# MUCHOMUSIC: EVALUATING MUSIC UNDERSTANDING IN MULTIMODAL AUDIO-LANGUAGE MODELS

**Benno Weck*[1]** **Ilaria Manco*[2]** **Emmanouil Benetos[2]**
**Elio Quinton[3]** **George Fazekas[2]** **Dmitry Bogdanov[1]**
[1]Universitat Pompeu Fabra, [2]Queen Mary University of London, [3]Universal Music Group
* equal contribution benno.weck01@estudiant.upf.edu, i.manco@qmul.ac.uk

## ABSTRACT

Multimodal models that jointly process audio and language hold great promise in audio understanding and are increasingly being adopted in the music domain. By allowing users to query via text and obtain information about a given audio input, these models have the potential to enable a variety of music understanding tasks via language-based interfaces. However, their evaluation poses considerable challenges, and it remains unclear how to effectively assess their ability to correctly interpret music-related inputs with current methods. Motivated by this, we introduce MuChoMusic, a benchmark for evaluating music understanding in multimodal language models focused on audio. MuChoMusic comprises 1,187 multiple-choice questions, all validated by human annotators, on 644 music tracks sourced from two publicly available music datasets, and covering a wide variety of genres. Questions in the benchmark are crafted to assess knowledge and reasoning abilities across several dimensions that cover fundamental musical concepts and their relation to cultural and functional contexts. Through the holistic analysis afforded by the benchmark, we evaluate five open-source models and identify several pitfalls, including an over-reliance on the language modality, pointing to a need for better multimodal integration. Data and code are open-sourced. [1]

## 1. INTRODUCTION

Combining the success of large language models (LLMs) with new advances in machine perception that have led to image, audio and video foundation models [1], multimodal LLMs are becoming influential across many fields [2–6]. Recently, models of this kind have started supporting the audio modality, with a subset also being applied to the music domain [7–13]. We refer to such models exhibiting audio understanding capabilities as *Audio LLMs*. In a nutshell, Audio LLMs consist of pre-trained LLMs whose input space has been expanded beyond text to include tokens from an audio encoder, granting them the ability to produce language outputs that require understanding of both modalities. While promising, these models also inherit many of the limitations of LLMs and little attention has so far been given to their evaluation. In most cases, current automatic evaluation relies on match-based metrics which measure the semantic or lexical overlap between model outputs and reference text. However, many works have pointed out deficiencies in this approach [14], which fails to capture the large space of acceptable language outputs admitted by open-ended tasks. For example, the question *"What are some possible uses for this music in a film or TV show?"* may be suitably answered in many different ways. Secondly, automatic music understanding evaluation via language is only supported by a handful of human-annotated datasets [15–17], of which only one [15] has widely been adopted in the context of Audio LLMs. Instead, many prior works have created a variety of ad-hoc datasets built upon synthetically generated captions from tags and other metadata [8, 9, 18] to train and evaluate their models, without explicit data validation mechanisms, which raises questions around their reliability. These three key issues, lack of standardisation, the inadequacy of text generation metrics, and the quality of annotations in current datasets, pose obstacles to the development of the field and has prompted some to resort to human evaluations [7], which can be costly and are hard to scale and reproduce.

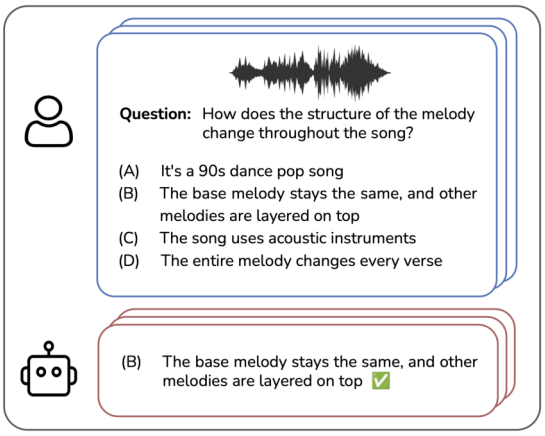


**Figure 1**. **Multiple-choice questions** in MuChoMusic have four answer options of different levels of difficulty.

[1] Data: https://doi.org/10.5281/zenodo.12709974, website: https://mulab-mir.github.io/muchomusic

In this paper, we present *MuChoMusic*, the first benchmark for evaluating music understanding in Audio LLMs. We design a test that is easy to evaluate by collecting a set of multiple-choice (MC) questions that are scrutinised by human annotators, on which simple classification accuracy can be obtained as a reliable indicator of music understanding over the categories covered by the test. The content of our benchmark is intended to be challenging, grounded in factual music knowledge, and tests core understanding and reasoning skills across several dimensions such as music theory, musical styles and traditions, historical and social contexts, structure and expressive analysis. Using MuChoMusic, we carry out a comprehensive evaluation of five existing Audio LLMs with music understanding capabilities. We envision that MuChoMusic will complement prior efforts to standardise music understanding evaluation [19–21] by including this new family of models and steering their early development towards robust progress.

