MuChin: A Chinese Colloquial Description Benchmark for Evaluating Language Models in the Field of Music

Author Name
Affiliation
email@example.com

Abstract

The rapidly evolving multimodal Large Language Models (LLMs) urgently require new benchmarks to uniformly evaluate their performance on understanding and textually describing music. However, due to semantic gaps between Music Information Retrieval (MIR) algorithms and human understanding, discrepancies between professionals and the public, and low precision of annotations, existing music description datasets cannot serve as benchmarks. To this end, we present MuChin, the first open-source music description benchmark in Chinese colloquial language, designed to evaluate the performance of multimodal LLMs in understanding and describing music. We established the Caichong Music Annotation Platform (CaiMAP) that employs an innovative multi-person, multi-stage assurance method, and recruited both amateurs and professionals to ensure the precision of annotations and alignment with popular semantics. Utilizing this method, we built a large-scale, private dataset with multi-dimensional, high-precision music annotations, the Caichong Music Dataset (CaiMD), and carefully selected 1,000 high-quality entries to serve as the test set for MuChin. Based on MuChin, we analyzed the discrepancies between professionals and amateurs in terms of music description, and empirically demonstrated the effectiveness of CaiMD for fine-tuning LLMs. Ultimately, we employed MuChin to evaluate existing music understanding models on their ability to provide colloquial descriptions of music. All data related to the benchmark and the code for scoring have been open-sourced[1].

1 Introduction

As Large Language Models (LLMs) have rapidly advanced, a multitude of LLMs have achieved notable results across various domains [Zhao *et al.*, 2023] and require comprehensive evaluation across benchmarks in different fields [Liang *et al.*, 2022; Huang *et al.*, 2023b; Chang *et al.*, 2023]. Thus, the advancement of LLMs and multimodal technologies necessitates the establishment of benchmarks within the field of music for a unified evaluation. Although benchmarks currently exist for evaluating music understanding models, such as MARBLE [Yuan *et al.*, 2023], which utilizes accuracy on downstream Music Information Retrieval (MIR) tasks as its metric, this does not comprehensively evaluate the capabilities of multimodal large language models.

Music description plays a crucial role in both music understanding [Manco *et al.*, 2021; Gardner *et al.*, 2023] and text-controlled music generation [Agostinelli *et al.*, 2023; Copet *et al.*, 2023]. However, there is currently a lack of benchmark specifically for colloquial music description, which is why we introduce MuChin, the first open-source benchmark for Chinese colloquial music description, with details provided in **Figure 1**.

As models for music understanding [Castellon *et al.*, 2021; Li *et al.*, 2023] and music generation [Zhang *et al.*, 2023; Wang *et al.*, 2023] have evolved, numerous datasets have been proposed, including those derived from Music Information Retrieval (MIR) algorithms or LLMs [Bertin-Mahieux *et al.*, 2011; Wang *et al.*, 2020; Lu *et al.*, 2023; Huang *et al.*, 2023a; Melechovsky *et al.*, 2023] as well as manually annotated datasets [Yang *et al.*, 2017; Bogdanov *et al.*, 2019; Schneider *et al.*, 2023; Zhu *et al.*, 2023; Wang *et al.*, 2022; Agostinelli *et al.*, 2023]. However, these datasets present certain issues that prevent them from serving as comprehensive benchmarks to thoroughly evaluate models' performance in understanding and describing music. Firstly, there is a considerable semantic gap between datasets obtained through algorithms and complex human descriptions. Secondly, current datasets annotated manually are confined to expert annotations and limited descriptive scopes, which significantly diverge from the descriptions provided by the general public [Amer *et al.*, 2013; Mikutta *et al.*, 2014]. And a detailed discussion will be presented in **Section 3.1**. Thirdly, due to limitations in algorithms' performance, datasets generated by MIR cannot achieve complete accuracy, and existing manually annotated datasets, where each entry is annotated by only one person, can also be prone to inaccuracies caused by human errors or biases.

To tackle these challenges, we need to engage both professionals and amateurs in annotating music. This approach will yield two distinct types of music descriptions: one, from pro-

