MQAD: A LARGE-SCALE QUESTION ANSWERING DATASET FOR TRAINING MUSIC LARGE LANGUAGE MODELS

Zhihao Ouyang, Ju-Chiang Wang, Daiyu Zhang, Bin Chen, Shangjie Li, Quan Lin

ByteDance

oyzhouyang@gmail.com

ABSTRACT

Question-answering (QA) is a natural approach for humans to understand a piece of music audio. However, for machines, accessing a large-scale dataset covering diverse aspects of music is crucial, yet challenging, due to the scarcity of publicly available music data of this type. This paper introduces MQAD, a music QA dataset built on the Million Song Dataset (MSD), encompassing a rich array of musical features - including beat, chord, key, structure, instrument, and genre — across 270,000 tracks, featuring nearly 3 million diverse questions and captions. MQAD distinguishes itself by offering detailed time-varying musical information such as chords and sections, enabling exploration into the inherent structure of music within a song. To compile MQAD, our methodology leverages specialized Music Information Retrieval (MIR) models to extract higher-level musical features and Large Language Models (LLMs) to generate natural language QA pairs. Then, we leverage a multimodal LLM that integrates the LLaMA2 and Whisper architectures, along with novel subjective metrics to assess the performance of MQAD. In experiments, our model trained on MQAD demonstrates advancements over conventional music audio captioning approaches. The dataset and codes are at https://github.com/oyzh888/MQAD.

***Index Terms*—** Query answering, Music Information Retrieval (MIR), LLM, dataset.

1. INTRODUCTION

Question-Answering (QA) systems provide an intuitive interface for interacting with and understanding music. While recent advancements in Large Language Models (LLMs), such as ChatGPT, have demonstrated ability to generate datasets in various domains [1, 2, 3], applications to Music Information Retrieval (MIR) remain underexplored. As study suggests [4], training LLMs on music-specific QA data enables models to perform a wide range of tasks, reflecting what we refer to as "emergent intelligence" [5, 6, 7, 8]–the ability to provide nuanced answers to complex music-related queries.

Recent advances in multimodal LLMs (MLLMs) have shown considerable promise [1, 2, 3]. Foundational models for vision-language integration are detailed in [9, 1, 10, 11], while [2, 4, 12, 13, 14] are highlighted as seminal works within the audio domain. Despite this progress, the MIR domain lacks a robust baseline, with [14] emerging as a foundational MLLM without offering open-source training dataset, underscoring the community's need for a comprehensive dataset to effectively train MLLMs. While LP-MusicCaps [15] and MU-LLaMA [4] have made notable contributions to the MIR community, they often fall short in handling intricate musical details such as chord progressions, song structure, and rhythmic changes. Existing datasets either focus on music tagging or lack the comprehensive QA capacity needed for in-depth music analysis.

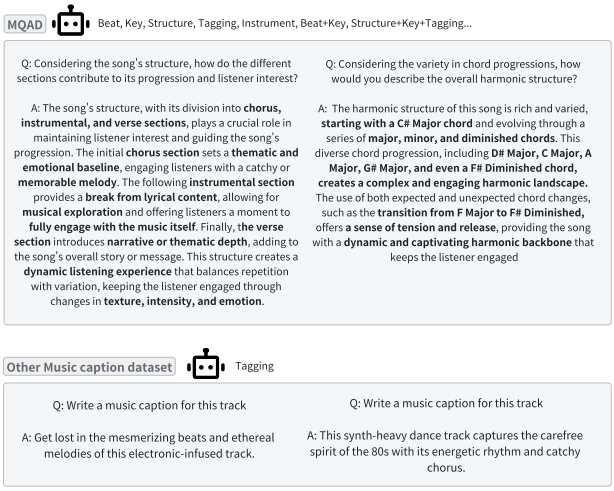

Fig. 1. The MQAD dataset offers diverse coverage of fine-grained MIR facets, making it ideal for training music LLMs.

To address these limitations, we introduce MQAD, a large-scale music QA dataset based on the Million Song Dataset (MSD) [16]. MQAD includes over 270,000 tracks and nearly 3 million QA pairs and captions, covering a wide spectrum of musical features such as beats, chords, key, structural sections, and notes of multiple instruments. As illustrated in Figure 1, the dataset enables LLMs to delve into the temporal and structural aspects of music, offering a richer understanding of its components.

