The Music Maestro or The Musically Challenged, A Massive Music Evaluation Benchmark for Large Language Models

Jiajia Li[1,2], **Lu Yang**[3], **Mingni Tang**[3], **Cong Chen**[4], **Zuchao Li**[3,*], **Ping Wang**[1,2,*], **Hai Zhao**[5]
[1]School of Information Management, Wuhan University, Wuhan, China
[2]Key Laboratory of Archival Intelligent Development and Service, NAAC
[3]School of Computer Science, Wuhan University, Wuhan, China
[4]School of Music, Shenyang Conservatory of Music, Shenyang, China
[5]Department of Computer Science and Engineering, Shanghai Jiao Tong University
{cantata, yang_lu, minnie-tang, zcli-charlie, wangping}@whu.edu.cn

Abstract

Benchmark plays a pivotal role in assessing the advancements of large language models (LLMs). While numerous benchmarks have been proposed to evaluate LLMs' capabilities, there is a notable absence of a dedicated benchmark for assessing their musical abilities. To address this gap, we present ZIQI-Eval, a comprehensive and large-scale music benchmark specifically designed to evaluate the music-related capabilities of LLMs. ZIQI-Eval encompasses a wide range of questions, covering 10 major categories and 56 subcategories, resulting in over 14,000 meticulously curated data entries. By leveraging ZIQI-Eval, we conduct a comprehensive evaluation over 16 LLMs to evaluate and analyze LLMs' performance in the domain of music. Results indicate that all LLMs perform poorly on the ZIQI-Eval benchmark, suggesting significant room for improvement in their musical capabilities. With ZIQI-Eval, we aim to provide a standardized and robust evaluation framework that facilitates a comprehensive assessment of LLMs' music-related abilities. The dataset is available at GitHub[1] and HuggingFace[2].

1 Introduction

In recent years, large language models (LLMs) have made significant advancements, revolutionizing various natural language processing tasks (Li et al., 2022b). These models have showcased their proficiency in tasks such as accessing and reasoning about world knowledge.

Benchmark evaluation has played a crucial role in assessing and quantifying the performance of LLMs across different domains. Specific benchmarks tailored to particular tasks such as coding (Austin et al., 2021), reading comprehension (Li et al., 2022a), and mathematical reasoning (Cobbe et al., 2021), in light of the advancements made by LLMs, are increasingly regarded as inadequate for assessing their comprehensive capabilities. Consequently, there has been a surge in the emergence of more comprehensive benchmarks (Liang et al., 2022; Srivastava et al., 2022).

However, both the specific and comprehensive benchmarks have failed to adequately address the musical capability of large language models. Music is an essential part of human life and culture, and assessing LLMs' comprehension and generation of music presents a unique and challenging task. This oversight emphasizes the necessity for a comprehensive evaluation framework specifically designed to capture the nuances of the musical domain.

Therefore, we present ZIQI-Eval, an extensive and comprehensive music benchmark specifically crafted to assess the music-related abilities of LLMs. ZIQI-Eval comprises a diverse range of questions, systematically organized into 10 major categories and 56 subcategories. These categories cover various aspects of music, including music theory, composition, genres, instruments, and historical context. In addition, this music benchmark actively contributes to the recognition of female music composers. By incorporating valuable content from these composers, it rectifies the gender disparity prevalent in historical literature, fostering advancement and inclusivity within the realm of music scholarship. With over 14,000 carefully crafted data entries, ZIQI-Eval provides a rich and extensive resource for evaluating LLMs' comprehension and generation of music-related content.

Utilizing ZIQI-Eval, we conducted a comprehensive experiment over 16 LLMs, comprising API-based models and open-source models, to evaluate the performance of LLMs in the realm of music. Specifically, we fed music knowledge or the first half of a musical score, along with four options, to the LLMs to assess their ability to select the correct option and provide meaningful explanations.

*Corresponding author.
[1]https://github.com/zcli-charlie/ZIQI-Eval
[2]https://huggingface.co/datasets/MYTH-Lab/ZIQI-Eval

Findings of the Association for Computational Linguistics: ACL 2024, pages 3246–3257
August 11-16, 2024

With an average F1 of just 58.68, even the top-performing model, GPT-4, falls short in demonstrating comprehensive music understanding and generation capabilities. This observation not only exposes the overlooked aspect of music in LLMs but also emphasizes the significance of ZIQI-Eval in bridging this gap and tackling the inherent challenges associated with it.

Our Contribution ***(1)*** We find that existing evaluations of the capabilities of LLMs have overlooked their musical abilities. Therefore, we propose ZIQI-Eval benchmark, a manually curated, large-scale, and comprehensive benchmark for evaluating music-related capabilities. It consists of 10 major categories and 56 subcategories, encompassing over 14,000 data entries. ***(2)*** We conduct evaluations on the music comprehension and music generation capabilities of 16 LLMs and find that almost all of them struggle to understand music effectively, let alone generate it. ***(3)*** We explore the issue of bias in LLMs' music capabilities, focusing on gender bias, racial bias, and region bias. Analysis indicates that over 35% of LLMs exhibit biases, with region bias being the most severe.