[1] https://muchin.midilib.com/

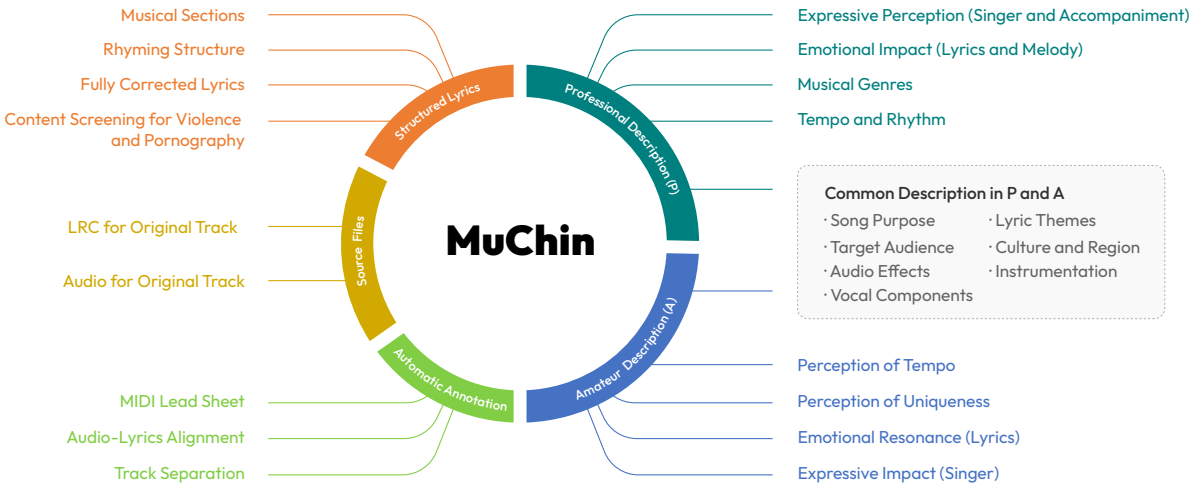

Figure 1: An overview of the MuChin benchmark. The Chinese Colloquial Descriptions consist of Description(A) and Common Description(P & A) annotated by amateur annotators. In addition, we recruit professional annotators to label Description(P), Musical Sections, and Rhyming Structures of the lyrics. And machine-annotated information such as MIDI is also incorporated. These enable MuChin to adapt to a wider range of benchmark tasks.

fessionals, will be rich in technical musical terms, while the other, from amateurs, will resonate with the general public's everyday language. Furthermore, we have introduced a sophisticated, multi-tiered quality assurance process involving multiple individuals at various phases to guarantee the precision of these annotations.

Building on this design, we created a platform that recommends widely-used music descriptors from the internet or specialized terms from the music industry, depending on the input from the annotator. This feature enables annotators to swiftly locate the precise descriptions they need. Additionally, the platform's backend employs a multi-layered, multi-person quality assurance process to verify the accuracy of the annotations. This approach enhances the efficiency, precision, and uniformity of the annotators' descriptions and ensures relevance to the general public by sourcing descriptive terms directly from the web.

With this platform, we have developed a comprehensive, highly accurate, and public-aligned dataset comprising over 100,000 entries, known as the Caichong Music Dataset (CaiMD). From this extensive collection, we meticulously selected 1,000 high-quality entries to serve as a test set, thereby establishing a benchmark for evaluating language models' capabilities in both generating and understanding music-related tasks. Given the precision of these annotated entries, they are also exceptionally suited for fine-tuning pre-trained large language models (LLMs) for a variety of music-related downstream tasks. To this end, we extracted an additional 9,000 entries from CaiMD and subsequently fine-tuned a LLM based on the Qwen [Bai *et al.*, 2023] model. This process demonstrated the effectiveness of our data on fine-tuning LLMs.

MuChin provides a new perspective on the performance of language models in the field of music, requiring the model not only to extract basic attributes from music and describe it from a professional point of view, but also to be able to align with the musical feelings of public users, and describe music in a popular way.

Our Contributions are:

1. We proposed and open-sourced MuChin: the first Chinese colloquial music description benchmark designed to more comprehensively assess the capabilities of multimodal LLMs in the field of music. Utilizing this benchmark, we evaluated the performance of existing music understanding models in terms of their ability to describe music colloquially, as well as the proficiency of current LLMs in generating structured lyrics.

2. We created the Caichong Music Annotation Platform (CaiMAP), implementing a multi-person, multi-stage quality assurance process to guarantee the precision and uniformity of annotations. This approach successfully facilitates efficient annotation of both professional and colloquial music descriptions, including musical sections and rhymes.

3. We built the Caichong Music Dataset (CaiMD): a large-scale, private dataset that is multi-dimensional and high-precision, aligned with the public. It contains over 100,000 entries of music annotations, encompassing information on both professional and colloquial descriptions. Through empirical studies, we demonstrated the effectiveness of the CaiMD on fine-tuning LLMs. Furthermore, we analyzed and verified the discrepancies between professionals and amateurs in terms of music understanding and description.

2 Establishment of MuChin Benchmark

To bridge the gap in benchmarks for language models within the domain of music, specifically targeting Chinese colloquial expressions, we curated and constructed an annotated dataset. This effort led to the creation of the MuChin benchmark.

2.1 Benchmark Tasks

To assess LLMs across multiple dimensions, we included a variety of tasks in our dataset, leading to the creation of MuChin, which is based on the following tasks.