## 2. RELATED WORK

In the music domain, Audio LLMs are commonly evaluated by assessing their text output in the context of a given task defined by an instruction template. Tasks are either designed to test whether the model is able to recognise predefined musical properties such as key (*"What is the key of this song?"*), genre, instrumentation, etc., or they probe for outputs that encompass a variety of musical concepts and that more closely resemble the dialogue format typical of chatbots. Tasks that fall under the former usually mirror canonical MIR tasks and their evaluation leverages standard metrics and benchmarks from the MIR literature. Evaluation of tasks that require broader understanding follows instead less established protocols. Prior works on Audio LLMs most commonly tackle this via two tasks, music captioning (*"Describe the contents of the provided audio in detail"*) [7–9, 11] and music question answering (*"What are some possible uses for this music in a film or TV show?"*) [8, 9]. To perform this kind of evaluation, the authors in [7, 9, 11] make use of the MusicCaps dataset [15], while others [8, 9] employ ad-hoc evaluation datasets created with the help of LLMs. In particular, Liu et al. [8] and Deng et al. [9] propose their own datasets for music question answering, MusicQA and MusicInstruct respectively. These are derived from captions in the MusicCaps dataset or tags from the MagnaTagaTune dataset [22] (MusicQA only), by augmenting them into music-question pairs via pre-trained LLMs. Similarly to these works, we also leverage LLMs to generate our set of questions and answers, but we follow a multiple-choice format to ensure meaningful evaluation and validate all generated data through human annotators to guarantee high data quality.

Finally, we note that concurrent work also proposes evaluation benchmarks for music understanding in LLM-based models [23–25], but these all differ from our work in significant ways: MuChin [23] includes only text in Chinese and does not follow a multiple-choice format, while both MusicTheoryBench (MTB) and ZIQI-Eval focus on the symbolic domain and address the evaluation of text-based LLMs. AIR-Bench [26] includes a small subset of music-related tasks, but puts its focus on audio understanding more generally. We provide an overview of key differences with other benchmarks in Table 1.

| Benchmark | Size | Source(s) | Audio | HC | MC |
|---|---|---|---|---|---|
| MusicQA [8] | 4.5k | MagnaTagATune | ✓ | ✗ | ✗ |
| MusicInstruct [9] | 61k | MusicCaps | ✓ | ✗ | ✗ |
| ZIQI-Eval [25] | 14k | Music books | ✗ | ✗ | ✓ |
| MTB [24] | 372 | (human-written) | ✗ | ✓ | ✓ |
| AIR-Bench [26] | 400 | MusicCaps | ✓ | ✗ | ✓ |
| MuChin [23] | 1k | *unknown* | ✓ | ✓ | ✗ |
| MuChoMusic | 1.2k | MusicCaps, SDD | ✓ | ✓ | ✓ |

**Table 1**. **Comparison of MuChoMusic to existing benchmarks.** HC: human-curated, MC: multiple-choice.

## 3. MUCHOMUSIC

Through MuChoMusic, we aim to alleviate three prominent issues in the evaluation of music understanding in Audio LLMs: a lack of standardisation, the inadequacy of existing text generation metrics, and the quality of current evaluation sets. We address the first two by adopting a multiple-choice format, while our methodical generation and validation procedure attends to the third issue by grounding the data in human-written descriptions and ensuring that the final questions and answers are correct and contextually relevant, as judged by multiple annotators.

### 3.1 Overview

MuChoMusic consists of 1,187 multiple-choice questions aimed at testing the understanding of 644 unique music tracks sourced from the MusicCaps [15] and the Song Describer Dataset [16]. We adopt a multiple-choice format in order to standardise evaluation and follow widespread practice in LLM-centric evaluation scenarios [27–30]. As illustrated in Figure 1, each question has four possible answers. One option is the correct answer, the other three are distractors. Inspired by [31], we structure these as follows: one does not fit the track of interest but is related to the question (*incorrect but related*), one correctly fits the audio, but does not address the question (*correct but unrelated*), and one does not apply to the track and is also irrelevant to the question (*incorrect and unrelated*).

**Evaluation dimensions** MuChoMusic is built from a diverse set of musical works and their detailed descriptions, and serves as a foundation for evaluating Audio LLMs across various dimensions of music comprehension. To delineate the specific evaluation dimensions encompassed by our benchmark, we develop a taxonomy consisting of two primary categories: *knowledge* and *reasoning*. Each category is further divided into several dimensions, informed by insights from national music education programs and existing research on music folksonomies [32]. This structured approach allows us to assess the depth and breadth of music-related competencies systematically, offering a holistic view of models' capabilities in the music domain.