2 Related Work

Music Comprehension Inspired by the field of natural language processing (NLP), previous studies represented music as embedding sequences for music understanding. Chuan et al. (2020) and Liang et al. (2020) partition music pieces into distinct, non-overlapping segments of fixed duration, and train embeddings for each segment.

Later, with the development of LLMs, recent research has utilized the modeling capabilities of these models to further enhance the understanding of music (Li et al., 2023b). MidiBERT (Chou et al., 2021) and MusicBERT (Zeng et al., 2021) both utilize pre-trained BERT to tackle symbolic-domain discriminative music understanding tasks. MusicBERT further designs OctupleMIDI encoding and bar-level masking strategy to enhance pre-training with symbolic music data. Gardner et al. (2023) extracts music-related information from an open-source music dataset and uses instruction-tuning to instruct their proposed model LLark to do music understanding, music captioning, and music reasoning. NG-Midiformer (Tian et al., 2023) first processes music pieces into sequences, followed by leveraging N-gram encoder to understand symbolic music.

CLaMP (Wu et al., 2023) retrieves symbolic music information through learning cross-modal representations between natural language and symbolic music.

Music Generation Before the proliferation of LLMs, there are some other traditional methods proposed for music generation, mainly falling into three categories: neural networks, neural audio codecs, and diffusion models.

Engel et al. (2019), Marafioti et al. (2020), Greshler et al. (2021) employ neural network architectures such as CNNs, RNNs, or GANs to achieve music generation. A neural audio codec typically contains an encoder and a decoder. Valenti et al. (2020) follows the typical structure. Some models such as Jukebox (Dhariwal et al., 2020), and MusicLM (Agostinelli et al., 2023) further insert a vector quantizer between the encoder and the decoder to learn a discrete latent representation. A diffusion model iteratively adds Gaussian noise and then learns to reverse the diffusion process to construct desired data samples from the noise. Kong et al. (2020) proposes DiffWave, a non-autoregressive model that converts the white noise signal into structured waveform through a Markov chain. Yang et al. (2023) utilizes latent diffusion approach to generate high-quality music.

Since the advent of LLMs, researchers gradually began to explore the application of LLMs in music domain (Zhang et al., 2020; Li et al., 2023c; Lu et al., 2023; Yuan et al., 2024a; Liu et al., 2024). AudioGen (Kreuk et al., 2022) and MusicGen (Copet et al., 2023) both use an autoregressive transformer-based decoder (Vaswani et al., 2017) that operates on the discrete audio tokens. Macaw-LLM (Lyu et al., 2023) incorporates visual, audio, and textual information by using an alignment module to unite multi-modal features to textual features for LLM to generate response. Some models (Deng et al., 2023; Hussain et al., 2023) exploit the potential of LLM to bridge multi-modal music comprehension and generation.

Benchmark Evaluations Benchmark evaluation plays a crucial role in assessing the development of LLMs. Previous specific benchmarking efforts (Hendrycks et al., 2021; Sakaguchi et al., 2020; Zhang et al., 2023) focused on evaluating certain capabilities of models in individual tasks or single-task types. However, with the advancement of LLMs, these benchmarks have become insuffi-

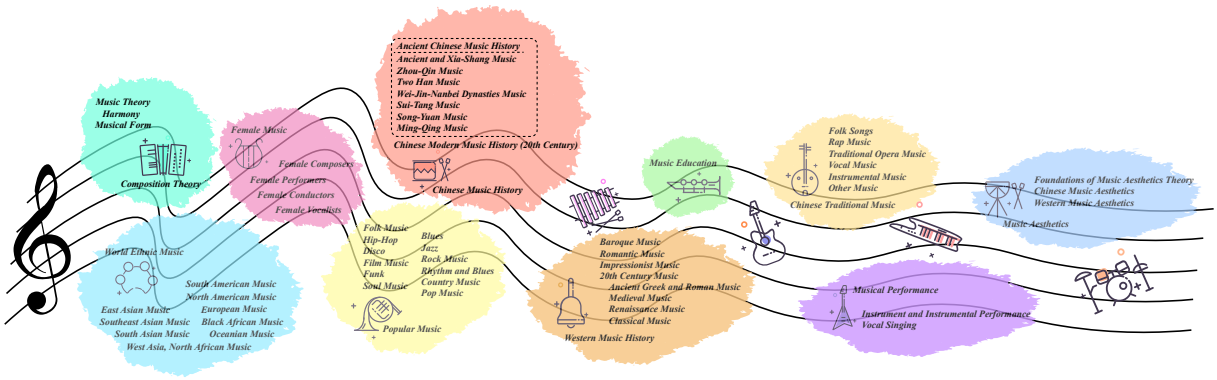


Figure 1: ZIQI-Eval task overview.

cient for comprehensive and accurate assessment of LLM capabilities. Consequently, researchers have proposed more comprehensive and challenging benchmarks (Hendrycks et al., 2020; Li et al., 2023a; Huang et al., 2023) to test whether LLMs possess general world knowledge and reasoning ability. Additionally, there are task-specific evaluations such as LawBench (Fei et al., 2023) and ArcMMLU (Zhang et al., 2023). However, whether in English or Chinese, there is currently only a few benchmarks in music domain (Yuan et al., 2024b; Downie et al., 2014), let alone for evaluating the musical abilities of LLMs, despite music being an important part of human life. Therefore, we propose ZIQI-EVAL, a benchmark for evaluating the musical abilities of LLMs, to fill the gap in benchmark evaluations of LLMs' musical capabilities.