Textual Description Task

Textual descriptions of music involve multi-dimensional representations, including auditory perception, emotions, and music classification. Annotators are required to label and write textual descriptions. Such annotated data sets the stage for benchmarking the ability of multimodal LLMs in understanding music, particularly in tasks like music emotion recognition and classification. Moreover, this data facilitates the evaluation of LLMs' capacity in processing descriptive music texts. Additionally, it can be used to fine-tune LLMs with music-related content.

When annotating textual descriptions, annotators are required to describe music from various aspects, as shown in **Figure 1**. To enhance the efficiency, accuracy, and consistency of annotations, and to align with the public, we built lexicons of music descriptive terms, including a popular term lexicon and a professional term lexicon. The former consists of popular music descriptive terms collected from the internet, while the latter contains keywords extracted from the descriptions of the open-source text-music dataset MusicCaps [Agostinelli *et al.*, 2023]. Annotators have the option to choose appropriate terms from an existing lexicon or, if they find the terms in the lexicon unsatisfactory, they can enhance the descriptions with their own contributions.

Lyric Generation Task

Lyric generation stands as a notable use case for LLMs within the music industry, requiring LLMs to have a profound comprehension of musical structures in order to produce well-organized lyrics. To facilitate this, we construct our dataset to include information on lyric structure, thereby setting a benchmark for assessing LLMs' proficiency in generating lyrics with clear structural distinctions. This involves meticulously defining each section of the lyrics.

Additionally, the ability of LLMs to generate lyrics that align with the theme and rhyme is also crucial. Thus, annotators are required to annotate the main themes and rhymes, as well as to correct any textual errors within the lyrics.

Tasks with Automatic Annotation

Tasks with automatic annotation are discussed in **Appendix A.**

2.2 Preparation and Settings

For the benchmark tasks delineated in **Section 2.1**, it is essential to annotate the data across the corresponding dimensions. Therefore, in this section, we will undertake data preprocessing, along with the recruitment and training of individuals, aiming to secure thorough and high-accuracy annotations.

Data Preprocessing

Data preprocessing, including **music genre clustering**, **track separation**, **audio-lyrics alignment**, and **automatic pre-annotation** is provided in **Appendix B**.

Recruitment and Training of Individuals

To annotate music using both amateur and professional descriptions, it is necessary to engage amateur music enthusiasts for annotating music with popular terms, and professionals – including music students and practitioners – as specialized annotators and quality assurance inspectors. Following this approach, we have recruited 213 individuals familiar with Chinese music through campus and public recruitment efforts. This group includes 109 amateur music enthusiasts and 104 professionals, consisting of 144 males and 69 females, with ages ranging from 19 to 35 years. We have organized these participants into four groups, each assigned specific tasks as follows:

- **Professional Group.** Annotate structures, rhymes and provide professional descriptions.
- **Amateur Group.** Provide colloquial descriptions.
- **Inspector Group.** Evaluate structure annotations, and score music descriptions.
- **Administrator.** Address and provide feedback on inquiries from various groups, and conduct random spot-checks of the groups' outcomes.

The **grouping** and **training method** for each group of individuals are detailed in the **Appendix E**.

2.3 Annotation and Assurance Pipeline

The subsequent phase involves annotation. We have devised an innovative multi-person, multi-stage assurance method aimed at improving quality of annotations and maximizing their accuracy. Additionally, this method serves to objectively evaluate the performance of annotators. Based on this method, we developed the **Caichong Music Annotation Platform (CaiMAP)**, which is introduced in **Appendix D.** The specific annotation pipeline is shown as Figure 2 and will be introduced in this section.

Screening & Structure Annotation Phase

In the screening phase, annotators are required to screen the data carefully. Songs with poor audio quality or content involving pornography or violence that are unsuitable for the dataset should be skipped.

In the structure annotation phase, the platform presents the complete lyrics sentence by sentence, and annotators are required to insert musical section tags between the lyrics. Annotators are also required to check the accuracy of the pre-annotated phonemes and rhymes for each line. If any inaccuracies are found, they should provide their own annotations.

Structure Quality Assurance Phase

To ensure the accuracy of the annotations, we implemented a quality assurance mechanism. Each piece of data undergoes annotation by two separate annotators. Subsequently, the platform autonomously verifies the congruence of the annotations. If they align, the platform seamlessly integrates the data into the dataset for the subsequent phase. In instances of disparities, both sets of annotations are referred to a quality assurance inspector for resolution. The inspector determines the correct annotation or submits an independent correction if necessary.

Figure 2: Pipeline of data annotation and assurance. Each annotated data undergoes 5 complex phases to ensure the accuracy. The figure shows the actual screenshots of the pages for each phase. For **software development** and **operation** details please refer to **Appendix D**.

Description Annotation Phase

Data that successfully clears the structure quality assurance phase becomes eligible for utilization in the music description phase. During this phase, to guarantee attentive listening and thoughtful music descriptions, annotators must listen to each song without interruption. Specifically, annotators are prohibited from writing any textual descriptions within the initial 30 seconds of the song. Copy and paste content is also not allowed. Additionally, limitations are imposed on the number of tags that can be entered and on the word count of user-generated entries.