Leveraging MQAD, we train a multimodal LLM that integrates the LLaMA2 [17] and Whisper [12] architectures, yielding state-of-the-art results in music captioning and question answering tasks. Additionally, we introduce a novel evaluation metric based on GPT-4 Turbo [18, 19, 20, 21], designed to simulate human judgment in assessing the quality of music QA systems.

In summary, our contributions are threefold: (1) we compile MQAD, the most extensive music QA dataset to date (with 3 million QA pairs); (2) we develop a multimodal LLM trained on MQAD, demonstrating significant improvements in music captioning; and (3) we introduce new evaluation metrics for assessing music QA systems from multiple perspectives.

Related Works The exploration of music through the lens of QA systems has gained increasing attention alongside the development of LLMs. The MUSIC-AVQA dataset was introduced to support spatio-temporal understanding of musical content [22], offer-

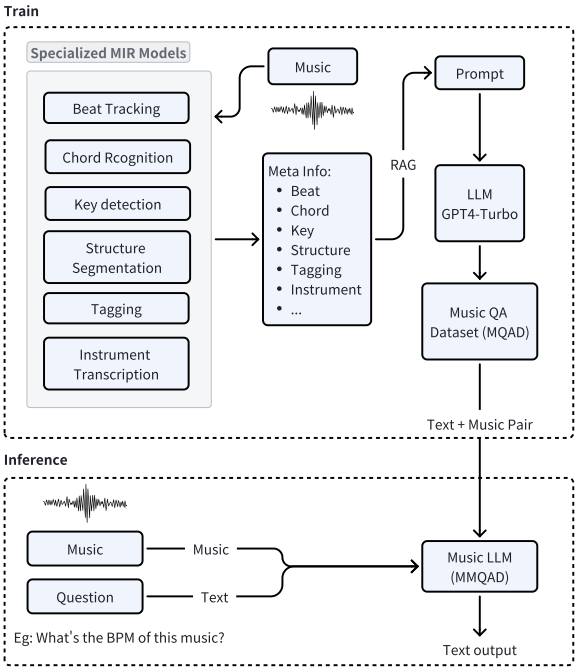

Fig. 2. MQAD dataset construction and MMQAD model training.

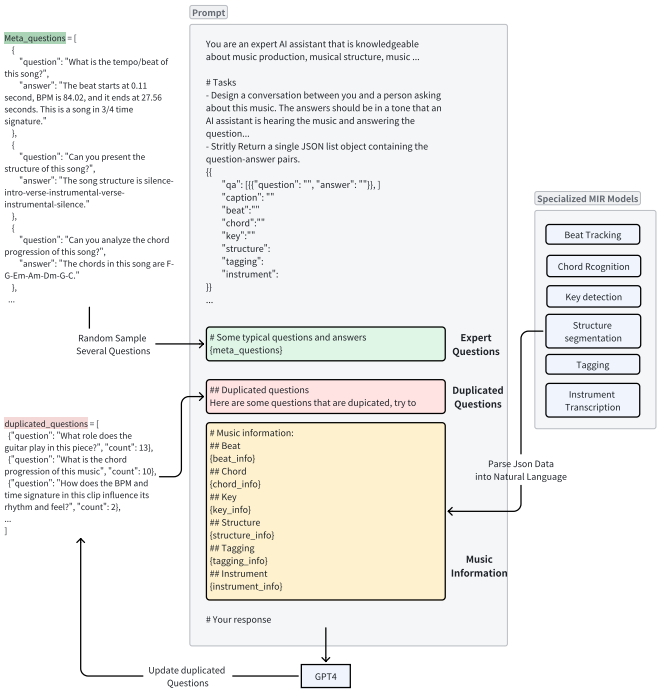

Fig. 3. RAG for generating QA pair and caption.

ing 45K QA pairs across 33 question templates that span multiple modalities and question types. However, its primary focus is on music performance in videos, limiting its scope. In [15], LLMs were employed to generate captions for music, enhancing existing music tagging datasets. However, their approach is limited to general music tags, such as genre, mood, instrument, and tempo, and lacks the capability to provide nuanced temporal music information. Similarly, [23] developed a QA agent system that uses GPT-4 Turbo to classify questions, combined with an autonomous MIR model to handle specific MIR tasks. However, it did not integrate music-specific knowledge into LLM or generate a dataset that could be used by the broader community to train music-focused LLMs. In contrast, [4] introduced a model aimed at improving music captioning through a music question-answering dataset. However, this dataset was small in scale and did not adequately address temporal aspects of music.

