SUPPLEMENTARY MATERIAL FOR 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

Contents

A Datasheet

A.1 Motivation

- **For what purpose was the dataset created?**

  The MuChoMusic dataset was created for the purpose of evaluating music understanding in multimodal audio-language models. Prior to this work, there were no benchmark datasets focusing on the music domain and suitable for testing models which take audio-text inputs and produce language outputs.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  The dataset is a result of a research collaboration undertaken by Universal Music Group International Limited (part of Universal Music Group), Queen Mary University of London, and Music Technology Group (Universitat Pompeu Fabra).

- **Who funded the creation of the dataset?**

  The dataset creation was funded by:

  - UK Research and Innovation [grant number EP/S022694/1]
  - Universal Music Group
  - The Musical AI project - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

- **Any other comments?**

  No.

A.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

  Each instance in the dataset represents one multiple-choice question and four associated answer options, written in English. Each question refers to the audio clip of a music recording from the MusicCaps [1] or the Song Describer Dataset (SDD). [2]

- **How many instances are there in total (of each type, if appropriate)?**

  There are 1,187 instances.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

  The instances included in the dataset are a sample of LLM-generated question-answer sets produced from a subset of the captions in the MusicCaps and Song Describer datasets. The sample used is not necessarily representative of the original datasets, for the following reasons: the original datasets do not uniformly contain captions that are suitable for question-answer generation (e.g. captions may be too short); some audio clips may be too noisy; the genre distribution may not be uniform.

- **What data does each instance consist of?**

  Each instance consists of raw text.

- **Is there a label or target associated with each instance?**

  Each instance consists of a text pair: a question and four answer options. The target is the first answer option, representing the correct answer.

- **Is any information missing from individual instances?**

  There is no missing information.

[1] https://doi.org/10.48550/arXiv.2301.11325
[2] https://doi.org/10.5281/zenodo.10072001

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

  Multiple instances of questions can refer to the same audio clip indicated by a unique identifier.

- **Are there recommended data splits (e.g., training, development/validation, testing)?**

  There is no recommended split, as the dataset is intended to be used solely for evaluation.

- **Are there any errors, sources of noise, or redundancies in the dataset?**

  There are no known errors in the dataset. However, each instance is associated to a set of evaluation dimensions automatically assigned via Gemini 1.0 Pro (version `gemini-1.0-pro-001`). Category assignments have not been manually verified, therefore some noise is expected due to inaccuracies in the model outputs.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

  The dataset is not fully self-contained. Specifically, the audio recordings are not included. The audio recordings associated with the Song Describer Dataset are available for download in the open access repository Zenodo at `https://doi.org/10.5281/zenodo.10072001`, and are guaranteed to be persistent. The audio items of the MusicCaps dataset are not readily available for download and are linked to YouTube videos, which are not guaranteed to be available indefinitely and are subject to the respective licenses.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?**

  There is no confidential information in the dataset.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

  To the best of our knowledge, the dataset does not contain data that can be considered offensive, insulting, threatening, or cause anxiety.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?**

  No.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

  No.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

  No.

- **Any other comments?**

  No.

A.3 Collection Process

- **How was the data associated with each instance acquired?**

  The data was synthetically generated by a large language model based on existing human-written captions, and later validated by human annotators.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses) or sensors, manual human curation, software programs, software APIs)?**

  The Gemini 1.0 Pro model (version `gemini-1.0-pro-001`) [3] was used as the large language model tasked to generate the question-answer text.

[3] `https://cloud.google.com/vertex-ai/generative-ai/docs/learn/model-versioning#gemini-model-versions`

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  Items were first selected from the original datasets as follows: for SDD, we selected tracks that have at least two captions, to ensure enough information is provided to the LLM to be able to formulate interesting and challenging questions; for MusicCaps, we only considered tracks from the 'genre-balanced' subset of the test split, excluding all tracks for which the labels indicate a low recording quality, to prevent large differences in audio quality. We further dropped items for which the corresponding YouTube video is no longer accessible. Finally, to reduce the volume of data in our validation procedure, we selected items according to two criteria: i) we excluded non-musical recordings and ii) limited the number of items in the most prevalent genres (e.g. electronic and rock). To obtain the genre labels for the audio files of both datasets we employ an off-the-shelf tagging model. [4] Through this curation process, we select 227 unique tracks from SDD and 497 from MusicCaps, supplementing the descriptions with short text labels associated to each track.

  After the data generation step, instances were further filtered based on human validation performed by 222 crowdworkers. During validation, participants were presented with a question, the corresponding audio clip, and all four answer options. They were 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 collected three to five annotations, stopping early if different annotators were in agreement. Consequently, we excluded questions from our final dataset for which i) less than 50% of the annotations indicated the intended correct answer or ii) more than 50% of the annotations marked any of the distractors as a plausible answer. The final dataset comprises 858 questions from MusicCaps descriptions and the remaining 329 from SDD captions.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  Crowdworkers were recruited via the Prolific platform to perform data validation. There were 222 participants in total, each paid a rate of £9 an hour. Participants were required to be above 18, be fluent in English, have an active interest in music, have no language-related disorders and have no hearing difficulties.

