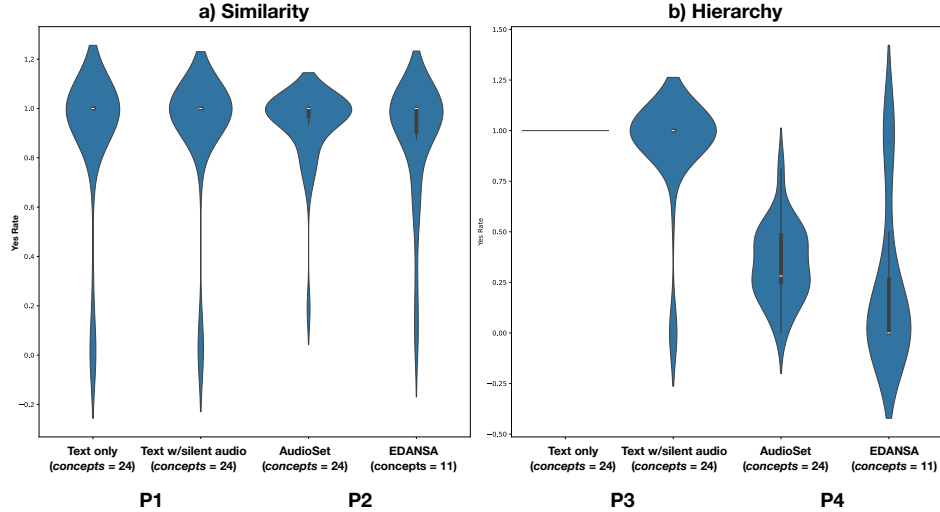
Figure 2: Results of experiment on concept representations in an audio MLLM. Subplot (a) shows the results for the similarity (synonyms and unrelated terms) category and subplot (b) for the hierarchy (hypernym) category. Each subplot shows violin plots of the yes rate for four different conditions: text-only prompting, text prompting with a silent audio file, text prompting with an audio file from AudioSet, and text prompting with an audio file from EDANSA. The prompts (P1–P4) are defined in in Table 3.

carefully crafted prompt, there remains the possibility that either the wording or content of the prompt was suboptimal. Similarly, Experiment 2 is limited to the carefully curated words and audio files in Appendices C and D, however, the amount of between-group consistency shown in the plots in Figure 2 for the synonymy experiment as well as the significant difference between groups for the hypernym experiment suggest that the dataset was sufficiently sized to capture the behavior related to LLM context representations that we were interested in exploring. A final limitation is our use of a single audio MLLM, LTU. This was largely constrained by available compute resources, but given the consistency in architecture across the three most currently popular audio MLLMs (Pengi, LTU, and SALMNON), it is likely that we would find similar results using any of them.

6   Conclusions and Future Work

In this paper, we evaluated whether an audio MLLM, specifically LTU, can exploit the reasoning capabilities of LLMs to learn in context through prompting (Experiment 1). We demonstrated the limitations of a current audio MLLM in leveraging its LLM's reasoning power via in-context prompting. This led to the design and implementation of Experiment 2 to examine the audio MLLM's concept representations with synonyms and hypernyms, demonstrating that the audio MLLM is not fully integrating text and audio information in a way that it can perform hierarchical-related reasoning on audio input. A common solution to address this reasoning in vision MLLMs is generating additional data pairs of text and images that require the model to pay attention to the missing ability, such as the order of the items in the images [Yuksekgonul et al., 2022]. This is a limited solution, which would only solve reasoning on tasks covered by the data pairs. There is potential to improve reasoning through finer-grained alignment between the modalities, although this is still an open area of research.

A better understanding of LLMs' reasoning capabilities, both in isolation and in the context of MLLMs, has the potential for broader societal impacts. This includes contributing to decoding what LLMs are and are not modeling, which can help delineate some of the known limitations that should be considered in their use. This could be used either positively or negatively, but we anticipate the development of such techniques would largely have a positive impact. This work is also the first step towards understanding how to leverage domain-specific description-based knowledge in audio classification with MLLMs, as exemplified by our in-context prompting experiment on grouse calls, to better proxy how humans learn.