2. METHODOLOGY

2.1. MQAD: Music QA Dataset

Data source The MQAD dataset is built upon the Million Song Dataset (MSD) [16], inheriting its vast array of genres and tags to ensure broad coverage. Our selection criteria is based on [24], yielding high-quality samples of about 270k tracks in total (approximately 20% of MSD).

Feature Extraction We utilized specialized MIR models for feature extraction, including beat tracking [25], chord and key detection [26], structure segmentation [27], and vocal and instrument transcription [28]. These models are based on Transformer architecture and demonstrate state-of-the-art or comparable performance on their respective benchmarks. The extracted features are formatted as follows: *Beat*: documented in line format, each beat shows its timestamp and beat count, where a beat_count of 1 indicates a downbeat [29]. *Chord*: noted with their starting and ending times and the chord name [30]. *Structure*: musical sections are listed line-by-line, each showing starting and ending times, with labels such as 'intro', 'verse', 'chorus', 'interlude', 'bridge', 'outro', and 'silence' [31]. *Key*: each musical section is accompanied with a key and a mode of major or minor. *Instrument Transcription*: the transcribed MIDI covers up to 12 tracks of instruments including vocals, bass, drums, guitar, and more [28], formatted in JSON. Each track entry includes a list of polyphonic notes with onset and offset times and a pitch. Empty lists denote the absence of certain instruments.

These textual representations of musical events provide a rich metadata layer, enhancing the generation of QA pairs and allowing LLMs to delve into the dynamic structural compositions of music within a song.

QA Pair Generation We used an LLM to generate text QA pairs from the extracted musical event data and the meta-information provided by MSD. To enhance the diversity and depth of the questions, we integrated a sophisticated Retrieval-Augmented Generation (RAG) system [18, 32], as illustrated in Figure 3. The generation process includes:

- *Backbone prompt*: Following the approach used by LLark [14], we inform GPT with the background, tasks, and system status, applying key modules to ensure high-quality, diverse content.
- *Meta Questions from Music Experts (green block)*: In collaboration with music experts, we developed meta questions covering a wide range of music perspectives. Randomly incorporating these questions into prompts ensures that each QA pair probes different musical aspects, adding specificity and depth.
- *Dynamic Prompt System (red block)*: To mitigate the tendency of LLMs like GPT-4 Turbo to generate repetitive responses, we implemented a dynamic prompting system that includes a de-

Dataset	# item	Duration (h)	C/A	Avg. Token
General Audio Domain				
AudioCaps [33]	51k	144.9	1	9.0±N/A
LAION-Audio [34]	630k	4325.4	1-2	N/A
WavCaps [35]	403k	7568.9	1	7.8±N/A
Music Domain				
MusicCaps [36]	6k	15.3	1	48.9±17.3
LP-MusicCaps-MC [15]	6k	15.3	4	44.9±21.3
LP-MusicCaps-MTT [15]	22k	180.3	4	24.8±13.6
LP-MusicCaps-MSD [15]	514k	4283.1	4	37.3±26.8
MQAD(QA)	804k	11804	4	102.6±23.8
MQAD-Full	3M	28556	11	∼100

Table 1. Comparison of different audio-caption datasets. 'C/A' is the number of captions per audio.

duplication block to suppress questions asked previously.

- *High Diversity Parameter*: We set a high temperature (0.95) using GPT-4 Turbo to further increase the diversity of the generated QA pairs, significantly surpassing the diversity with GPT-3.5.

These optimizations greatly reduce the question duplication rate from over 5% to less than 0.05%. The combined use of a RAG system, expert-derived meta questions, and GPT-4 Turbo makes MQAD one of the most comprehensive music QA datasets available.

Dataset Statistics Table 1 compares various audio-caption datasets. MQAD-Full comprises 4 QA pairs and 7 captions per track in 270k MSD audio clips, totaling 3 million items—5× larger than comparable datasets and 500× larger than MusicCaps [36], one of the most widely used MIR datasets. An example comparison is illustrated in Figure 1. Due to computational constraints, this work focuses on the MQAD(QA) subset, representing 36% (4/11) of the dataset, containing only QA pairs. The remaining 64% (7/11), mainly captions, is reserved for future use. Table 1 highlights the detailed and extensive QA pairs of MQAD, which offers rich MIR information.