Description Quality Assurance Phase

Since music description annotation involves subjective judgments and is challenging to assess, the platform employs a randomized selection process, choosing 20% of the annotation results from each annotator for submission to quality assurance inspectors for scoring. These scores are then logged in the platform's backend. Annotated data that successfully pass the sampling quality assurance are submitted into the dataset, whereas those that do not meet the standards are rejected.

Admin Spot-Check & Settlement Phase

Administrators can monitor the real-time progress of each group's work and make payments accordingly, depending on the outcomes of quality assurance checks. Annotators who consistently achieve high pass rates for their annotations will be rewarded additionally, whereas those with lower pass rates will incur penalties, thus motivating them to annotate diligently.

To determine whether the inspectors are competent in their work, administrators also have the access to randomly selected samples of their work for secondary verification.

All the qualified annotated data are incorporated into the **CaiMD**. We provide the subsequent **data processing procedures**, **examples**, and an **overview** in **Appendix F**.

3 Experiments

In this section, we will begin by examining the disparities between professionals and amateurs, thereby underscoring the importance of alignment with public perception. Following that, we will choose several recent language models as benchmarks, encompassing both generative language and music comprehension models. We will then assess their ability to comprehend music, understand musical descriptions, and perform downstream tasks. Through these experiments, our goal is to evaluate the effectiveness of recent language models in the realm of music and to demonstrate our benchmarking approach.

3.1 Discrepancies Between Professionals and Amateurs

To illustrate the substantial disparity between the comprehension and description of music by professionals and amateurs, highlighting the inability of professional descriptions to resonate with the public, we conducted an experiment to gauge the differences in how these two groups articulate various musical attributes across various dimensions.

Analysis Metrics

When a specific type of musical attributes is selected, we calculate the semantic similarity between professionals and amateurs across various dimensions, utilizing the **Semantic Similarity Score** metric which will be detailed in **Section 3.3**.

Results

The results of the discrepancies between professionals and amateurs across various dimensions are as **Figure 3**.

From **Figure 3(a)**, it is evident that there is minimal variance in the multidimensional descriptions of most music genres between the two groups. However, notable disparities arise in their perception of expression in Jazz and Rock, implying significant differences in understanding and describing of expression within progressive genres between professionals and amateurs.

From **Figure 3(b)**, a greater discrepancy between professionals and amateurs is apparent in their interpretations of songs evoking calm and angry emotions, in contrast to those evoking happiness. This underscores the impact of emotions on the comprehension divide between the two groups.

Figure 3(c) reveals substantial disparities in the semantic similarity distribution across various song purposes. This discrepancy suggests that professionals and amateurs have distinct dimensional understandings of music tailored to different intents.

Considering these findings, it becomes evident that professionals and amateurs exhibit varying levels of interpretative disparities across diverse dimensions and music types. Therefore, a comprehensive music description benchmark should accommodate both groups' perspectives.

3.2 Generative LLMs

We utilize MuChin to evaluate existing LLMs in structured lyric generation, including Qwen [Bai *et al.*, 2023],