Acknowledgments and Disclosure of Funding

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, December 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023a.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023b.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Enis Berk Çoban, Megan Perra, Dara Pir, and Michael I Mandel. EDANSA-2019: The ecoacoustic dataset from arctic north slope Alaska. In *Proceedings of the Workshop on the Detection and Classification of Acoustic Scenes and Events*, 2022.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *arXiv preprint arXiv:2305.11834*, 2023. URL https://arxiv.org/abs/2305.11834.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Lizhou Fan, Wenyue Hua, Xiang Li, Kaijie Zhu, Mingyu Jin, Lingyao Li, Haoyang Ling, Jinkui Chi, Jindong Wang, Xin Ma, et al. NPHardEval4V: A dynamic reasoning benchmark of multimodal large language models. *arXiv preprint arXiv:2403.01777*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech*, pages 571—-575, 2021.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023a.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023b.

Susan Jean Hannon, Perri K Eason, and Kathy Martin. Willow ptarmigan (lagopus lagopus), version 1.0. *Birds of the World*, 2020. Available from: `https://doi.org/10.2173/bow.wilpta.01`.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Theo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. Visqa: X-raying vision and language reasoning in transformers. *Transactions on Visualization and Computer Graphics*, 28(1):976–986, 2021.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up"' with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Weak supervision helps emergence of word-object alignment and improves vision-language tasks. *arXiv preprint arXiv:1912.03063*, 2019.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should VQA expect them to? In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2776–2785. IEEE/CVF, June 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and enhancement of semantic grounding in large vision-language models. In *Proceedings of the AAAI-ReLM Workshop*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3195–3204. IEEE/CVF, 2019.

George A Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. What is the limitation of multimodal LLMs? a deeper look into multimodal LLMs through prompt probing. *Information Processing & Management*, 60(6):103510, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

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

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2023.

Chen Shani, Jilles Vreeken, and Dafna Shahaf. Towards concept-aware large language models. *arXiv preprint arXiv:2311.01866*, 2023.

Dadallage AR Silva, Spencer Whitehead, Christopher Lengerich, and Hugh Leather. Collat: On adding fine-grained audio understanding to language models using token-level locked-language tuning. *Advances in Neural Information Processing Systems*, 36, 2023.

Shezheng Song, Xiaopeng Li, and Shasha Li. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023. URL https://arxiv.org/pdf/2310.13289.

Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. AVicuna: Audio-visual LLM with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5238–5248. IEEE/CVF, 2022.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *Proceedings of the International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 4651–4659. IEEE, 2016.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a GPT4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.

Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. LOVM: Language-only vision model selection. *Advances in Neural Information Processing Systems*, 36, 2023.