2.2. MMQAD: Multimodal LLM

We developed the MMQAD model to validate the MQAD dataset, using LLAMA2-7B [17] as the LLM backbone and Whisper [12] as the audio encoder. Following practices in [2, 4, 12, 13, 14], input text and audio are tokenized by LLAMA2 and Whisper encoders, respectively, and processed by LLAMA2. LLAMA2's strong performance, community support, and resource efficiency made it the ideal choice, enabling MMQAD to handle both textual and auditory inputs for diverse MIR tasks.

The training of MMQAD was conducted on the MQAD dataset, utilizing a subset (i.e., 'qa' key) of the question set specifically tailored for our QA use cases. For all experiments, the input to the encoder is an audio clip of up to 30 seconds at a 16kHz sampling rate, converted to a log-scaled mel spectrogram with 80 mel bins, a 25 ms Hann window, a 10 ms stride, and a 10 ms hop size. All models were trained using the AdamW optimizer with a learning rate of 1e-4. We employed a cosine learning rate decay to zero after a warm-up period of 1000 steps. Two scenarios of training processes are considered: pre-training and fine-tuning. For pre-training, we used a batch size of 256, and the models were trained for 32,768 steps. For fine-tuning and transfer learning, we used a batch size of 64 and trained for 10 epochs. Beam search with 5 beams was employed for the inference of all models. We trained the self-attention layers using LoRA [37] and froze the Whisper encoder to conserve computational resources. For fine-tuning, we slightly decreased the learning rate.

2.3. Subjective and Objective Metrics

For evaluation, we primarily focus on objective metrics as suggested in [15], which include BLEU1 to 4 (B1– 4), METEOR (M), ROUGE-L (R-L), and BERTScore. Other than that, we propose a (pseudo-) subjective metric that leverages LLM as an judger. Research in the NLP domain has shown a high correlation between human judgment and assessments made by LLMs [18, 19, 20, 21], supporting the use of this metric to evaluate the overall quality of MMQAD outputs from a comprehensive perspective. Specifically, we employed GPT4-Turbo to compare predicted answers with the ground truth across eight distinct musical dimensions: Average Accuracy, Average Keywords Matching, and various Average Music Metrics including beat tracking, structural segmentation, chord and key detection, instrumentation, genre, and cultural appropriateness.

3. EXPERIMENTS

We compared MMQAD and MusicCaps-LP models using the test sets of LP-MusicCaps-MC and MQAD.

3.1. Datasets and Models

Three datasets are studied in this work: LP-MusicCaps-MSD, LP-MusicCaps-MC, and our MQAD. *LP-MusicCaps-MSD* consists of approximately 445K training examples. This dataset serves as a pre-training resource for music captioning tasks. *LP-MusicCaps-MC* is more concise dataset with about 2.6K training examples and 2.8K test examples. This is used to assess the model's performance in both supervised and zero-shot settings. *MQAD* is our primary dataset, featuring approximately 804K training QA pairs derived from 213K songs. The validation set contains about 46K QA pairs from 12K songs, and the test set includes 110K QA pairs across 28K songs, with a subset of 500 high-quality questions, providing a wide spectrum for comprehensive evaluation.

Since QA tasks include captioning, we adapted the LP-MusicCaps-MC test set into a QA format for fair comparison. This was achieved by appending the prompt *"write a music caption for this track"* to allow the MMQAD model to process it as a music captioning task. In contrast, MusicCaps-LP, being a native music captioning model designed exclusively for audio input, does not require any textual prompts.

For pre-training, three models are defined as follows: *MMQAD-B* is pre-trained on the combination of LP-MusicCaps-MSD training set and MQAD training set. *MMQAD-C* is pre-trained exclusively on the MQAD training set. *MMQAD-D* is pre-trained on the LP-MusicCaps-MSD training set alone.

For fine-tuning, we define *MMQAD-C+F*, where we further fine-tuned *MMQAD-C* on LP-MusicCaps-MC training set, to verify improvement in music captioning under supervised conditions.

3.2. Result for Music Captioning Task

Table 2 shows the music captioning performance on the LP-MusicCaps-MC test set. In the **pre-training setting**, *MMQAD-C* outperforms other models, particularly in METEOR and ROUGE-L scores, underscoring the advantages of utilizing the MQAD dataset.