3 ZIQI-Eval Benchmark

3.1 Dataset Curation

General Principle This dataset integrates the renowned music literature database Répertoire International de Littérature Musicale (RILM), providing a broad research perspective and profound academic insights. The inclusion of "The New Grove Dictionary of Music and Musicians" injects the essence of musical humanism into the dataset. Furthermore, dozens of domestic and foreign monographs, such as "Music in Western Civilization" by Paul Henry Lang, the availability of past exam materials from Baidu Wenku, and the advanced data processing capabilities of GPT-4 (Achiam et al., 2023), collectively enhance the data integrity and reliability of the benchmark.

Data Statistics ZIQI-Eval dataset consists of two parts: music comprehension question bank and music generation question bank.

The music comprehension question bank which is presented in the form of multiple-choice questions consists of 10 major categories and 56 subcategories, encompassing 14244 data entries. It not only includes traditional classifications such as music performance, composition theory, and world ethnic music, but also covers popular music, Western music history, Chinese music history, Chinese traditional music, music aesthetics, and music education. The topics range from popular music, rock music, blues, to female music and more. Additionally, the dataset adopts a decentralized design philosophy, fully showcasing the diversity and inclusiveness of global music cultures.

The music generation question bank consists of 200 questions, testing the ability of music continuation. Considering the difficulty in the evaluation of the generated music, the music generation questions are also presented in the form of multiple-choice questions.

We conduct a comprehensive evaluation of LLMs' music capabilities across the entire dataset. It is worth mentioning that this music dataset has made positive contributions in highlighting female music composers. By including relevant content about female composers, it addresses the gender imbalance in historical literature and promotes progress and inclusivity in the music academic community. This initiative not only reflects the benchmark's profound recognition of gender equality issues but also demonstrates its efforts in advancing the diversification of the music field.

3.2 Evaluation Criteria

The evaluation is divided into two parts: music comprehension test and music generation test. The music comprehension test aims to assess the LLMs' music comprehension abilities, specifically their

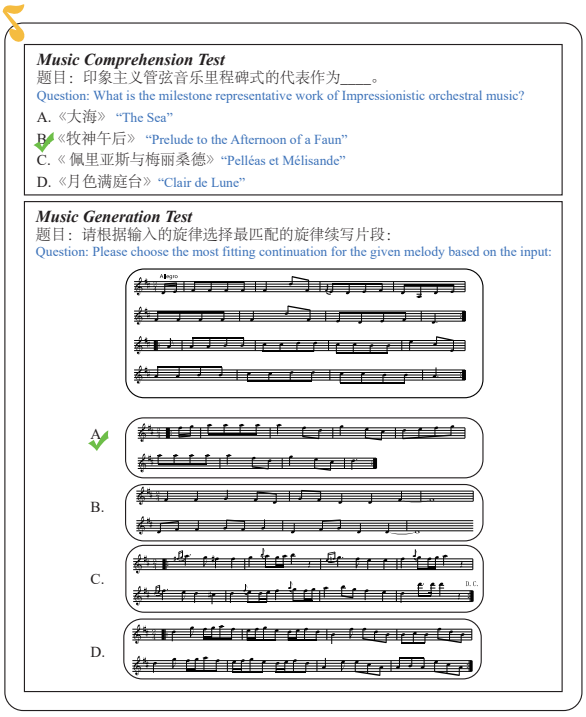


Figure 2: Examples of music comprehension and music generation test.

understanding of music harmony, melody, and rhythm. The music generation test, on the other hand, seeks to evaluate the LLMs' capacities for music generation, namely their ability to generate music across diverse styles and genres.

Music Comprehension Test We turn the music-related knowledge into the question stem and provide them with options to LLMs, making LLMs to choose the right answer. For example, as shown in Figure 2, take "What is the milestone representative work of Impressionistic orchestral music?" as the stem, "The Sea", "Prelude to the Afternoon of a Faun", "Pelléas et Mélisande", and "Clair de Lune" as the options, we examine whether LLMs can select the right answer "Prelude to the Afternoon of a Faun".

Music Generation Test Given that most LLMs can only accept textual inputs, we utilize ABC notation to convert the musical scores of audio into a textual format, which serves as the input for LLMs. We partition the sheet music written in ABC notation into two segments. The initial segment serves as the question, while the subsequent segment presents four alternative options, also in ABC notation, for the potential continuation of the composition. Then we make LLMs discern the most likely continuation fragment, assessing their music continuation ability. For instance, as shown in Figure 2, we split the original score, and test whether LLMs have the ability to choose the most fitting option. To facilitate understanding, we visualize the ABC notation.