A Prompt used for In-Context Learning sub-experiment in Experiment 1

```
# source for Grouse description:
# https://birdsoftheworld.org/bow/species/wilpta/cur
# /introduction#vocal
{
    'default':
        'write an audio caption describing the sound',
    'label_in_prompt':
        'Write an audio caption describing the sound. \
   Could the sound be Silence, Biological soundscape,\
   aviation noise, Rainfall, Grouse, Insect,\
   Songbird, Duck and/or Goose and/or Swan,\
   anthropogenic noise, physical soundscape, Bird, or Wind?',
    'grouse01':
        """provide labels for the audio file, could it be a Grouse?
here is detailed information on Grouse call types:
Both Sexes. (1) Kok is a short (50 ms) clucking call; \
volume and frequency vary with intensity of arousal\
(2) Ko-ko-ko is a low-amplitude call given \
in prolonged bouts, sounding like low growls; \
(3) Krrow is a medium-length (50-300 ms)\
call that rises quickly and falls slowly in frequency: In males, \
sounds like bugow; in females, like meow; given during aggressive \
disputes \
(4) Kohwa, Kohway, and Kohwayo are medium-length (100 ms) \
calls often given in association with \
Krrow during aggressive interactions. \
(5) Aroo is a variable call with falling \
and rising pattern of frequency modulation at varying rate and \
intensity. \
(6) Rattle (song on the ground) is a long (800 ms), accelerating \
string of short elements similar to Kok\
(7) Flight Song (Aerial Bek) is a 2-part vocalization; first part \
is a decelerating series of modified Ko-ko-ko calls, second part \
a decelerating Kohwa  alternatively described as a series of nasal \
barks, typically a few single notes followed by a rattling sequence \
of 8-12 guttural notes, and several doubled notes on landing,
''whek!..whek-kekekrrrrekek-kek...koh-wa..koh-wa..koh-wa". \
(8) Scream is a brief, high-frequency\
call with indeterminate harmonic structure. (9) Hiss is a band of \
white noise about 2 s long.\
list labels for the audio file:
""",
    'grouse02_label_in_prompt':
        """provide labels for the audio file, could it be a Grouse?
here is detailed information on Grouse call types:
Both Sexes. (1) Kok is a short (50 ms) clucking call; \
volume and frequency vary with intensity of arousal\
(2) Ko-ko-ko is a low-amplitude call given \
in prolonged bouts, sounding like low growls; \
(3) Krrow is a medium-length (50-300 ms)\
call that rises quickly and falls slowly in frequency: In males, \
sounds like bugow; in females, like meow; given during aggressive \
disputes \
(4) Kohwa, Kohway, and Kohwayo are medium-length (100 ms) \
calls often given in association with \
Krrow during aggressive interactions. \
(5) Aroo is a variable call with falling \
```

```
and rising pattern of frequency modulation at varying rate and \
intensity.
(6)  Rattle (song on the ground) is a long (800 ms), accelerating \
string of short elements similar to Kok\
(7) Flight Song (Aerial Bek) is a 2-part vocalization; first part \
is a decelerating series of modified Ko-ko-ko calls, second part \
a decelerating Kohwa  alternatively described as a series of \
nasal barks, typically a few single notes followed by a rattling  \
sequenceof 8-12 guttural notes, and several doubled notes on \
landing, ''whek!..whek-kekekrrrrekek-kek..koh-wa..koh-wa..koh-wa".\
(8) Scream is a brief, high-frequency\ call with indeterminate \
harmonic structure.\
(9) Hiss is a band of white noise about 2 s long.\
 Write an audio caption describing the sound. \
 Could the sound be Silence, Biological soundscape,\
 aviation noise, Rainfall, Grouse, Insect,\
 Songbird, Duck and/or Goose and/or Swan,\
 anthropogenic noise, physical soundscape, Bird, or Wind?'
"""
}
```

B   Alternative labels considered in Experiment 1

Original label	Alternatives
Silence	Silence
Biophony	Biological soundscape, animal chorus, wildlife sounds, ecosystem acoustics
Aircraft	airplane, aviation noise, air traffic noise
Rain	Rainfall, raindrops, rain pattering
Grouse	Grouse
Bug	Insect, bug, entomological sounds, insect calls
Songbird	Songbird
DGS	Duck and/or Goose and/or Swan
Anthropophony	human-made noise, industrial noise, anthropogenic noise
Geophony	natural ambient sounds, non-biological soundscape, physical soundscape
Bird	Bird
Wind	Gust sounds, blowing wind, aeolian sound

C   Words used for concepts in Experiment 2

Category	Label	Synonym(s)	Hypernym(s)	Unrelated
biophony	bird	fowl	vertebrate	speech
		avian	craniate	speaking
		aves	chordate	wind
			animal	breathing
biophony	cattle	cows	bovine	working
		oxen	bovid	grumbly
		bos taurus	ruminant	melodic
			animal	wind
biophony	dog	canis familiaris	canine	comforting
		domestic dog	canid	speech
			domestic animal	brief
			animal	music