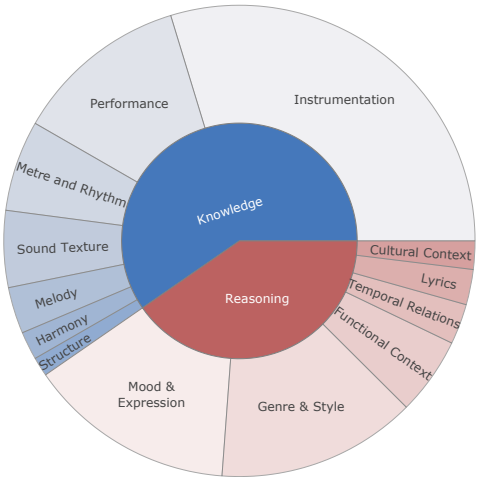


**Figure 2**. **Distribution of evaluation dimensions** covered by MuChoMusic across knowledge and reasoning.

In the *knowledge* category, questions probe a model's ability to recognise pre-acquired knowledge across various musical aspects: (i) melody, (ii) harmony, (iii) metre and rhythm, (iv) instrumentation, (v) sound texture, (vi) performance, and (vii) structure. Questions that test *reasoning* are instead designed to require the synthesis and analytical processing of multiple musical concepts: (i) mood and expression, (ii) temporal relations between elements, (iii) interpretation of lyrics, (iv) genre and style, (v) historical and cultural context, and (vi) functional context. An example of reasoning might involve using an understanding of tempo, chord quality, and instrumentation in concert to ascertain the mood of a music piece. Each question can cover multiple dimensions and their categorisation is obtained automatically, as described in Section 3.2. Figure 2 shows the coverage of the two categories and their respective dimensions within the benchmark. Over half the questions test at least one aspect of musical knowledge, such as features relating to instrumentation or performance characteristics, while 44% are dedicated to probing reasoning skills. While the distribution of dimensions within each category is not balanced, we note that this reflects the distribution of different musical concepts within music captions [16], resulting in categories such as instrumentation, mood and genre appearing more frequently.

### 3.2 Dataset construction

To build our dataset, we automatically transform human-written music captions into multiple-choice questions. These are then carefully validated by multiple human annotators, alongside the associated audio, in order to filter out invalid, ambiguous or irrelevant questions resulting from inaccuracies or hallucinations in the model output.

**Data sources** We source our data from music caption datasets as we aim for elaborate and linguistically diverse information about the music. Currently, only two captioning datasets provide sufficiently detailed music descriptions, namely the Song Describer Dataset (SDD) and MusicCaps. SDD contains 2-minute-long music clips with

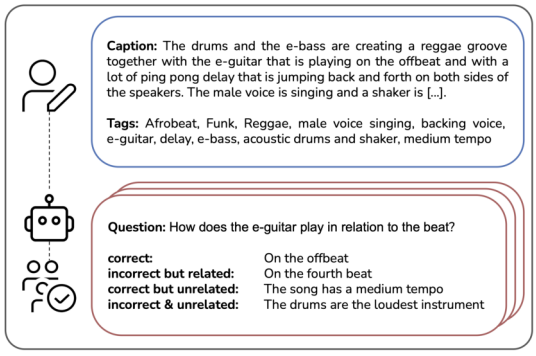


**Figure 3**. **QA generation and validation pipeline**. Example shown here is from MusicCaps [15].

single-sentence captions crowd-sourced from music enthusiasts, while the captions in MusicCaps, describing 10-second audio snippets, are written by professional musicians. From SDD, we select all tracks that have at least two captions, to ensure enough information is provided to the model to be able to formulate interesting and challenging questions. While this is not possible for the MusicCaps dataset, where only one caption is available for each track, we note that descriptions are, on average, longer than in SDD and designed to be more comprehensive. From the genre-balanced subset of the MusicCaps test split, we exclude all tracks for which the labels indicate a low recording quality, to prevent differences in audio quality from affecting the results. For both datasets, we employ a state-of-the-art genre tagging model [33] to identify non-musical tracks and to sub-sample songs from the most common genres (e.g. rock and electronic). Through this curation process, we select 227 unique tracks from SDD and 497 from MusicCaps. We supplement the descriptions with short text labels taken from the dataset itself in the case of MusicCaps and from a related dataset for SDD [34].