4 Experiments

4.1 Setup

Baselines We comprehensively assess 16 LLMs, including API-based models and open-source models. The API-based models contain GPT-4 (gpt-4-1106-preview) (Achiam et al., 2023), GPT-3.5-Turbo (OpenAI, 2022), Claude-instant-1.2 (Anthropic, 2022), and ERNIE-Bot (Baidu, 2023) series. The open-source models contain Aquila-7B (WUDAO, 2023), Bloomz-7B (Muennighoff et al., 2022), ChatGLM2-6B (THUDM, 2023), Mixtral-8x7B (Jiang et al., 2024), XuanYuan-70B (Zhang and Yang, 2023), Qwen (Bai et al., 2023) series, and Yi (Young et al., 2024) series.

Metrics We use a regular expression R, namely $r'[ABCD]'$, to match the answer and consider the first uppercase letter $\in$ {'A', 'B', 'C', 'D'} matched as the response. We define Accuracy (Acc.) as the proportion of correctly answered questions among all questions. Precision is the proportion of correctly answered questions among the questions predicted as A/B/C/D. Recall is the proportion of correctly answered questions among the total number of questions that should be answered as A/B/C/D. In this case, the total number of questions that should be answered as A/B/C/D is actually the total number of questions, so the recall metric is equivalent to the accuracy metric. F1 score is the weighted harmonic mean of precision and recall. The specific formulas for these metrics are as follows:

$$\tilde{X} = G(X)$$
$$\hat{y} = R\left(\tilde{X}\right)$$
$$Precision = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)}{V}$$
$$Recall(Acc.) = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)}{N}$$
$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

where $G(\cdot)$ is the LLM generation process, $\tilde{X}$ is the generated string, $R(\cdot)$ is applying the regular expression for answer retrieval, $\hat{y}$ is the predicted

Models	Music Comprehension Test			Music Generation Test		
	Precision	Recall (Acc.)	F1	Precision	Recall (Acc.)	F1
GPT-4	63.15	62.93	63.04	55.15	53.50	54.31
GPT-3.5-Turbo	55.96	50.18	52.91	31.77	30.50	31.12
Claude-instant-1.2	64.20	45.86	53.50	25.13	25.00	25.06
ERNIE-Bot	74.91	49.96	59.94	30.05	29.00	29.52
ERNIE-Bot-Speed	41.81	31.18	35.72	42.00	42.00	42.00
ERNIE-Bot-Turbo	50.90	47.88	49.34	25.50	25.50	25.50
ERNIE-Bot-8k	53.62	53.17	53.39	30.11	26.50	28.19
Aquila-7B	46.57	29.06	35.79	22.50	9.00	12.86
Bloomz-7B	35.11	31.97	33.47	29.46	19.00	23.10
ChatGLM2-6B	65.80	39.82	49.61	24.80	15.50	19.08
Mixtral-8x7B	43.58	43.39	43.49	31.00	31.00	31.00
XuanYuan-70B	80.73	37.70	51.40	23.60	21.00	22.22
Qwen-7B	35.95	14.54	20.70	24.05	9.74	13.87
Qwen-14B	30.04	17.98	22.49	23.66	15.50	18.73
Yi-6B	60.00	11.06	18.68	0.00	0.00	0.00
Yi-34B	32.24	16.76	22.06	12.12	2.00	3.43

Table 1: Main results(%) of the Music Comprehension Test and Music Generation Test in ZIQI-Eval. Part 1: API-based models; Part 2: Open-source models.

answer, V is the number of questions predicted as A/B/C/D, N is the total number of the questions, and $\mathbb{I}(\cdot)$ is the indicator function.

4.2 Results

Table 1 presents the main results of ZIQI-Eval. Based on the results, we can find that:

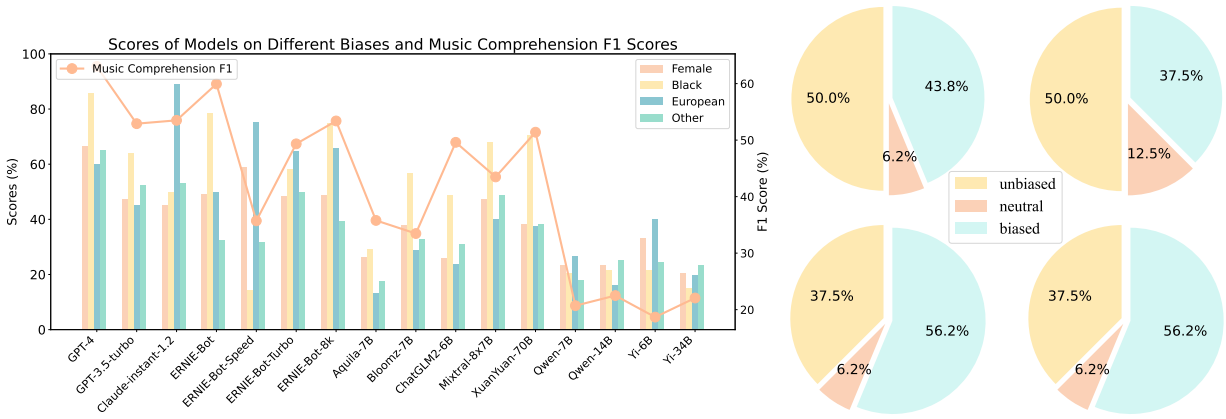
I. Overall, the performance of all LLMs on the ZIQI-Eval benchmark is poor. In both music comprehension test and music generation test, the majority of LLMs have not surpassed the passing threshold of 60. Their F1 scores generally hover between 30 and 50, performing only marginally better than random selection. Even the top-performing model, GPT-4, achieved F1 of only 63.04 and 54.31 in the respective tests. This glaring discrepancy highlights the inadequate consideration given to music abilities within current LLM models and underscores the formidable challenges posed by the ZIQI-Eval benchmark.

II. API-based models perform better than open-source models. In the evaluation of music comprehension test, API-based models generally exhibit higher F1 compared to open-source models. The F1 of API-based models is basically distributed between 50 and 70, while open-source models mostly range between 20 and 50. Only specific open-source models like XuanYuan-70B can achieve an F1 higher than 50.

In the evaluation of music generation questions, API-based models apparently outperform open-source models with significantly higher F1. The highest F1 achieved by an API-based model is 54.31, surpassing the highest F1 of 31.00 in open-source models.

III. The music capabilities of LLMs are dependent but not solely on parameter scale. There is a certain degree of relationship between the musical ability and parameter scale of LLM models within the same series, while the musical ability of LLM models from different series is not strongly correlated with parameter scale.

The Qwen series and Yi series LLMs consistently show improvements in both music comprehension and generation F1. Contrary to expectations, the model with significantly different parameter scales, ChatGLM2-6B and Yi-34B, exhibits higher F1 in music comprehension for the ChatGLM2-6B model, surpassing Yi-34B by 27.55. Even among models with similar parameter scales, there can be considerable differences in performance. For example, the Yi-6B model achieves a music comprehension F1 of only 18.68, while ChatGLM2-6B achieves an F1 of 49.61, resulting in a significant difference of 30.93 between the two F1 scores.



(a) F1 (b) Proportion

Figure 3: Performance of LLMs on gender bias, racial and region bias. Subfigure (a) shows F1 scores of every LLM regarding biases. A line graph is plotted using the average F1 of each LLM to show LLM's overall bias condition. Subfigure (b) depicts the distribution of the biases. Top left: gender bias, top right: racial bias, bottom left: European region, bottom right: other regions.

IV. The instruction-following abilities of LLMs are not directly linked to their music capabilities. The precision scores of LLMs are strongly correlated with their instruction-following abilities. However, a strong instruction-following capability does not necessarily indicate strong musical capabilities in LLMs. Some LLMs may score highly in terms of precision, but they struggle to effectively comprehend and generate music. Yi-6B serves as a clear example where the comprehension precision score reaches 60.00, but the F1 is only equivalent to 18.68.

V. The music generation capabilities of LLMs are in need of improvement. Even though some LLMs demonstrate a decent understanding of music, their music generation capabilities still have room for improvement. In general, the F1 for music generation test in LLMs are lower compared to music comprehension test. The difference can be quite significant, such as ERNIE-Bot-8k, where the score for music comprehension test is higher by 25.20 compared to music generation test.

5 Analysis

In addition to the overall evaluation of LLMs on the dataset, we are also interested in the models' ability for specific categories.

5.1 Does LLM show any bias towards questions related to women?

We compare the F1 of LLMs in the female music theme with the average F1 obtained by LLMs in the female music theme to analyze whether there is bias in LLMs towards female music. We categorize LLMs into three groups: LLMs without gender bias (above the average F1), LLMs with no significant bias (deviating within a range of $\pm 5.0\%$ from the average F1), and LLMs with gender bias (below the average F1).

Models	Female	Black	Region	
			European	Other
GPT-4	66.67	85.71	60.00	64.97
GPT-3.5-turbo	47.37	64.15	45.09	52.49
Claude-instant-1.2	45.16	50.00	88.89	53.13
ERNIE-Bot	49.04	78.57	49.96	32.48
ERNIE-Bot-Speed	59.09	14.29	75.28	31.61
ERNIE-Bot-Turbo	48.26	58.18	64.80	50.03
ERNIE-Bot-8k	48.95	75.00	65.92	39.49
Aquila-7B	26.31	29.27	13.25	17.45
Bloomz-7B	37.87	56.60	28.90	32.94
ChatGLM2-6B	25.88	48.89	23.84	30.95
Mixtral-8x7B	47.32	67.86	40.00	48.76
XuanYuan-70B	38.35	70.54	37.70	38.17
Qwen-7B	23.40	20.51	26.67	18.11
Qwen-14B	23.45	21.74	16.30	25.22
Yi-6B	33.26	21.62	40.00	24.39
Yi-34B	20.63	15.00	19.85	23.29
Average	40.06	48.62	43.53	36.47

Table 2: Results(%) of Female Music F1, Black African Music F1, F1 of European Music and other regions' music in World Ethnic Music.