In the **fine-tuning setting**, the models enhanced through fine-tuning generally show improved performance, with *MMQAD-C+F* particularly excelling. This highlights the effectiveness of the MQAD dataset in refining music captioning tasks.

Comparing the performance of *MMQAD-C* and *MMQAD-D*, we observe differences between MQAD and LP-MusicCaps-MSD,

Model	Supervised Metrics(%)							Length
	B1	B2	B3	B4	M	R-L	BERT-S	Avg.Token
Baseline								
Supervised Model	28.51	13.76	7.59	4.79	20.62	19.22	87.05	46.7±16.5
Pre-training (Zero-shot Captioning)								
Tag Concat [2, 13]	4.33	0.84	0.26	0.00	3.10	2.01	79.30	23.8±12.1
Template [14]	7.22	1.58	0.46	0.00	5.28	6.81	81.69	25.8±12.4
K2C-Aug [22]	7.67	2.10	0.49	0.10	7.94	11.37	82.99	19.9±7.6
LP-MusicCaps[15]	19.77	6.70	**2.17**	**0.79**	12.88	13.03	84.51	45.3±28.0
MMQAD-B (Ours)	13.16	4.18	1.52	0.54	11.24	13.91	85.18	24.9±5.4
MMQAD-C (Ours)	**21.55**	**7.16**	1.58	0.44	**19.41**	**15.94**	**85.41**	**75.3±12.1**
MMQAD-D (Ours)	10.65	3.17	1.00	0.31	9.52	12.60	85.19	28.0±10.5
Fine-tuning (Transfer Learning)								
Tag Concat [2, 13]	28.65	14.68	8.68	5.82	21.88	21.31	87.67	41.8±14.3
Template [14]	28.41	14.49	8.59	5.78	21.88	21.25	87.72	41.1±13.2
K2C-Aug [22]	29.50	14.99	8.70	5.73	21.97	20.92	87.50	44.1±15.0
LP-MusicCaps[15]	29.09	14.87	**8.93**	**6.05**	**22.39**	21.49	87.78	42.5±14.3
MMQAD-C+F (Ours)	**30.30**	**15.49**	8.80	5.63	22.25	**21.56**	**87.78**	**45.12±13.71**

Table 2. Music captioning results on the MusicCaps test set.

Model	Supervised Metrics(%)							Length
	B1	B2	B3	B4	M	R-L	BERT-S	Avg.Token
LP-MusicCaps[15]	15.24	4.74	1.17	0.28	12.14	13.72	84.48	50.12 ± 14.06
MMQAD-B (Ours)	**51.86**	**35.01**	**25.59**	**19.78**	**40.23**	**36.13**	**91.32**	$\mathbf{75.14 \pm 13.95}$
MMQAD-C (Ours)	51.47	34.65	25.35	19.58	40.04	35.77	91.31	74.85 ± 14.06
MMQAD-D (Ours)	7.33	2.18	0.72	0.25	9.78	13.39	85.49	32.93 ± 18.31

Table 3. Music QA results on the MQAD test set.

suggesting that combining datasets, as in MMQAD-B, may not yield significant performance improvements. This highlights that MMQAD, leveraging a large-scale LLM like LLAMA2-7B, requires diverse QA data for effective training. The fixed-question approach in LP-MusicCaps-MSD may limit its performance. For details on the 'Supervised Model' in Table 2, see [15].

Furthermore, we observe MMQAD's performance appears to be inferior to other models in BLEU-3 and BLEU-4. This can be attributed to its tendency to generate more detailed responses, averaging 75.3 tokens per output. This verbosity particularly affects the results of longer n-grams (i.e., BLEU-3 and BLEU-4), where the lengthy generated text results in lower scores despite the correctness of the content. In contrast, shorter n-grams like BLEU-1 are less impacted by such verbosity.

3.3. Results for the Music QA Task

Evaluating the music question-answering task poses significant challenges, necessitating both subjective and objective approaches. Given the limited scope of the LP-MusicCaps-MC, which contains fewer than 6,000 entries, it does not sufficiently challenge a model's comprehensive QA capabilities. Therefore, we compiled the MQAD test set, which includes 100K samples featuring detailed MIR questions such as chord progression and music structure.

