What do MLLMs hear? Examining reasoning with text and audio encoder 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]
johanna.devaney@brooklyn.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 generating descriptions of images or sound recordings. We evaluate how MLLMs separate representation of auditory and textual information may sever the reasoning pathway between the audio encoder and the LLM component. Through a captioning-based classification experiment with similar and hierarchical textual relationships, we demonstrate that audio MLLMs cannot fully leverage their LLMs' text-based reasoning when generating audio captions.

1 Introduction

Humans can learn from descriptions of events 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 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 acts 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. 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 low-resource classes using only descriptions of the given classes for both zero-shot learning. MLLMs

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

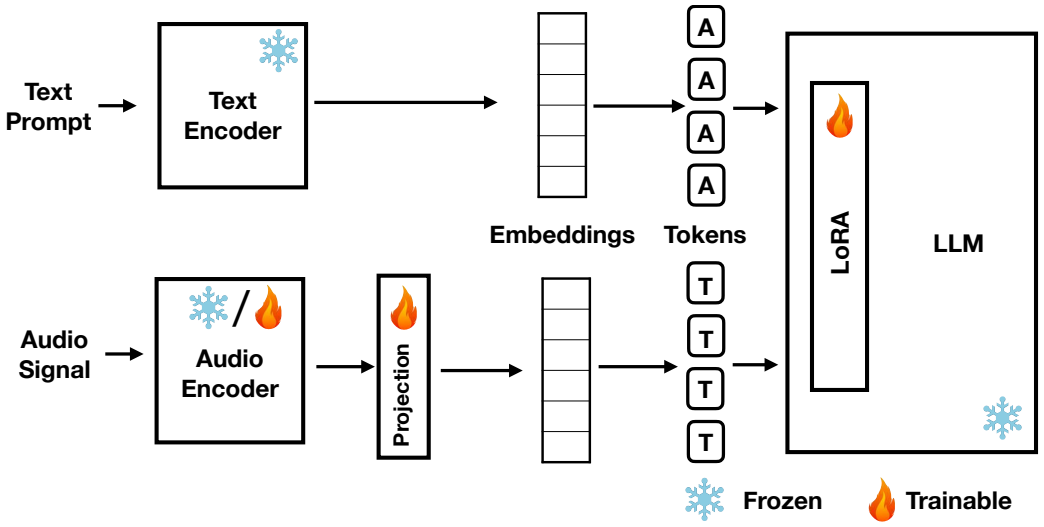

Figure 1: Generic audio MLLM architecture, specific components may vary with specific models. The snowflake represents components that are typically frozen and the flame represents those that are typically trained (or can be in the case of the '/').

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. Previously MLLMs' have demonstrated reasoning capabilities *about* their image or audio inputs [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, suggests that MLLMs 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.

2 Reasoning in multimodal large language models (MLLMs)

2.1 Visual reasoning in MLLMs

Much of the work on reasoning in vision MLLMs builds on the Visual Question Answering (VQA) dataset [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]. They also can understand 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 performing worse than LLMs on instruction following [Zeng et al., 2023] and an inability to leverage textual relationships present in the LLM in the visual modality [Lu et al., 2024]. While training 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 a 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.,]. 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. 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, rather than rely simply on the language model, in order to minimize hallucinations. In evaluating LTU's reasoning capabilities, Gong et al. [2023b] were 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.

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 [Gong et al., 2023a]. 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 transcribed text or captions. To address this limitation, activation tuning is applied, which is lowering the scaling factor of the LoRA method. As with 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.

3 Examining concept representations in an audio MLLM

To understand the reasoning capabilities of MLLMs, we want to determine if these models are leveraging the rich information embedded within the audio modality during their textual reasoning processes, or if they are merely mapping the audio information to individual keywords, or captions. Ideally when an LLM is augmented with additional modalities, such as audio, it should be able to answer questions leveraging the new modality. For example, audio MLLMs should utilize the general properties encapsulated in the audio embeddings to facilitate reasoning with audio. To investigate whether MLLMs are doing this, we have designed an input and expected output that necessitates the activation of reasoning abilities. By modifying the input, we aim to uncover the triggers for 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 cataloged 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

Table 1: 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}?

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].

3.1 Methodology

LLMs model semantic relationships and can answer questions requiring reasoning, such as understanding synonym and hypernym relations. Synonyms represent a type of semantic relationship where two different terms share a similar meaning. Hypernyms represent another 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'. We use this property of the synonym and hypernym relationships to construct the similarity- and hierarchy-related text prompts from Table 1. For synonyms we used P1 in Table 1 for text-based queries and P2 for audio-based queries. For hierarchy, we used P3 for text-based queries and P4 for audio queries. We expect MLLMs to be able to answer P1 and P3 correctly in the text domain if they understand the relationship between these concepts. If this is true, we are interested in evaluating whether the models have learned the semantic relationship between the concepts from textual data such that they can transfer this understanding to be able to reason with audio data, i.e., correctly answer P2 and P4. 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 carefully selected and iteratively refined our prompts, experimenting with numerous versions, including those suggested by the original model authors. The four prompts we ultimately chose (Table 1) were selected because they consistently produced the best results.