Because we do not fine-tune LLMs, the results reflect the inherent biases of the LLMs themselves. According to the results of Table 2, 50.00% of the models have no gender bias, 6.25% of the models are neutral or have no significant bias, and 43.75% of the models have gender bias. LLMs with F1

lower than the average F1 tend to overlook relevant content related to female music themes. Yi-34B and Aquila-7B, in particular, have significantly lower scores than the average, indicating a notable gender bias issue in these two models.

5.2 Does LLM exhibit bias toward different races?

We calculate the F1 of LLMs for the subtopic of Black African music, using the same partitioning method as for determining gender bias, to assess whether there is racial bias in LLMs. According to the results of Table 2, 50.00% of the models have no racial bias, 12.50% of the models are neutral or have no significant bias, and 37.50% of the models have racial bias. The F1 of ERNIE-Bot-Speed and Yi-34B are below the mean by 34.33 and 33.62 respectively, indicating a significant racial bias in these two models.

5.3 Does LLM display bias in terms of region?

We seek to investigate whether LLMs are influenced by Eurocentrism, which positions Europe as the cultural and knowledge center, potentially leading to lower evaluations or neglect of contributions from non-central regions and resulting in biases against these regions. To assess the presence of region bias, we computed the F1 for the European Music subtheme within World Ethic Music, and the average F1 for other subthemes within World Ethic Music. Among the LLMs, 43.75% exhibited higher F1 rates in European Music compared to other regional music, while 37.5% of LLMs demonstrated higher F1 rates in European Music than the average F1 rate within European Music. These findings suggest that LLMs are influenced by Eurocentrism and exhibit bias towards non-central regions. Surprisingly, ERNIE-Bot-Speed and Claude-instant-1.2 exhibit significantly higher F1 scores in European music compared to other regions, 43.67 and 35.76 respectively, demonstrating a clear regional bias inclination.

From Figure 3(b), it is evident that LLMs demonstrate similar tendencies towards gender bias, racial bias, and region bias, displaying a trend where both ends (with bias and without bias) are relatively higher, while the middle (neutral) is lower. Some LLMs have F1 scores significantly lower than the mean, such as Yi-34B with an F1 lower than the mean by over 50%, suggesting that LLMs still have a long way to go in eliminating biases, as shown in Figure 3(a). It is worth noting that LLMs with a propensity for gender bias are likely to exhibit racial bias and region bias as well, as evidenced by models such as Aquila-7B, Qwen series, and Yi series. Consequently, it is imperative for future developments in LLMs to address biases comprehensively, not limited to gender, racial and region biases.

6 Futher Analysis

6.1 Phenomenon Analysis of LLMs

To further explore the generation capabilities of LLMs in the realm of music, we conducted an in-depth analysis of the responses provided by each model. Our findings categorize the existing LLMs into three distinct types:

I. Lack of melodic understanding: This type includes LLMs that demonstrate a complete lack of comprehension regarding musical notation. When faced with questions that require the continuation of a melody after a format transformation, these models predominantly resort to evasion, often responding with statements like "Unable to determine, need more information." They fail even to understand the format of the input melody. ChatGLM2-6B and Aquila-7B are prototypical examples of this type, characterized by a high frequency of evasive responses, resulting in a significantly low efficacy in their replies. A notable phenomenon is their tendency to "guess" by consistently selecting option A, leading to most responses without any analytical explanation. For instance, in the responses from ChatGLM2-6B, option A was chosen up to 60%. Besides a preference for option A, Aquila-7B also shows a partiality towards option D.

II. Limited appreciation, misaligned with human preferences: A representative model in this type is ERNIE-Bot-8K. This model provides highly interpretable analyses for each option of every question, offering seemingly logical explanations concerning melody, rhythm, and pitch. However, the model's performance, with F1 barely exceeding that of random selection, underscores the challenge of encapsulating the subjective essence of music appreciation through algorithmic processes. This discrepancy not only highlights the limitations of current AI models in understanding complex, subjective domains but also underscores the need for more sophisticated approaches that can better capture the intricacies of human preferences.

