

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

Zihao Wang<sup>1,2</sup>, Shuyu Li<sup>2</sup>, Tao Zhang<sup>2</sup>, Qi Wang<sup>2</sup>, Pengfei Yu<sup>2</sup>, Jinyang Luo<sup>2</sup>, Yan Liu<sup>2</sup>, Ming Xi<sup>2</sup> and Kejun Zhang<sup>1\*</sup>

<sup>1</sup>Zhejiang University, <sup>2</sup>DuiNiuTanQin Co., Ltd.

carlwang@zju.edu.cn {lsyxary, zhangtao8, duoluo7161, cgoxopx, rockyoungljy}@gmail.com,  
liuyan@liao.com, ximing88@gmail.com, zhangkejun@zju.edu.cn

## 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 Cai-chong 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 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 annotated data 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<sup>1</sup>.

## 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 benchmarks 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

\*Corresponding Author

<sup>1</sup><https://github.com/CarlWangChina/MuChin/>

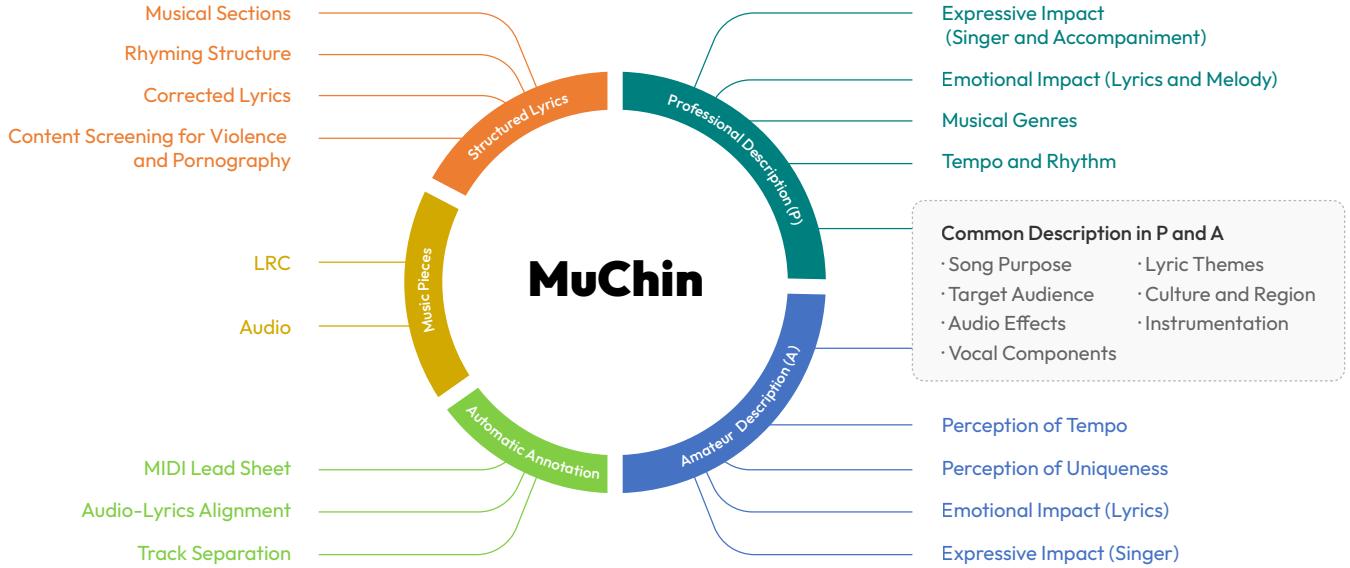


Figure 1: An overview of the MuChin benchmark. The Chinese Colloquial Descriptions consist of Description(A) and Common Description(P & A) annotated by amateur annotators. In addition, we recruit professional annotators to label Description(P), Musical Sections, and Rhyming Structures of the lyrics.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

yield two distinct types of music descriptions: one, from professionals, 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 precision 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, 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 illustrate this point, we have fine-tuned an LLM with an additional dataset, thereby demonstrating its efficacy.

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 dataset that is multi-dimensional and high-precision, aligned with the public. It contains 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, precision, 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-precision 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. Music pieces 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 music piece without interruption. Specifically, annotators are prohibited from writing any textual descriptions within the initial 30 seconds of the music piece. 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 music pieces 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.