**QA generation** We generate the question-answer sets by instructing Gemini 1.0 Pro [3] to formulate question and answer options for a given human-written caption. To leverage the model's in-context learning capability, we prompt it with a detailed task description and three examples of input (description and tags) and expected output. In addition to the question and answer pairs, we ask the model to start its output with a summary of the provided information about the music recording and to interleave the distractor answer options with explanations of their suitability. This way of prompting is inspired by the chain-of-thought methodology and helps to elicit the best model responses [35, 36]. This way, we obtain three multiple choice questions from each description on average and collect a total of 2,091 question-answer pairs. An example of the generated questions is shown in Figure 3.

**Data validation** In order to ensure that questions and answers in our benchmark are factually accurate, aptly written and that each question can be correctly answered based on the available audio, we validate all sets of questions via human annotators. For this step, we recruit 222 participants via the Prolific platform (www.prolific.com). During annotation, a question, the corresponding audio clip, and

all four answer options are presented to the participants in random order, for a total of 30 to 50 question items. Participants are then asked to select all options that correctly answer the question or skip the question by indicating that they are unable to provide an answer or that the question is not valid. Following this procedure, for each question, we collect three to five annotations, stopping early if different annotators are in agreement. This task setup is intended to vet questions and detect those that do not adhere to the intended multiple-choice format, either because the expected correct answer is not the only plausible option or because any one of the distractors is more likely. Consequently, we exclude questions from our final dataset for which i) less than 50% of the annotations indicate the intended correct answer or ii) more than 50% of the annotations mark any of the disctractors as a plausible answer. The final dataset comprises 858 questions from MusicCaps descriptions and the remaining 329 from SDD captions.

**Question categorisation** Once questions are validated, we categorise them according to our taxonomy outlined in Section 3.1. To achieve this, we employ Gemini 1.0 Pro, this time prompting it to automatically label each question with one or more of the evaluation dimensions. The prompt includes the full taxonomy including detailed descriptions of all dimensions, a chain-of-thought instruction, and a single question with only the correct answer. The produced output contains an explanation of the categories and dimensions assigned to each question.

## 4. BENCHMARKING WITH MUCHOMUSIC

We now demonstrate the use of our benchmark, describing our proposed evaluation protocol and metrics, and then detailing our experiments on benchmarking Audio LLMs.

### 4.1 Evaluation Protocol

In multiple-choice-based evaluation, a model is provided with a question and a set of answer options, and is then tasked with selecting the most suitable answer. In practice, this can be accomplished in different ways [29]. In our experiments, we adopt *output-based* evaluation: given a music clip and an associated question-answer set, the language output produced by the model is mapped to one of the candidate options by string matching. Another common approach in MC evaluation is to determine the selected answer through the conditional log likelihood scores of the tokens forming each of the different options. While this can help estimate uncertainty and confidence in the model predictions, in our experiments, we explore only the output-based setting, for three reasons: (1) this corresponds to real-world use of the models, as interactions usually take the form of a conversation; (2) it has a lower computational cost; (3) prior work has demonstrated that sentence probabilities are not necessarily indicative of the probabilities assigned to the answers [37]. To extract the selected answer from the generated outputs, we match either the option identifier (*A*, *B*, *C* or *D*) or the full answer text, if one and only one is given in the output.

| Model | Audio encoder | LLM |
|---|---|---|
| MusiLingo [9] | MERT [39] | Vicuna 7B [40] |
| MuLLaMa [8] | MERT [39] | LLaMA-2 7B [41] |
| M2UGen [12] | MERT [39] | LLaMA-2 7B [41] |
| SALMONN [11] | BEATS [42] & $\text{Whisper}_{\text{large-v2}}$ [43] | Vicuna 7B [40] |
| Qwen-Audio [13] | $\text{Whisper}_{\text{large-v2}}$ [43] | Qwen 7B [44] |

**Table 2**. **Overview of models** we evaluate in our study.

**Metrics** We look at two main metrics to measure model performance on our benchmark: accuracy and instruction following rate (IFR). Accuracy is given by the percentage of correctly answered questions out of the total set of questions. IFR is given by the percentage of generated answers that correspond to one of the given options. In both cases, finegrained scores can be obtained by considering only the subset of questions covering at least one of the available evaluation dimensions shown in Figure 2.

**Adaptation** An important design factor in the evaluation of LLM-based models is adaptation [29], the process of adapting the input to a suitable format. While the format of the audio input is typically fixed by the model design, text inputs allow for more flexibility and different prompting techniques have been shown to significantly influence model's behaviour [35, 36, 38]. Beyond simply passing the question and answer options as the input text, corresponding to *zero-shot prompting*, an effective alternative strategy is to leverage *few-shot in-context learning* (ICL), whereby the model is presented with a set of reference inputs that exemplify the task prior to being shown the question of interest. We experiment with in-context learning in our experiments, providing between 0 and 5 examples in the text input. In the interest of standardisation and to ensure a fair comparison between the models, unless otherwise specified, we keep the prompt selection fixed in our final experiments, following an initial exploration.