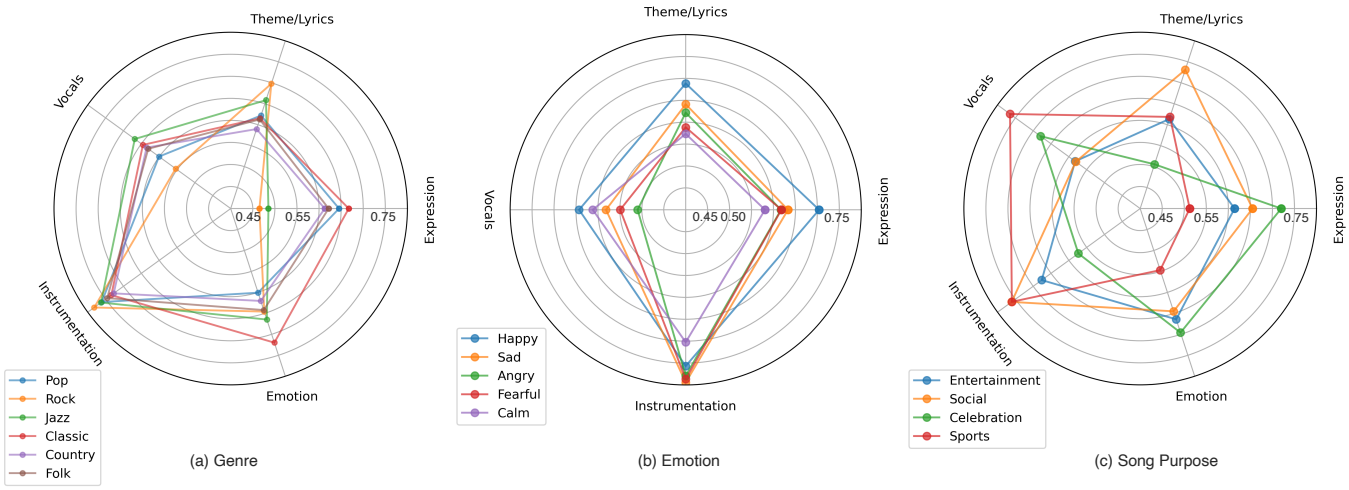

Figure 3: Semantic similarity scores between professionals and amateurs. When a specific type of music is selected, we calculate the similarity between the two groups in various dimensions, for which the **calculation method** is discussed in **Section 3.3**. As a **smaller** value signifies a **larger discrepancy**, the experimental results in this figure reveal significant gaps between the two groups across several specific dimensions.

Baichuan-2 [Baichuan, 2023], GLM-130B [Zeng *et al.*, 2022], and GPT-4 [Achiam *et al.*, 2023]. Moreover, taking into account that Qwen is primarily trained on a Chinese corpus and excels in Chinese language environments, we further refined Qwen by fine-tuning it with 9,000 entries from the CaiMD dataset. Subsequently, we evaluated the performance of this fine-tuned Qwen model on MuChin to assess both the efficacy of our dataset in fine-tuning language and music models, as well as the fine-tuned model's proficiency in comprehending music descriptions and executing associated tasks.

Evaluation Metrics

In assessing the performance of LLMs, we prompt them with music description inputs, asking for structured lyrics along with musical sections and rhymes. While the lyrical content should present subjective diversity, the structural integrity remains objective. Hence, our evaluation primarily centers on the accuracy of the lyric structure rather than its content. We introduce an evaluation method that measures the likeness between the model-generated lyrics and the ground truth across six dimensions outlined below.

- **Song Level.** Song structure similarity measures the similarity between the generated lyrics and the ground truth in terms of overall structure.
- **Section Level.** Section structure similarity measures the similarity between the generated lyrics and the ground truth in terms of musical section labels, order, and the number of sections.
- **Phrase Level.** Phrase structure similarity measures the similarity in the number of phrases within each musical section compared to the ground truth.
- **Word Level.** Word structure similarity measures the similarity between the generated lyrics and the ground truth in terms of the number of words per corresponding phrase.
- **Rhyming Fitting Accuracy.** Rhyme fitting accuracy measures the degree to which the generated lyrics match the ground truth, in terms of end-of-line rhymes.
- **Rhyming Proportion Reasonableness.** To further measure the reasonableness of rhyming, we set an additional award score based on the proportion of rhyming sentences within the overall lyrics, to evaluate the reasonableness of the rhyming proportion in the generated lyrics.

The overall similarity is calculated by computing a weighted average, with weights of 0.10, 0.325, 0.175, 0.20, and 0.20 assigned respectively to the first five dimensions: song, section, phase, word, and rhyming fitting. Additionally, an extra weight of 0.10 is allocated to assess the reasonableness of rhyming proportions.

After comprehensive consideration, the Gestalt algorithm [Ratcliff *et al.*, 1988], which is a universal algorithm for string matching and similarity calculation, is suitable for our lyric evaluation task. Based on the Gestalt algorithm, we propose a scoring algorithm to assess the similarity between generated lyrics and actual lyrics.

The **calculation of the scores** of different dimensions is detailed in **Appendix G**.

Results

Table 1 presents the similarity scores across various dimensions for structured lyrics generated by the selected LLMs in a one-shot scenario, utilizing music descriptions as provided prompts. Notably, all models achieve commendable results. We can observe that among the base models, the overall score increases with the expansion of parameter size. Thanks to its vast parameter size and extensive training data, GPT-4 significantly outperforms the other three models across most dimensions. However, the fine-tuned Qwen, despite having fewer parameters, notably surpasses the untuned base models in overall score and demonstrates a substantial lead in every dimension. This underscores the significant impact of

Model		GPT-4	GLM-4	Baichuan-2	Qwen	
					Base Model	CaiMD Fine-tuned
Parameter Size		1800B	130B	53B	14B	14B
Overall Score		67.08(±6.23)	54.93(±16.46)	49.19(±15.85)	48.31(±13.39)	**85.24(±11.65)**
Structure Similarity	Song Level	2.50(±1.16)	2.29(±0.97)	2.32(±0.99)	2.58(±1.51)	**4.69(±2.38)**
	Section Level	32.40(±0.41)	28.20(±6.75)	28.83(±8.02)	26.49(±4.92)	**32.14(±0.91)**
	Phrase Level	15.52(±2.19)	12.93(±4.31)	12.74(±4.36)	11.59(±3.80)	**17.01(±0.80)**
	Word Level	0.36(±0.79)	0.15(±0.39)	0.01(±0.02)	0.10(±0.23)	**9.12(±5.92)**
Rhyming	Fitting Accuracy	13.88(±3.05)	9.61(±5.17)	4.84(±4.72)	8.01(±4.36)	**16.30(±2.94)**
	Proportion Reasonableness	2.40(±2.66)	1.74(±2.65)	0.45(±1.96)	1.29(±1.89)	**5.98(±4.03)**