Educational Qualifications and Major	Music Major?	With Music Background?	Score Ranges	Average Score (Precision)
High-school	✗	✗	21 - 51	34.50
Undergraduate	✗	✗	26 - 62	39.91
Master	✗	✗	29 - 60	45.80
Ph.D.	✗	✗	26 - 51	44.00
Undergraduate	✗	✓	27 - 70	45.83
Master	✗	✓	51 - 65	57.75
Ph.D.	✗	✓	64	64.00
High-school	✓	✓	26 - 71	37.81
Undergraduate	✓	✓	30 - 88	51.52
Master	✓	✓	35 - 89	64.75
Ph.D.	✓	✓	28 - 88	64.91
GPT-4	-	-	-	67.54

Table 3: Comparison between GPT-4 and humans with different musical backgrounds and educational qualifications.

	master	Ph.D.	GPT-4
Western Music History	66.28	64.20	75.00
Popular Music	54.78	50.00	81.82
World Ethnic Music	48.54	51.52	61.11
Chinese Traditional Music	61.62	68.94	50.00
Chinese Music History	81.38	79.72	71.43
Female Music	53.29	52.27	75.00
Black African Music	82.89	72.73	100.00
Musical Performance	78.29	68.18	100.00
Music Education	52.63	63.64	50.00
Music Aesthetics	42.54	51.52	50.00
Composition Theory	63.16	67.53	50.00
Film Music	46.05	27.27	100.00
Music Generation	7.89	18.18	25.00

Table 4: Comparison of performance between highly educated individuals and GPT-4 in each category.

III. Relatively good appreciation skills: GPT-4 stands out as a typical example of this type. Its responses consider aspects such as melodic coherence, stylistic similarity, and the seamless integration of musical structures, aligning to a certain extent with human preferences. Further analysis of the questions GPT-4 answered reveals a incorrectly strong inclination towards musical continuity. In many instances, it is observed that GPT-4 prioritized coherence, which leads to the selection of incorrect options.

6.2 Analysis of GPT-4

Taking GPT-4 as a case study, we have gained further insights into the performance of LLMs in the realm of music. The performance of GPT-4 in the domains of female music and world ethnic music indicates a commendable understanding of specific musical areas, reflecting GPT-4's focus on diversity and inclusivity.

GPT-4 has demonstrated exceptional performance in the realm of popular music, achieving scores close to 90. This may be due to the abundant and accessible resources in popular music, including lyrics, genres, and artist information. The popularity and media coverage of pop music may also have facilitated the model's learning efficiency in this field.

It has also scored high in western music history and musical performance, showcasing its capability in processing music history and practical music-making. The higher scores in western music history over all other regions suggest a certain degree of region bias.

In Chinese music history and Chinese traditional music, GPT-4 demonstrates relatively low performance, revealing its deficiencies in handling Chinese musical content.

In the area of music aesthetics, GPT-4 scored low, revealing a significant weakness. This may be attributed to the complexity and subjectivity of music aesthetics, which might surpass the model's ability to learn from existing textual materials, indicating that there is room for improvement in the model's perception, evaluation, and theoretical analysis of music.

Through analysis, we identify that GPT-4 tends to make errors in several distinct categories, primarily falling into three types:

Matching Errors: This category encompasses questions related to musical knowledge, specifically matching-type queries, such as identifying the first Hungarian national opera or the composer of "The Song of the Red Flag". GPT-4's responses often affirmatively state incorrect options, indicating inaccuracies within its knowledge base for specific factual information.

Comprehension Errors: These errors involve

understanding specific musical terminologies and the relationships between certain concepts. Questions like "What function of art does edutainment refer to?" or "What role do work songs play in labor as a genre of folk music?" exemplify where GPT-4 misinterprets multiple word meanings, leading to a misunderstanding of the intended concept. This suggests a need for improvement in GPT-4's understanding within the musical domain.

Reasoning Errors: In instances where GPT-4 correctly understands the question and possesses the relevant knowledge background, errors occur during the reasoning or calculation process, resulting in incorrect conclusions. An example can be seen in questions involving the calculation of musical intervals, where GPT-4 confuses semitones and whole tones. This indicates a gap in GPT-4's ability to perform downstream tasks that require precise musical logical deductions.

7 Human Experiments

We also conduct relevant user studies, and we choose the Precision metric to compare the performance of SOTA LLM (GPT-4) and humans. Due to time and cost limitations, we randomly selected 2 questions for each subcategory (including music comprehension questions and music generation questions), totaling 114 questions. We distributed surveys and recruited 165 individuals with and without a musical background, of different educational qualifications, to answer the questions. Among them, 32.1% are non-music majors without any musical background, 8.5% are non-music majors with some musical background, and 59.4% are music majors with a musical background. As for educational qualifications, 21.2% have a high school diploma, 31.5% have a bachelor's degree, 35.2% have a master's degree or are currently pursuing one, and 12.1% have a doctorate or are currently pursuing one. The results can be seen in Table 3 and Table 4. It can be observed that:

I. Individuals with a musical background tend to score higher. Specifically, the order of scores is as follows: Music majors with a musical background > Non-music majors with some musical background > Non-music majors without a musical background.

II. Individuals with higher educational qualifications tend to score higher. Specifically, the order of scores is as follows: Individuals with a doctorate degree or currently pursuing a doctorate > Individuals with a master's degree or currently pursuing a master's > Undergraduate students > High school students.