### 4.2 Models

In our evaluation, we consider three music-specific models, MuLLaMA [8], MusiLingo [9], and M2UGen [12], and two general-audio LLMs which can be applied to music, as reported in their respective papers, SALMONN [11] and Qwen-Audio [13]. To the best of our knowledge, these are all the existing Audio LLMs which can be applied to music and for which open-source weights are available. These all share a similar architectural design and are composed of a backbone LLM, an audio encoder and a lightweight learnable adapter module to align embeddings produced by the audio encoder to the input space of the LLM, based on either the LLaMA-adapter [45] (MuLLaMA, MusiLingo, M2UGen) or a Q-Former network [46] (SALMONN). An overview of the backbones used in each model is provided in Table 2. All systems are trained via instruction tuning [38, 47] and all employ a combination of different instruction datasets, often in multiple training stages including pre-training and fine-tuning. For all models, we follow the official implementation and use default

| Model | Accuracy | | | IFR |
|---|---|---|---|---|
| | All | Knowledge | Reasoning | All |
| MusiLingo [9] | 21.1 | 22.0 | 19.2 | 71.6 |
| MuLLaMa [8] | 32.4 | 32.3 | 31.3 | 79.4 |
| M2UGen [12] | 42.9 | 44.9 | 41.2 | 96.4 |
| SALMONN [11] | 41.8 | 41.0 | 43.3 | 99.8 |
| Qwen-Audio [13] | 51.4 | 51.1 | 51.0 | 89.7 |
| *Random guessing* | 25.0 | 25.0 | 25.0 | 100.0 |

**Table 3**. Overall **benchmarking results**.

inference settings. We repeat all experiments 3 times, randomly shuffling the order in which answer options are presented, and report average performance across all runs.

## 5. RESULTS AND DISCUSSION

In this section, we first presents findings from our benchmarking experiments, with the goal of elucidating the current state of music understanding in Audio LLMs. We then illustrate how MuChoMusic can be used to derive new insights via a diagnostic analysis, and discuss key takeaways.

### 5.1 Benchmarking Results

We report results for all models in Table 3, showing the overall accuracy score alongside detailed scores on knowledge and reasoning questions, and the instruction following rate (IFR). Figure 4 presents a breakdown of accuracy scores along all reasoning and knowledge dimensions. Unless otherwise specified, we show one-shot performance for all models, as we find this to be the overall optimal setting, as we discuss in more detail in Section 5.2. From this, we observe that current models generally perform poorly across all settings and along all evaluation dimensions. Among these, Qwen-Audio stands out with a score of 51.4%. Surprisingly, with the exception of M2UGen, music-specialised models generally perform worse than general-audio ones, in some cases performing only marginally above or even below random performance. As evidenced by the IFR, these models struggle to output answers in the correct format, which in turn negatively impacts their accuracy score. As shown later in Section 5.3, we find that, when none of the answer options is selected by the model, this is often due to *auditory hallucinations*, *language hallucinations* or *training biases*.

### 5.2 Analysis and Discussion

We now investigate factors influencing performance along different axes by using our benchmark as a diagnostic tool.

**Are models sensitive to prompts?** We first study the effect of varying the number of in-context examples. As shown in Figure 5, providing a single example is occasionally beneficial to accuracy and IFR, but with both the difference magnitude and overall impact differing between models. Additionally, this trend does not hold after the one-shot setting, and we see no consistent improvement when using a larger number of examples. Interestingly,

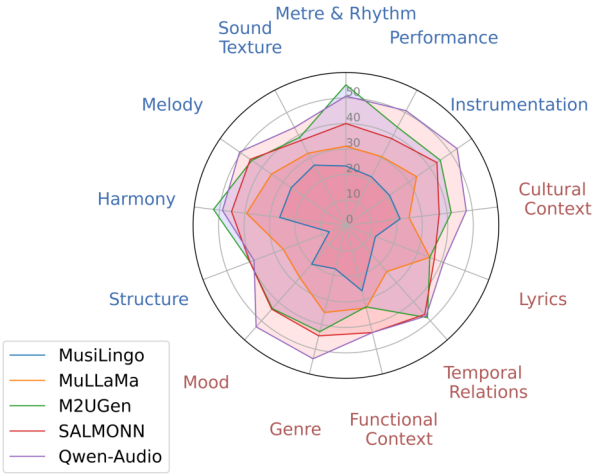


**Figure 4**. **Finegrained accuracy** across evaluation dimensions in knowledge (labelled in blue) and reasoning (red).

we observe that, for M2UGen, Qwen-Audio and MuLLaMa, changes in accuracy from zero- to one-shot prompts are accompanied by a reduction in variance, suggesting that ICL can help minimise variability in the model output. While we do not explore this in our experiments, we also hypothesise that the advantages of ICL may become more prominent through multimodal few-shot prompting [48, 49], which we leave for future work.