- **Over what timeframe was the data collected?**

  The raw data was generated in a single session in April 2024. Data validation was done over the course of two weeks in April 2024.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?**

  The project was approved by the Queen Mary Devolved School Research Ethics Committee (QMERC reference number: QMERC20.565.DSEECS24.006). As part of the review, the following documentation was provided: a participant information sheet, a consent form, and an application form with detailed questions about the data collection procedure and potential risks.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

  The data for the validation was collected directly from participants via the Prolific platform. [5]

  Yes, individuals were notified via the text below:

  > Calling all music enthusiasts! Participate in our online study and contribute to academic research on music understanding in machines.
  >
  > What to expect:
  >
  > 1. Listen to short music clips.
  > 2. Answer multiple-choice questions after each clip.
  > 3. Replay clips as needed.
  > 4. Tick all correct answers or flag questions you cannot answer.
  > 5. Use headphones for optimal experience.
  >
  > Your inputs are used to validate AI-generated questions that will be used to evaluate machine learning models. Join now and help us figure out if machines can actually understand music!

[4] `https://essentia.upf.edu/models.html#discogs-effnet`

[5] `https://www.prolific.com/`

- **Did the individuals in question consent to the collection and use of their data?**

  Yes, consent was provided on the data collection platform by ticking the checkboxes with the following statements:

  1. I confirm that I have read the Participant Information Sheet dated 19.02.2024 version 0.2 for the above study; or it has been read to me. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.
  2. I understand that my participation is voluntary and that I am free to stop taking part in the study at any time without giving any reason and without my rights being affected.
  3. I understand that my data will be accessed by the research team.
  4. I understand that my data will be securely stored in secure database server within the UK and in accordance with the data protection guidelines of the Queen Mary University of London in fully anonymised form.
  5. I understand that I can access the information I have provided and request destruction of that information at any time prior to 12.04.24. I understand that following 12.04.24, I will not be able to request withdrawal of the information I have provided.
  6. I understand that the researcher will not identify me in any publications and other study outputs using personal information obtained from this study.
  7. I understand that the information collected about me will be used to support other research in the future, and it may be shared.
  8. I agree to take part in the above study.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

  Participants were made aware of the option to withdraw from the annotation procedure at any point.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**

  No analysis of the potential impact was conducted.

- **Any other comments?**

  No.

A.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

  After validation, questions were categorised according to a predefined taxonomy. To achieve this, we employed Gemini 1.0 Pro, this time prompting it to automatically label each question with one or more of the evaluation dimensions.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

  The raw data was saved together with the output of the labeling process to produce the final dataset.

- **Is the software that was used to preprocess/clean/label the data available?**

  The code is made available as part of the accompanying GitHub repository. [6]

- **Any other comments?**

  No.

[6] `https://github.com/mulab-mir/muchomusic`

A.5 Uses

- **Has the dataset been used for any tasks already?**

  The dataset has been used for the evaluation of audio LLMs as described in the associated publication.

- **Is there a repository that links to any or all papers or systems that use the dataset?**

  A repository that links to papers and systems that use the dataset will be made publicly available following the official release of the dataset.

- **What (other) tasks could the dataset be used for?**

  The dataset is intended purely for evaluation and can be used more broadly for the task of music question answering.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

  No.

- **Are there tasks for which the dataset should not be used?**

  We aim for this dataset to be used for evaluation and benchmarking of audio-language models, thus we discourage using it for training.

- **Any other comments?**

  No.

A.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

  The dataset will be publicly available online.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**

  The dataset will be distributed via Zenodo and will have a digital object identifier (DOI). [7] Additionally, a GitHub repository with download scripts and examples of usage code will also be provided.

- **When will the dataset be distributed?**

  The dataset will be distributed in July 2024.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

  The dataset will be available under the CC BY-SA 4.0 license. [8]

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

  The data is associated with instances from the MusicCaps and Song Describer datasets and is therefore subject to the respective licenses. These can be found in the original dataset repositories. In addition to this, audio instances from the MusicCaps dataset are linked to YouTube videos, and may therefore not be accessible in all countries.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

  No.

- **Any other comments?**

  No.

[7] `https://doi.org/10.5281/zenodo.12709974`
[8] `https://creativecommons.org/licenses/by-sa/4.0/`

A.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

  The dataset will be supported and maintained by Queen Mary University of London and the Music Technology Group (Universitat Pompeu Fabra). The dataset will be hosted on Zenodo and supporting code will be hosted on GitHub.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  Queries about the dataset can be submitted by opening an issue on GitHub or emailing the dataset curators (i.manco@qmul.ac.uk, benno.weck01@estudiant.upf.edu).

- **Is there an erratum?**

  An erratum will be provided in the GitHub repository as necessary.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

  The dataset will be updated to correct errors if necessary. Future versions of the dataset will be released via Zenodo.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

  No personal data was collected and therefore there are no applicable limits on the retention of the data.

- **Will older versions of the dataset continue to be supported/hosted/maintained?**

  Older versions of the dataset will continue to be hosted as they are permanently archived on Zenodo. Updated versions are clearly marked through DOI versioning which also illustrates the obsolescence of the older versions. Any code related to the dataset will be updated to support only the most recent version. This will be explicitly mentioned in the repository.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

  No.