For **objective evaluation**, Table 3 presents the QA performance comparison on the MQAD test set. Comparing between LP-MusicCaps and MMQAD shows that MMQAD-B achieves the best results across all metrics, slightly surpassing MMQAD-C. However, MMQAD-D underperforms significantly due to its training data being limited to music captions, which do not adequately test its broader QA capabilities.

For **subjective evaluation**, considering the costs associated with using GPT-4, we selected 500 representative cases from our MQAD test set that span a broad spectrum of MIR questions. As indicated in Table 4, MMQAD-B stands out as the top performer, which is in line with the findings in Table 3. Overall, our models consistently exhibit superior performance across various MIR dimensions, demonstrating the efficacy of our testing suite in evaluating QA quality across a diverse range of metrics.

Metric(%)	-B	-C	-D
Accuracy	**87**	86	26
Keywords Match	**89**	87	28
Beat Tracking	**85**	**85**	24
Structural Segment	**87**	**87**	22
Key Detection	**82**	80	21
Genre and Mood	**86**	85	32
Instrument Presence	**85**	**85**	28
Cultural Appropriateness	**77**	76	22

Table 4. Subjective music QA results of models MMQAD-B, MMQAD-C, and MMQAD-D using GPT-4 on MQAD test set.

4. CONCLUSION

We have presented the MQAD dataset and the MMQAD model, which establish a new paradigm for research in MIR by leveraging the power of large-scale, diverse QA datasets and multimodal LLMs. By offering a detailed exploration of music through the lens of question-answering systems, we have demonstrated the potential for significant advancements in how machines understand and interact with music. For future work, our dataset could enhance text-to-music generation by integrating with MMQAD, enabling more nuanced user queries for temporal control, which may improve the quality and precision of generated outputs.

5. REFERENCES

[1] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[2] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *NeurIPS*, vol. 36, 2024.

[4] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," in *ICASSP*, 2024, pp. 286–290.

[5] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, 2023.

[6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[7] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[8] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[10] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[13] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[14] J. Gardner, S. Durand, D. Stoller, and R. Bittner, "LLark: A multimodal foundation model for music," *arXiv preprint arXiv:2310.07160*, 2023.

[15] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *ISMIR*, 2023.

[16] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *ISMIR*, 2011.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., "LLaMA 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "Gpteval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[19] T.-Y. Hsu, C.-Y. Huang, R. Rossi, S. Kim, C. L. Giles, and T.-H. K. Huang, "GPT-4 as an effective zero-shot evaluator for scientific figure captions," *arXiv preprint arXiv:2310.15405*, 2023.

[20] V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer, "Is GPT-4 a reliable rater? evaluating consistency in gpt-4 text ratings," *arXiv preprint arXiv:2308.02575*, 2023.

[21] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[22] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *CVPR*, 2022, pp. 19108–19118.

[23] D. Yu, K. Song, P. Lu, T. He, X. Tan, W. Ye, S. Zhang, and J. Bian, "Musicagent: An ai agent for music understanding and generation with large language models," *arXiv preprint arXiv:2310.11954*, 2023.

[24] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1208–1212, 2017.

[25] Y.-N. Hung, J.-C. Wang, X. Song, W.-T. Lu, and M. Won, "Modeling beats and downbeats with a time-frequency transformer," in *ICASSP*, 2022, pp. 401–405.

[26] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, "SpecTNT: A time-frequency transformer for music audio," in *ISMIR*, 2021.

[27] J.-C. Wang, Y.-N. Hung, and J. B. Smith, "To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions," in *ICASSP*, 2022, pp. 416–420.

[28] W.-T. Lu, J.-C. Wang, and Y.-N. Hung, "Multitrack music transcription with a time-frequency perceiver," in *ICASSP*, 2023.

[29] M. F. Matthew E. P. Davies, Sebastian Bock, *Tempo, Beat and Downbeat Estimation*, ISMIR Tutorial, Nov. 2021.

[30] J. Pauwels, K. O'Hanlon, E. Gómez, M. Sandler, et al., "20 years of automatic chord recognition from audio," in *ISMIR*, 2019.

[31] J.-C. Wang, J. B. Smith, and Y.-N. Hung, "MuSFA: Improving music structural function analysis with partially labeled data," *ISMIR Late Breaking & Demo*, 2022.

[32] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[33] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[34] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023.

[35] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[36] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, et al., "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRa: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.