**How do models respond to different distractors?** Next, we shift our attention to examining how distractors in our benchmark influence the difficulty of the task. To this end, we ablate answer options corresponding to the different kinds of distractors described in Section 3.2, and present the model with only two or three answer options. In Figure 6(a) we show how performance is affected when using only one distractor alongside the correct option, always randomising their order. From this, we observe that the two distractors containing information which is not related to the question (CU and IU) have a similar effect, while including the *incorrect but related* (IR) option consistently makes the task more challenging. This phenomenon persists when adding a second distractor (not shown here), with combinations which include IR invariably leading to worse performance. Intuitively, the two *unrelated* options can be ruled out based on the text input only, while selecting the correct answer between two options that appear relevant requires engaging multimodal understanding to relate information in the audio content to the text in the question. Crucially, this indicates that models particularly struggle to discern between options that are equally plausible based on the text input only, suggesting that less attention is given to the audio content. This forms the basis of our hypothesis that current Audio LLMs are characterised by a strong language bias, leading to poor performance in tasks that are more audio-dependent. We test this hypothesis in the next section.

**Do models actually pay attention to the audio?** In order to verify whether the audio input is effectively being ignored or is overshadowed by its text counterpart, we de-

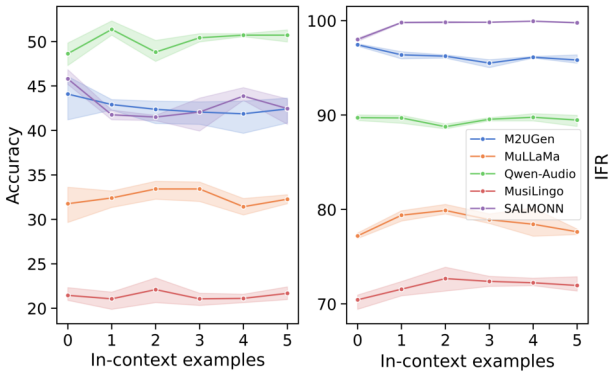


**Figure 5**. **Effect of the number of in-context examples** on accuracy (left) and instruction-following rate (right).

vise a simple test, which we call *audio attention test*, where we replace the audio clip corresponding to a given question with either white Gaussian noise or a randomly chosen track from the dataset. In order to pass this test, a model should display a statistically significant drop in performance when either form of audio perturbation is used, compared to its baseline performance. We showcase results on this test in Figure 6(b). From this, we clearly see that, with the exception of SALMONN and Qwen-Audio, all models fail the audio attention test, and the severity of this failure is often negatively correlated to their overall performance on the benchmark (see Table 3). This confirms that current Audio LLMs are biased towards textual information, often choosing answers that score well under their language prior. Additionally, it provides an explanation for their low performance on the benchmark, as this is effectively bounded by the maximum score they can attain mostly based on the language input. We argue that this constitutes a major pitfall in the design and training procedure of these models, which results in music understanding abilities that do not match the expected performance, as obtained through prior evaluations.

### 5.3 Failure Modes

While the core goal of our benchmark is to provide standardised automatic evaluation to objectively measure general music understanding capabilities, we argue that it can also constitute a useful tool for qualitative assessment. We showcase three examples here, focusing on the two lowest-performing models. While this is not an exhaustive analysis, these examples offer a bird's-eye view of how language pre-training biases percolate through multimodal training, resulting in failures to attend to the inputs in our evaluation. To describe these, we borrow terminology from [50].

**Auditory hallucination** One of the ways models fail to provide a suitable answer falls under the category of *auditory hallucination*, whereby a response includes references to musical elements that are not present in the audio. For example, when asked about an accompaniment instrument, models with this type of hallucination may ignore any suitable option provided (*"acoustic guitar"* or *"strings"*), instead answering *"The song is accompanied by a piano."*, when the audio clip clearly contains no piano.

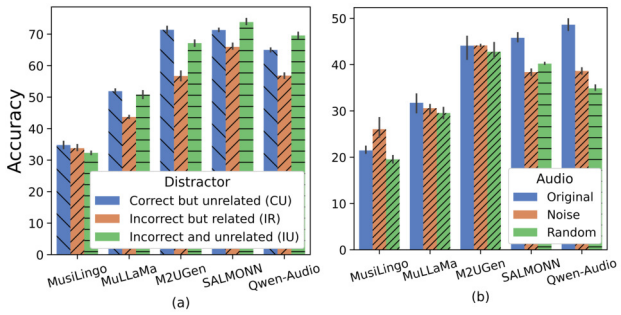


**Figure 6**. (a) **Effect of using different types of distractors**: models tend to perform worst when tasked with distinguishing between two related answers. (b) **Audio attention test**: only some models display a significant drop in performance when provided with incorrect audio inputs. For these experiments, we adopt zero-shot prompting.