Table 1: Evaluation results of the selected LLMs on the benchmark of structured lyric generation. The results are **calculated** by the **formula** detailed in **Appendix G**. A larger value indicates a higher degree of similarity to the corresponding dimension of the actual lyrics, signifying better quality of the generated structured lyrics. For base models, the highest score in each dimension is underlined.

fine-tuning in enhancing the model's capability to comprehend music descriptions and generate structured lyrics. It also suggests considerable potential for improvement in current LLMs within the field of music, emphasizing the importance of CaiMD and MuChin in advancing the development of Chinese LLMs in this domain.

3.3 Music Understanding Models

Analogous to pre-trained language models in NLP, such as BERT [Devlin *et al.*, 2018], a proficient pre-trained music understanding model should be able to effectively represent information across various dimensions within the music, allowing it to be extracted using a simple shallow neural network acting as a decoder. In our benchmark tailored for Chinese music description, we primarily evaluate the capabilities of music understanding models in music description. We select widely employed music understanding models as baselines and evaluate their performance on MuChin. The recent music understanding models include MERT-95M, MERT-330M [Li *et al.*, 2023], Jukebox-5B [Castellon *et al.*, 2021], Music2Vec [Li *et al.*, 2022] and EnCodec [Défossez *et al.*, 2022]. And considering that Jukebox-5B is a pre-trained generative model, not originally designed for music understanding, we use the method in [Castellon *et al.*, 2021] to encode audio with Jukebox-5B.

Evaluation Metrics

To assess the effectiveness of music understanding models, we feed music audio into them and obtain their respective encoding sequences. Subsequently, for each model, we utilize a classifier comprising an average pooling layer and 5 linear layers to extract 10 sets of descriptive music tags corresponding to the dimensions of its output encoding sequences.

- **Semantic Similarity Score.** The BGE model [Xiao *et al.*, 2023], as a general word vector embedding model, has demonstrated impressive performance on various tasks. We utilize the bge-large-zh-v1.5 model to calculate the semantic similarity between the generated and original tags.

For each set of test data, we can ascertain the semantic similarity between them by encoding the tags into embeddings using the BGE model and computing the outer product of these embeddings. Then we sequentially enumerate each generated tag against the original tags, calculate the Semantic Similarity Scores between them, and then obtain the average of all the values as the score of a specific model.

Results

Table 2 demonstrates the semantic similarity scores of the five selected models. It can be observed that, MERT, which encodes both audio and music attributes, performs best in understanding and describing music. Thanks to its massive number of parameters and volume of training data, Jukebox also achieves commendable results. However, as its architecture does not emphasize music attributes, its performance does not reach its full potential.

Moreover, for MERT-95M and MERT-330M, despite their scores being relatively close, we still observe the inverse-scaling effect across multiple dimensions, consistent with the phenomenon mentioned in the paper of MERT [Li *et al.*, 2023]. Specifically, for objective music attributes such as rhythm and instrumentation, MERT-330M performs better, but for most subjective descriptive dimensions, MERT-95M shows superior performance. Therefore, we hypothesize that, in line with the descriptions in the MERT paper, as the amount of data and the number of parameters increase, MERT incorporates more music attribute information, which makes it easier for the model to extract music attributes. However it may lead to a dilution of some audio description-related information. This also indicates that the music attributes extracted by MIR cannot be directly used for music description benchmarks.

4 Related Work

Datasets Based on MIR Algorithms. Datasets based on MIR algorithms employ existing MIR algorithms to extract musical attributes from symbolic music or music audio. And then the attributes are either incorporated into complete descriptive texts or regarded as descriptive tags. MSD [Bertin-Mahieux *et al.*, 2011] collects a million of music data, along with audio, MIDI, and tags retrieved by Echo Nest Analyze API[2] (MIR toolkit). POP909 [Wang *et al.*, 2020] presents