- **Any other comments?**

  No.

B Evaluation Dimensions

Each question tests a set of understanding skills along several dimensions. Each dimension falls under one of two categories: knowledge and reasoning.

- **Knowledge** refers to the ability to recognise pre-acquired musical concepts within a given music clip (e.g. recognising chord progression, recognising an instrument, etc.)
- **Reasoning** refers to the ability to combine knowledge of several musical concepts and typically requires analysis, synthesis, and evaluation (e.g. use the combination of tempo recognition, chord quality and instrumentation to determine the mood of the clip)

B.1 Musical Knowledge Dimensions

Musical knowledge includes both common concepts that are familiar to anyone, including people with no formal music education, and concepts acquired through music education (music literacy).

- Melody
  - Identification of specific musical elements responsible for melody
  - Interval between two notes
  - Contour (ascending, descending, scalar, arpeggiated)
  - Understanding of scale degrees and tonal relationships within melodies
  - Ability to identify and analyse motifs, sequences, themes, and melodic development
- Harmony
  - Identification of specific musical elements responsible for harmony
  - Chord quality (major, minor, diminished, augmented)
  - The function of a given chord within the harmonic progression (tonic, dominant, subdominant)
  - Cadences and harmonic rhythm.
  - Harmonic Analysis: Analysing the harmonic progression, chordal relationships, and tonal framework within a piece of music.
- Metre and Rhythm
  - Time signature
  - Tempo
  - Rhythmic patterns
  - Identification of specific musical elements responsible for metre and rhythm
  - Rhythmic techniques (syncopation, polyrhythm, ostinato, etc.)
- Instrumentation
  - To identify the instrument(s) playing
  - To recognize the family of instruments (e.g., brass, woodwinds, strings) heard in the excerpt
- Sound Texture
  - Timbre
  - Audio effects and their use
  - Sound effects (natural sounds and everyday noises)
- Performance
  - Recording setup
  - Live performance
  - Vocal techniques (e.g. chanting, rapping, falsetto)

  - Dynamics (Variations of loudness)
  - Expressive techniques (dynamics, articulations) used in the excerpt
- Structure
  - The presence of specific sections (e.g., exposition, development, recapitulation in a sonata-allegro form).
  - Transitional passages or changes in thematic material within the excerpt
  - Awareness of sectional divisions (segment type: intro, verse, chorus, etc.)

B.2 Music Reasoning Dimensions

- Mood and expression
  - Emotion, mood or atmosphere conveyed by the music
  - Identification of specific musical elements to convey a certain mood or expression (e.g. expressive performance, phrasing, dynamics)
- Temporal relations between elements
  - Identification of sections, thematic material, and developmental processes
  - Temporal ordering of music events
  - Interaction between different musical elements
- Lyrics
  - Content and theme of the lyrics
  - Emotion, mood or atmosphere conveyed by the lyrics
  - Identify primary language of the lyrics
- Genre and style
  - Recognise characteristics of various musical genres (e.g., jazz, blues, rock, classical)
  - Identify the musical style associated with the excerpt
- Historical and cultural context
  - Identify different historical periods (Baroque, Classical, Romantic, Contemporary, etc.)
  - Relate musical terms related to specific historical periods or styles (e.g., opera, sonata, concerto, symphony) to the audio content
  - Identify the musical tradition the piece belongs to
  - Identify musical influences
- Functional context
  - Understand everyday contexts and activities the music relates to based on its characteristics
  - Recognise possible roles of the music piece in various aspects of life, such as entertainment, relaxation, and cultural expression (e.g. background, TV advert)

C Additional Details on the Evaluation

C.1 Comparison of Benchmarked Models

Here we take a better look at the differences between the models included in our evaluation, from the perspective of model architecture, training data and training mechanism used.

Model	Audio encoder	LLM	Adapter	Dataset	Training	MT
MusiLingo	MERT ❄	Vicuna ❄	Single linear layer + temporal compression, output concatenated to text	MusicInstruct: 60k Q&A pairs (ChatGPT)*	pretraining, instruction tuning	✗
MuLLaMa	MERT ❄	LLaMA-2 ❄	3-layer MLP, output multiplied by queries in self-attention of top LLM layers	MusicQA: 113k Q&A pairs (MPT-7B)*	instruction tuning	✗
M2UGen	MERT ❄	LLaMA-2 ❄	3-layer MLP, output injected into top LLM layer	MUCaps: 220k text-music pairs (MuLLaMa)*	finetuning	✓
SALMONN	BEATS & ❄ Whisper	Vicuna 🔥 (LoRA)	Window-level Q-former	2.3M Q&A pairs (ChatGPT)*	pretraining, instruction tuning, task overfitting	✓
Qwen-Audio	Whisper 🔥	Qwen ❄	-	140k hours of audio	pretraining, supervised finetuning	✓

Table 1. **Overview of models** we benchmark in our study. * indicates synthetic data. ❄ indicates frozen weights and 🔥 fine-tuning/training.