15

Category	Label	Synonym(s)	Hypernym(s)	Unrelated
biophony	insect	bug	arthropod invertebrate animal	wind authoritative characterized speech
anthrophony	aircraft	airplane airship aeroplane	craft vehicle conveyance transport	speech natural male rich
anthrophony	car	motorcar automobile auto machine	motor vehicle vehicle conveyance transport	speech police generic music
anthrophony	fireworks	pyrotechnics	low explosive explosive	speech speaking warm generic
anthrophony	alarm	alert	signal sign	speaking car vehicle authoritative
geophony	rain	rainfall rainwater	precipitation downfall weather atmospheric condition	surface thunder human authoritative
geophony	wind	air current current of air	weather weather condition atmospheric condition	microphone male rich instrument
geophony	thunder	boom	thunderstorm electrical storm storm atmospheric phenomenon	rolling speech footsteps whistling
geophony	waterfall	falls	water	speech male music man

D Concepts and audio files used in Experiment 2

Category	Label	AudioSet ID	EDANSA_IDs
biophony	bird	-XilaFMUwng -qS77R0Y1K8 12T-9dLEbY8 1dH-lZ8TNLU	INP-AR-03_20190617_220000_8m_30s__8m_40s S4A10227_20190611_043000_22m_19s__30m_34s_splt-21 S4A10301_20190613_000000_12m_50s__13m_0s S4A10301_20190613_000000_7m_30s__7m_40s
biophony	cattle	sbpW3Z87Nbc z3YihIejSIA UYBuKiXo92s KksMNKXuiNw	
biophony	dog	20qZLse0acs 8CrTpWNBiTo E6QQRZHrx6s KRdvyjpQfoI	
biophony	insect	9j_FItO0jt8	SINP03/SINP-03_20190704_210000_1m_30s__1m_40s

Category	Label	AudioSet ID	EDANSA_IDs
		zPSH6-UC4Og	S4A10327_20190725_104602_45m_50s__46m_0s
		5j_v9dhjbdU	anwr_41_S4A10273_20190707_183000_exact_2019-07-07
		QBj5dyzsJkY	SINP03/SINP-03_20190708_173000_15m_20s__15m_30s
anthrophony	aircraft	-OVb-UG8yJw	S4A10361_20210515_010002_41m_30s__41m_40s
		-ocADGlyaHc	S4A10272_20190509_073000_39m_20s__39m_30s
		7S88FsFE5EE	18/2019/S4A10280_20190525_104602_33m_22s__33m_32s
		DU3cNZdlylQ	S4A10298_20210730_060002_55m_50s__56m_0s
anthrophony	car	xRonpWC3SvY	S4A10443_20200428_100412_2m_0s__2m_10s
		-aOxR6ILsw8	S4A10291_20191010_144000_2m_0s__2m_10s
		4TshFWSsrn8	
		Kwpn3utYEHM	
anthrophony	fireworks	L6QtigLJD_4	
		l7RTgupQWcc	
		UxEyOSK9nxo	
		AJRD-zU2Akw	
anthrophony	alarm	3o-q-VMhyA8	
		FBut7W5XwnA	
		T_FZMsRHzLc	
		fcsGkE89Qi8	
geophony	rain	96HJ2f5dj6U	S4A10273_20190803_050000_42m_24s__57m_34s_splt-29
		fvQeqBqqcVw	S4A10273_20190803_050000_42m_24s__57m_34s_splt-53
		johz0yXuORc	AR01/2018/INP-AR-01_20180817_020000_7m_51s__8m_1s
		fwas0HLGbqM	S4A10287_20190803_050000_rain02_splt-2
geophony	wind	A74lbeD1k1o	S4A10273_20190803_093000_55m_0s__56m_50s_splt-2
		CkutJYIfghs	anwr_37_S4A10279_20190603_043000_exact_2019-06-03_04-38-36_0m_0s__0m_10s
		AkUDv7JexjQ	S4A10295_20190708_000000_49m_50s__50m_0s
		zzbTaK7CXJY	dempster/25/2020/S4A10334_20200415_140002_2m_54s__3m_4s
geophony	thunder	0439dMJj-FY	
		przrSPZgOkY	
		ZBaYrfz5afo	
		54wNjdYr8ww	
geophony	waterfall	FF2bhR7s3VY	
		JfDeETDDwhM	
		VMbJTgzMhKE	
		hfIfBPkH8Fo	