[2]https://developer.spotify.com/

Model		Jukebox	MERT-330M	MERT-95M	Music2Vec	EnCodec
Parameter Size		5B	330M	95M	95M	56M
Data (h)		60 $\sim$ 120k	160k	17k	1k	1k
Professional Description	Average Score-P	0.5490(±0.1458)	0.5586(±0.1433)	**0.5640(±0.1425)**	0.5474(±0.1417)	0.4583(±0.1377)
	Tempo & Rhythm	0.4610(±0.1016)	**0.4650(±0.1013)**	0.4607(±0.0958)	0.4604(±0.1026)	0.4587(±0.1092)
	Emo. Impact (L & M)	0.5312(±0.0939)	0.5350(±0.0903)	**0.5396(±0.0857)**	0.5311(±0.0924)	0.4860(±0.0920)
	Cult. & Reg.	0.5166(±0.2107)	0.5340(±0.2139)	**0.5390(±0.2110)**	0.5120(±0.2094)	0.4072(±0.1261)
	Vocal Components	0.5464(±0.1953)	0.5550(±0.1957)	**0.5713(±0.1989)**	0.5356(±0.1926)	0.4230(±0.1361)
	Song Purp.	0.5810(±0.2191)	0.5864(±0.2166)	**0.6040(±0.2230)**	0.5664(±0.2144)	0.4630(±0.1504)
	Mus. Genres	0.4600(±0.1239)	0.4644(±0.1172)	**0.4692(±0.1158)**	0.4610(±0.1207)	0.4297(±0.1219)
	Exp. Perc. (S & A)	0.9146(±0.0541)	0.9280(±0.0476)	**0.9310(±0.0447)**	0.9190(±0.0576)	0.7085(±0.2888)
	Tgt. Aud.	0.4521(±0.1471)	0.4656(±0.1459)	**0.4683(±0.1417)**	0.4565(±0.1514)	0.3623(±0.0980)
	Instrum.	0.5083(±0.1647)	**0.5180(±0.1587)**	0.5156(±0.1592)	0.5063(±0.1727)	0.4043(±0.1426)
	Audio Eff.	0.5195(±0.1476)	0.5356(±0.1458)	**0.5425(±0.1483)**	0.5244(±0.1539)	0.4404(±0.1122)
Amateur Description	Average Score-A	0.5894(±0.1353)	0.5900(±0.1284)	**0.5923(±0.1284)**	0.5770(±0.1417)	0.4602(±0.1449)
	Perc. of Tempo	**0.4600(±0.1521)**	0.4540(±0.1475)	0.4580(±0.1456)	0.4463(±0.1407)	0.4065(±0.0994)
	Emo. Reson. (L)	0.5977(±0.1780)	0.5894(±0.1798)	**0.6006(±0.1780)**	0.5806(±0.1827)	0.4430(±0.1320)
	Cult.& Reg.	0.4565(±0.1013)	0.4539(±0.0975)	**0.4575(±0.0949)**	0.4510(±0.1023)	0.4324(±0.0972)
	Vocal Components	**0.5195(±0.1208)**	0.5190(±0.1216)	0.5186(±0.1227)	0.5117(±0.12 00)	0.4795(±0.0950)
	Song Purp.	0.5240(±0.2377)	0.5210(±0.2356)	**0.5410(±0.2422)**	0.5201(±0.2428)	0.3801(±0.1532)
	Perc. of Uniq.	0.5356(±0.2076)	0.5356(±0.2115)	**0.5547(±0.2085)**	0.5060(±0.1942)	0.4130(±0.1191)
	Exp. Impact (S)	0.9404(±0.0328)	0.9385(±0.0315)	**0.9460(±0.0315)**	0.9297(±0.0477)	0.7144(±0.2640)
	Tgt. Aud.	0.4417(±0.1041)	0.4448(±0.1114)	**0.4530(±0.0951)**	0.4353(±0.1220)	0.3933(±0.1075)
	Instrum.	0.7144(±0.0737)	**0.7153(±0.0537)**	0.6787(±0.0333)	0.6807(±0.1059)	0.4219(±0.2092)
	Audio Eff.	0.7056(±0.1448)	**0.7275(±0.1465)**	0.7144(±0.1326)	0.7110(±0.1586)	0.5176(±0.1725)

Table 2: Evaluation results of selected music understanding models on the benchmark. The metrics of description presented in the table can be referenced to the **descriptive dimensions** of P and A on the right side of **Figure 1**. After encoding music by the models, we employ an MLP to output descriptive tags corresponding to these dimensions. The **pipeline** of this process can be found in **Appendix H**. The method for calculating the **semantic similarity** scores between the model's output results and the test set labels can be referenced in **Section 3.3**.

a dataset containing audio, lead sheets, and other music attributes like keys and beats. MuseCoco [Lu *et al.*, 2023] and Mustango [Melechovsky *et al.*, 2023] extract features from the original audio and then utilize ChatGPT to incorporate them as descriptions. MuLaMCap in Noise2Music [Huang *et al.*, 2023a] utilizes an LLM to generate a set of music descriptive texts, and then employs MuLan [Huang *et al.*, 2022], a text-music embedding model to match these texts with the music audio in the datasets.