III. LLMs have an advantage in terms of knowledge breadth, while music master's and doctoral students have an advantage in their specialized domain knowledge. GPT-4 demonstrates music abilities that are equivalent to, or even surpass, the average level of a music doctoral student (including those currently pursuing a doctorate). This clearly proves the exceptional music capabilities of LLMs. In this regard, the remarkable abilities of LLMs are highly commendable, and it is believed that they will continue to bring us more surprises in the future. However, on the other hand, in certain knowledge domains (such as Chinese Traditional Music, Chinese Music History, Music Education, Composition Theory etc.), it is evident that the scores of music master's or doctoral students are higher. This may be due to their focused and in-depth research in a specific field during the master's and doctoral stages.

8 Conclusion and Future Work

Our research sheds light on the oversight of existing evaluations in recognizing the musical abilities of large models. To address this gap, we introduce ZIQI-Eval, a comprehensive benchmark that encompasses 10 major categories and 56 subcategories, comprising over 14,000 data entries. Notably, this benchmark also actively contributes to the acknowledgment of female music composers, rectifying the gender disparity and promoting inclusivity. We conducted a comprehensive experiment involving 16 LLMs, including both API-based and open-source models, to assess their performance in the domain of music. The results indicate that there is significant scope for enhancing the musical capabilities of existing LLMs. We intend to create a multimodal benchmark to evaluate the musical expertise of LLMs in the future.

Acknowledgement

Jiajia Li, Lu Yang and Mingni Tang made equal contributions to this paper. We are appreciative of the anonymous reviewers for their helpful suggestions that have improved the quality of our paper.

Limitations

Our research to date has been exclusively focused on objective questions, without delving into the study of subjective questions. One limitation of our current music benchmark is the absence of multimodal data. While the benchmark may excel in evaluating and comparing the quality and creativity of musical compositions based on audio data alone, it fails to incorporate other essential aspects of the music experience, such as visual elements or textual information.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Anthropic. 2022. Introducing claude. *Anthropic Blog*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baidu. 2023. Wenxin yiyan. *Baidu Blog*.

Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, Yi-Hsuan Yang, et al. 2021. Midibert-piano: large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv:2107.05223*.

Ching-Hua Chuan, Kat Agres, and Dorien Herremans. 2020. From context to concept: exploring semantic relationships in music with word2vec. *Neural Computing and Applications*, 32:1023–1036.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2023. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

J Stephen Downie, Xiao Hu, Jin Ha Lee, Kahyun Choi, Sally Jo Cunningham, and Yun Hao. 2014. Ten years of mirex: reflections, challenges and opportunities. In *ISMIR 2014*, pages 657–662. ISMIR.

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2023. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*.

Gal Greshler, Tamar Shaham, and Tomer Michaeli. 2021. Catch-a-waveform: Learning to generate audio from a single short example. *NeurIPS*, 34:20916–20928.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *NeurIPS*.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M ^{2} UGen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022a. MultiSpanQA: A dataset for multi-span question answering. In *ACL*, pages 1250–1260.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, et al. 2023b. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*.

Zuchao Li, Ruhan Gong, Yineng Chen, and Kehua Su. 2023c. Fine-grained position helps memorizing more, a novel music compound transformer model with feature interaction fusion. In *AAAI*, pages 5203–5212.

Zuchao Li, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2022b. Text compression-aided transformer encoding. *TPAMI*, 44(7):3840–3857.

Hongru Liang, Wenqiang Lei, Paul Yaozhu Chan, Zhenglu Yang, Maosong Sun, and Tat-Seng Chua. 2020. Pirhdy: Learning pitch-, rhythm-, and dynamics-aware embeddings for symbolic music. In *MM*, pages 574–582.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP*, pages 286–290. IEEE.

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. 2023. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Andres Marafioti, Piotr Majdak, Nicki Holighaus, and Nathanaël Perraudin. 2020. Gacela: A generative adversarial context encoder for long audio inpainting of music. *IEEE*, 15(1):120–131.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI Blog*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *AAAI*, 34(05):8732–8740.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

THUDM. 2023. ChatGLM2. https://github.com/THUDM/ChatGLM2-6B.

Jinhao Tian, Zuchao Li, Jiajia Li, and Ping Wang. 2023. N-gram unsupervised compoundation and feature injection for better symbolic music understanding. *arXiv preprint arXiv:2312.08931*.

Andrea Valenti, Antonio Carta, and Davide Bacciu. 2020. Learning style-aware symbolic music representations by adversarial autoencoders. *arXiv preprint arXiv:2001.05494*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*, 30.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. 2023. Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval. *arXiv preprint arXiv:2304.11029*.

WUDAO. 2023. Aquila. *Github repository*.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

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

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, et al. 2024b. Marble: Music audio representation benchmark for universal evaluation. *NeurIPS*, 36.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. MusicBERT: Symbolic music understanding with large-scale pre-training. In *ACL*, pages 791–800.

Shitou Zhang, Zuchao Li, Xingshen Liu, Liming Yang, and Ping Wang. 2023. Arcmmlu: A library and information science benchmark for large language models. *arXiv preprint arXiv:2311.18658*.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *CIKM*, pages 4435–4439.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *AAAI*, pages 9628–9635.