We have constructed a concise benchmark that comprises 12 concept words. Each word is associated with up to 4 synonyms and 4 hypernyms. We also curated a negative example of the synonyms and hypernyms. This resulted in 159 word relationships (see Appendix A for a full list). If the MLLMs are reasoning, we expect them to respond positively to all prompts involving correct concept hypernyms and synonyms, and negatively to prompts involving unrelated terms. For each word, we employ 4 audio files, each repeated 4 times, resulting in 16 queries. To interpret these outputs, we employ regular expressions to discern whether it is responding positively or negatively, framing this as a binary classification problem. Our samples are carefully selected from the evaluation set of AudioSet Gemmeke et al. [2017] and EDANSA [Çoban et al., 2022]. We specifically opted for samples that contain only the sound of the target label, deliberately excluding any other sound events. EDANSA is a bioacoustics audio dataset collected in Alaska. Since the LTU and SALMONN models do not use EDANSA, nor any other ecological soundscapes, in their training set, it is an ideal choice for testing their out-of-distribution performance. A full list of the audio files we used is available in Appendix B. We utilized two NVIDIA A40 GPUs, each with 48GB of memory. We only run models in inference mode, adding up to less than 30 hours of GPU time.

3.2 Results

In the similarity category, shown in the top of Table 2, LTU performed comparably across both conditions (P1 and P2) while the performance in SALMONN deteriorated when the task relied on audio captions (P2). In the hierarchy category, shown in the bottom of Table 2, the performance of LTU and SALMONN both deteriorated when relying on audio captions (P4).Notably, in the presence

Table 2: The table displays F1 scores, precision, and recall for LTU and SALMONN models in two tasks: similarity (synonyms, P1) and hierarchy (hypernyms, P3). Four conditions are tested: text-only, text with silent audio file, text with AudioSet audio (P2 and P4, italicized), and text with EDANSA audio (italicized).

			LTU			SALMONN		
Experiment	Prompt	Condition	F1	Precision	Recall	F1	Precision	Recall
Similarity	P1	Text	0.85	0.76	0.96	0.87	0.91	0.83
		Silent audio	0.88	0.82	0.95	0.87	0.79	0.96
	P2	*AudioSet*	*0.79*	*0.81*	*0.76*	*0.56*	*0.39*	*1*
		EDANSA	*0.84*	*0.85*	*0.84*	*0.59*	*0.42*	*1*
Hierarchy	P3	Text	0.98	0.95	1	0.92	0.95	0.90
		Silent audio	0.93	0.95	0.91	0.95	0.93	0.97
	P4	*AudioSet*	*0.47*	*0.84*	*0.33*	*0.66*	*0.5*	*1*
		EDANSA	*0.37*	*0.95*	*0.23*	*0.67*	*0.5*	*1*

of audio (P2 and P4), SALMONN exhibited a tendency to affirm all queries, as indicated by a recall of 1. A visualization of these results with violin plots is available in Appendix C.

3.3 Discussion

Our results demonstrate a performance disparity in MLLMs when they are tasked with answering questions with text prompts versus audio captions. For LTU, recall deteriorates when using audio captions, while for SALMONN, precision deteriorates. SALMONN also tends to answer 'yes' to questions when audio is introduced (including silent audio), which artificially increases its recall values. Thus, while both LTU and SALMONN effectively leverage reasoning capabilities with text, their performance wanes when presented with sound, suggesting a lack of connections between audio and textual concepts. The LTU results show consistent performance on similarity for both in-distribution (AudioSet) and out-of-distribution (EDANSA) data, suggesting that the root of the problem lies with the integration of the LLM component.

4 Conclusions

In this paper, we designed and implemented an experiment to examine the audio MLLM's concept representations with synonyms and hypernyms. We demonstrated that the audio MLLMs do not fully integrate text and audio information in a way that it can perform hierarchical-related reasoning on audio input. One limitation of our experiment is that it is limited to the carefully curated words and audio files in Appendices A and B, however, the amount of between-group consistency shown in Table 2 and Appendix C suggest that the dataset was sufficiently sized to capture the behavior related to MLLM context representations that we were interested in exploring.

A common solution to address 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. We believe that a better solution is training datasets with integrated text and audio tokens (e.g.,"Do you hear the <sound>? It is coming from the garden."), as has been demonstrated on standard audio tasks, such as captioning and few-shot audio classification [Liang et al., 2023].