Datasets Based on Manual Annotation. Some datasets based on manual annotations collect descriptions or tags from music websites, while others include data annotated by professional musicians. Hooktheory[3] is a music website where users upload audio with their annotations such as melodies, chords, and beats. MTG [Bogdanov *et al.*, 2019] and Môusai [Schneider *et al.*, 2023] use corresponding tags of music on music websites as descriptive tags, while ERNIE-Music [Zhu *et al.*, 2023] uses comments of music as music descriptions, and establish datasets upon these. Musiclm [Agostinelli *et al.*, 2023] presents a dataset, MusicCaps, including music descriptions annotated by professional musicians.

[3]https://www.hooktheory.com/

Existing Benchmarks in the Field of Music. There are several benchmarks for specific domains in the field of music. Sheet Sage [Donahue and Liang, 2021] presents a benchmark for melody transcription, while GTZAN [Sturm, 2013] presents a test set for music genre classification. MARBLE [Yuan *et al.*, 2023] is a comprehensive benchmark for music understanding models on 4 levels of downstream MIR tasks. However, there is a lack of comprehensive benchmarks focusing on colloquial music description.

5 Conclusion

In this study, we developed an annotation platform called CaiMAP to create a dataset of music descriptions in colloquial Chinese language, termed CaiMD. Leveraging these resources, we introduced the MuChin benchmark, which offers a novel perspective on the performance of language models in the realm of music. MuChin challenges models not only to provide professional-level descriptions of music but also to align with public perceptions.

Despite our efforts to make MuChin as comprehensive and inclusive as possible, it solely addresses tasks related to understanding and generating music descriptions. As such, it does not fully capture the overall capabilities of models in the field of music.

References

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

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

[Amer *et al.*, 2013] Tarek Amer, Beste Kalender, Lynn Hasher, Sandra E Trehub, and Yukwal Wong. Do older professional musicians have cognitive advantages? *PloS one*, 8(8):e71630, 2013.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Baichuan, 2023] Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[Bertin-Mahieux *et al.*, 2011] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. *ISMIR*, 2011.

[Bogdanov *et al.*, 2019] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *ML4MD Machine Learning for Music Discovery Workshop at ICML2019*. ICML, 2019.

[Castellon *et al.*, 2021] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. *arXiv preprint arXiv:2107.05677*, 2021.

[Chang *et al.*, 2023] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

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

[Défossez *et al.*, 2022] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[Défossez, 2021] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[Donahue and Liang, 2021] Chris Donahue and Percy Liang. Sheet sage: Lead sheets from music audio. *Proc. ISMIR Late-Breaking and Demo*, 2021.

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

[Huang *et al.*, 2022] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022.

[Huang *et al.*, 2023a] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023.

[Huang *et al.*, 2023b] 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. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, 2023.

[Li *et al.*, 2022] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, et al. Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. *arXiv preprint arXiv:2212.02508*, 2022.

[Li *et al.*, 2023] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Yike Guo, and Jie Fu. Mert: Acoustic music understanding model with large-scale self-supervised training, 2023.

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

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

[Manco *et al.*, 2021] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[McAuliffe *et al.*, 2017] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502, 2017.

[Melechovsky *et al.*, 2023] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*, 2023.

[Mikutta *et al.*, 2014] CA Mikutta, Gieri Maissen, Andreas Altorfer, Werner Strik, and Thomas König. Professional musicians listen differently to music. *Neuroscience*, 268:102–111, 2014.

[Ratcliff *et al.*, 1988] John W Ratcliff, David Metzener, et al. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46, 1988.

[Rouard *et al.*, 2023] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023.

[Schneider *et al.*, 2023] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Môusai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.

[Sturm, 2013] Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

[Wang *et al.*, 2020] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. *arXiv preprint arXiv:2008.07142*, 2020.

[Wang *et al.*, 2022] Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, et al. Songdriver: Real-time music accompaniment generation without logical latency nor exposure bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1057–1067, 2022.

[Wang *et al.*, 2023] Zihao Wang, Le Ma, Chen Zhang, Bo Han, Yikai Wang, Xinyi Chen, HaoRong Hong, Wenbo Liu, Xinda Wu, and Kejun Zhang. Remast: Real-time emotion-based music arrangement with soft transition. *arXiv preprint arXiv:2305.08029*, 2023.

[Xiao *et al.*, 2023] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[Yang *et al.*, 2017] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

[Yuan *et al.*, 2023] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, et al. Marble: Music audio representation benchmark for universal evaluation. *arXiv preprint arXiv:2306.10548*, 2023.

[Zeng *et al.*, 2022] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[Zhang *et al.*, 2023] Chen Zhang, Yi Ren, Kejun Zhang, and Shuicheng Yan. Sdmuse: Stochastic differential music editing and generation via hybrid representation. *IEEE Transactions on Multimedia*, pages 1–9, 2023.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[Zhu *et al.*, 2023] Pengfei Zhu, Chao Pang, Shuohuan Wang, Yekun Chai, Yu Sun, Hao Tian, and Hua Wu. Ernie-music: Text-to-waveform music generation with diffusion models. *arXiv preprint arXiv:2302.04456*, 2023.