**Language hallucination** Another instance of hallucination concerns mundane statements that deviate from the topic of the question altogether. Among others, an observed case of this failure mode is a statement of the form *"The song has a clear and coherent rhythm structure"* to a question specifically asking about the *"type of drum beat"*.

**Training data bias** The last failure mode we encounter is related to a bias towards frequent patterns occurring in the training data. While some of the benchmarked models undergo a stage of training that includes instruction-tuning examples with questions and answers, occasionally they still produce trivial outputs. For example, when asked *"What is the intended purpose of this song?"*, a model with this type of bias may answer *"The intended purpose of this song is not mentioned in the caption"*. Reviewing MusicQA, used in training MuLLaMa and MusiLingo, reveals that a high number of the LLM-generated training examples mention similar phrases, thus likely biasing the model towards this type of uninformative but highly likely output.

## 6. CONCLUSION

We have presented MuChoMusic, a multiple-choice music question answering benchmark designed to test music understanding in Audio LLMs. From an evaluation of five state-of-the-art systems, we find that our benchmark acts as a challenging and informative test, and that current models do not yet leverage both the audio and text modalities fully. All questions in our benchmark are synthesised from human-written music descriptions and manually reviewed to guarantee high data quality. A categorisation of the questions highlights that MuChoMusic offers a broad coverage of areas targeted by current models, and additionally pinpoints gaps that could guide future developments in the field. While we demonstrate that our approach leads to new insights, we note that the multiple-choice format presents some limitations [51]. Therefore evaluation on MuChoMusic should be complemented via further benchmarking efforts to address additional aspects of music understanding through different tasks and formats.

## 7. ETHICS STATEMENT

### 7.1 Annotator welfare

Prior to participation, the annotation experiment described in Section 3.2 was approved by the Queen Mary Ethics of Research Committee to ensure alignment with ethical guidelines and protections for human subjects in research. We did not collect any personal data from our annotators, safeguarding their privacy and confidentiality. Annotators were fully informed about the objectives of the research, the nature of their tasks, and the use of their annotations, underpinning their informed consent before contributing to the project. In an effort to provide a fair compensation for their contributions, annotators were paid £9 per hour.

### 7.2 Biases and fairness

In constructing the MuChoMusic benchmark, our data collection strategy included sourcing music tracks from a variety of backgrounds, acknowledging the inherent challenges in representing the rich diversity of global music cultures within our dataset. We recognise that our initiative does not fully balance the benchmark across all genres, languages, and cultural backgrounds, and annotations were conducted exclusively in English due to logistical constraints, highlighting areas for future expansion and improvement.

## 8. ACKNOWLEDGEMENTS

IM is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Universal Music Group. EB is supported by RAEng/Leverhulme Trust research fellowship LTRF2223-19-106.

## 9. REFERENCES

[1] Rishi Bommasani et al. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021).

[2] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. “Grounding Multimodal Large Language Models to the World”. In: *The Twelfth International Conference on Learning Representations*. 2024.

[3] Gemini Team et al. “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805* (2023).

[4] OpenAI et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

[5] Jean-Baptiste Alayrac et al. “Flamingo: a Visual Language Model for Few-Shot Learning”. In: *Advances in Neural Information Processing Systems*. 2022.

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. “Qwen-vl: A frontier large vision-language model with versatile abilities”. In: *arXiv preprint arXiv:2308.12966* (2023).

[7] Josh Gardner, Simon Durand, Daniel Stoller, and Rachel Bittner. “LLark: A Multimodal Instruction-Following Language Model for Music”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024.

[8] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. “Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024.

[9] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. “MusiLingo: Bridging Music and Text with Pre-trained Language Models for Music Captioning and Query Response”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, 2024.

[10] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. “Pengi: An Audio Language Model for Audio Tasks”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[11] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. “SALMONN: Towards Generic Hearing Abilities for Large Language Models”. In: *The Twelfth International Conference on Learning Representations*. 2024.

[12] Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. “M$^2$UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models”. In: *arXiv preprint arXiv:2311.11255* (2023).

[13] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models”. In: *arXiv preprint arXiv:2311.07919* (2023).

[14] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Joshua P. Gardner, Rohan Taori, and Ludwig Schmidt. “VisIT-Bench: A Dynamic Benchmark for Evaluating Instruction-Following Vision-and-Language Models”. In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023.

[15] Andrea Agostinelli et al. “Musiclm: Generating music from text”. In: *arXiv preprint arXiv:2301.11325* (2023).