Overall, a better understanding of LLMs' reasoning capabilities, both in isolation and in the context of MLLMs, has the potential to improve understanding of generative technologies. This includes contributing to decoding what LLMs are and are not modeling, which can help delineate some of the limitations that should be considered in their use.

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.

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`.

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.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

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.

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. 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.

Jinhua Liang, Xubo Liu, Wenwu Wang, Mark D Plumbley, Huy Phan, and Emmanouil Benetos. Acoustic prompt tuning: Empowering large language models with audition capabilities, november 2023. *URL http://arxiv.org/abs/2312.00249*, 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, 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.

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.

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.

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 Words used for concepts in Experiment

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

B Concepts and audio files used in Experiment

Category	Label	AudioSet ID	EDANSA_IDs
biophony	bird	-XilaFMUwng	INP-AR-03_20190617_220000_8m_30s__8m_40s
		-qS77R0Y1K8	S4A10227_20190611_043000_22m_19s__30m_34s_splt-21
		12T-9dLEbY8	S4A10301_20190613_000000_12m_50s__13m_0s
		1dH-lZ8TNLU	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
		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	

C Visualization of Experiment Results

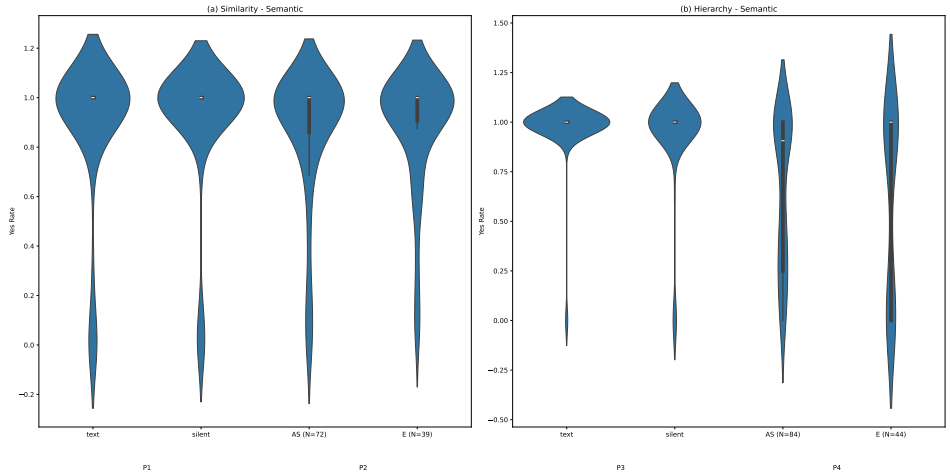

Figure C1: Visualization of results of experiment on concept representations on LTU. Subplot (a) shows the results for the similarity (synonyms and unrelated terms) category and subplot (b) for the hierarchy (hypernym) category. Each subplot shows violin plots of the corectness rate for four different 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 1.

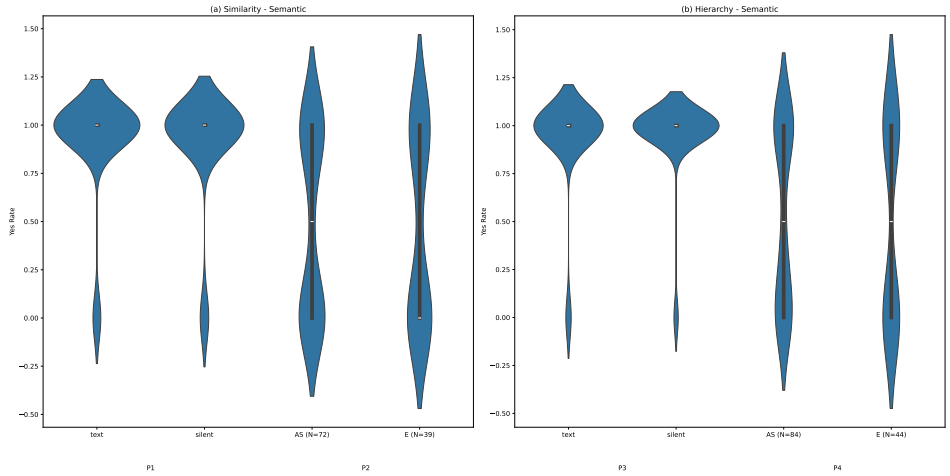

Figure C2: Visualization of results of experiment on concept representations on SALMONN. Subplot (a) shows the results for the similarity (synonyms and unrelated terms) category and subplot (b) for the hierarchy (hypernym) category. Each subplot shows violin plots of the corectness rate for four different 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 1.