[16] Ilaria Manco et al. “The Song Describer Dataset: a Corpus of Audio Captions for Music-and-Language Evaluation”. In: *Machine Learning for Audio Workshop at NeurIPS 2023*. 2023.

[17] Daniel McKee, Justin Salamon, Josef Sivic, and Bryan Russell. “Language-Guided Music Recommendation for Video via Prompt Analogies”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

[18] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. “LP-MusicCaps: LLM-Based Pseudo Music Captioning”. In: *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023*. 2023.

[19] Rachel M. Bittner, Magdalena Fuentes, David Rubinstein, Andreas Jansson, Keunwoo Choi, and Thor Kell. “mirdata: Software for Reproducible Usage of Datasets”. In: *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*. 2019.

[20] Ruibin Yuan et al. “MARBLE: Music Audio Representation Benchmark for Universal Evaluation”. In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023.

[21] Christos Plachouras, Pablo Alonso-Jiménez, and Dmitry Bogdanov. “mir_ref: A Representation Evaluation Framework for Music Information Retrieval Tasks”. In: *37th Conference on Neural Information Processing Systems (NeurIPS), Machine Learning for Audio Workshop*. 2023.

[22] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. “Evaluation of algorithms using games: The case of music tagging”. In: *Proceedings of the 10th ISMIR Conference*. 2009.

[23] Zihao Wang, Shuyu Li, Tao Zhang, Qi Wang, Pengfei Yu, Jinyang Luo, Yan Liu, Ming Xi, and Kejun Zhang. “MuChin: A Chinese Colloquial Description Benchmark for Evaluating Language Models in the Field of Music”. In: *arXiv preprint arXiv:2402.09871* (2024).

[24] Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. “Chatmusician: Understanding and generating music intrinsically with llm”. In: *arXiv preprint arXiv:2402.16153* (2024).

[25] Jiajia Li, Lu Yang, Mingni Tang, Cong Chen, Zuchao Li, Ping Wang, and Hai Zhao. “The Music Maestro or The Musically Challenged, A Massive Music Evaluation Benchmark for Large Language Models”. In: *arXiv preprint arXiv:2406.15885* (2024).

[26] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. “AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension”. In: *arXiv preprint arXiv:2402.07729* (2024).

[27] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. “Measuring Massive Multitask Language Understanding”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021.

[28] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. “Agieval: A human-centric benchmark for evaluating foundation models”. In: *arXiv preprint arXiv:2304.06364* (2023).

[29] Percy Liang et al. “Holistic Evaluation of Language Models”. In: *Transactions on Machine Learning Research* (2023).

[30] Aarohi Srivastava et al. “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Transactions on Machine Learning Research* (2023).

[31] Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. “STARC: Structured Annotations for Reading Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[32] Mohamed Sordo, Fabien Gouyon, Luís Sarmento, Òscar Celma, and Xavier Serra. “Inferring Semantic Facets of a Music Folksonomy with Wikipedia”. In: *Journal of New Music Research* 42.4 (2013).

[33] Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. “Music Representation Learning Based on Editorial Metadata from Discogs”. In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*. 2022.

[34] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. “The MTG-Jamendo Dataset for Automatic Music Tagging”. In: *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*. 2019.

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. “Large Language Models are Zero-Shot Reasoners”. In: *Advances in Neural Information Processing Systems*. 2022.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. "Chain of Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. 2022.

[37] Joshua Robinson and David Wingate. "Leveraging Large Language Models for Multiple Choice Question Answering". In: *The Eleventh International Conference on Learning Representations*. 2023.

[38] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. "Finetuned Language Models are Zero-Shot Learners". In: *International Conference on Learning Representations*. 2022.

[39] Yizhi LI et al. "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training". In: *The Twelfth International Conference on Learning Representations*. 2024.

[40] Wei-Lin Chiang et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. 2023.

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[42] 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 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023.

[43] 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 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023.

[44] Jinze Bai et al. "Qwen Technical Report". In: *arXiv preprint arXiv:2309.16609* (2023).

[45] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. "LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention". In: *The Twelfth International Conference on Learning Representations*. 2024.

[46] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. "Bootstrapping Vision-Language Learning with Decoupled Language Pre-training". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[47] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. "Self-Instruct: Aligning Language Models with Self-Generated Instructions". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023.

[48] Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. "Towards multimodal in-context learning for vision & language models". In: *arXiv preprint arXiv:2403.12736* (2024).

[49] 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". In: *The Twelfth International Conference on Learning Representations*. 2024.

[50] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. "HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[51] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. "Large Language Models Are Not Robust Multiple Choice Selectors". In: *The Twelfth International Conference on Learning Representations*. 2024.