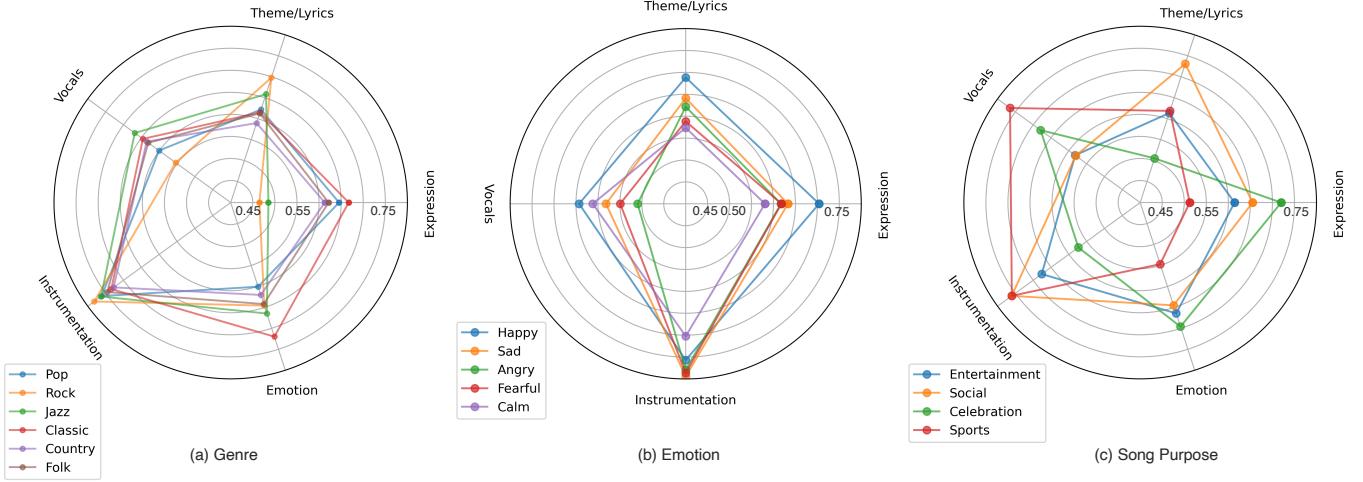


Figure 3: Semantic similarity scores between professionals and amateurs. When a specific type of music is selected, we calculate the similarity between the two groups in various dimensions, for which the **calculation method** is discussed in **Section 3.3**. As a **smaller** value signifies a **larger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

### 3.2 Generative LLMs

We utilize MuChin to evaluate existing LLMs in structured lyric generation, including Qwen [Bai *et al.*, 2023], 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 another batch of data. Subsequently, we evaluated the performance of this fine-tuned Qwen model on MuChin to assess both the efficacy of the data 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 reward 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, phrase, 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 mod-

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

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.

els in overall score and demonstrates a substantial lead in every dimension. This underscores the significant impact of 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 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.*, 2019], 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 encoded 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 encoded 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

Model	Jukebox	MERT-330M	MERT-95M	Music2Vec	EnCodec	
Parameter Size	5B	330M	95M	95M	56M	
Data (h)	60 ~ 120k	160k	17k	1k	1k	
Average Score-P	0.5490( $\pm 0.1458$ )	0.5586( $\pm 0.1433$ )	<b>0.5640(<math>\pm 0.1425</math>)</b>	0.5474( $\pm 0.1417$ )	0.4583( $\pm 0.1377$ )	
Tempo & Rhythm	0.4610( $\pm 0.1016$ )	<b>0.4650(<math>\pm 0.1013</math>)</b>	0.4607( $\pm 0.0958$ )	0.4604( $\pm 0.1026$ )	0.4587( $\pm 0.1092$ )	
Emo. Impact (L & M)	0.5312( $\pm 0.0939$ )	0.5350( $\pm 0.0903$ )	<b>0.5396(<math>\pm 0.0857</math>)</b>	0.5311( $\pm 0.0924$ )	0.4860( $\pm 0.0920$ )	
Cult. & Reg.	0.5166( $\pm 0.2107$ )	0.5340( $\pm 0.2139$ )	<b>0.5390(<math>\pm 0.2110</math>)</b>	0.5120( $\pm 0.2094$ )	0.4072( $\pm 0.1261$ )	
Professional Description	Vocal Components	0.5464( $\pm 0.1953$ )	0.5550( $\pm 0.1957$ )	<b>0.5713(<math>\pm 0.1989</math>)</b>	0.5356( $\pm 0.1926$ )	0.4230( $\pm 0.1361$ )
Song Purp.	0.5810( $\pm 0.2191$ )	0.5864( $\pm 0.2166$ )	<b>0.6040(<math>\pm 0.2230</math>)</b>	0.5664( $\pm 0.2144$ )	0.4630( $\pm 0.1504$ )	
Mus. Genres	0.4600( $\pm 0.1239$ )	0.4644( $\pm 0.1172$ )	<b>0.4692(<math>\pm 0.1158</math>)</b>	0.4610( $\pm 0.1207$ )	0.4297( $\pm 0.1219$ )	
Exp. Impact (S & A)	0.9146( $\pm 0.0541$ )	0.9280( $\pm 0.0476$ )	<b>0.9310(<math>\pm 0.0447</math>)</b>	0.9190( $\pm 0.0576$ )	0.7085( $\pm 0.2888$ )	
Tgt. Aud.	0.4521( $\pm 0.1471$ )	0.4656( $\pm 0.1459$ )	<b>0.4683(<math>\pm 0.1417</math>)</b>	0.4565( $\pm 0.1514$ )	0.3623( $\pm 0.0980$ )	
Instrum.	0.5083( $\pm 0.1647$ )	<b>0.5180(<math>\pm 0.1587</math>)</b>	0.5156( $\pm 0.1592$ )	0.5063( $\pm 0.1727$ )	0.4043( $\pm 0.1426$ )	
Audio Eff.	0.5195( $\pm 0.1476$ )	0.5356( $\pm 0.1458$ )	<b>0.5425(<math>\pm 0.1483</math>)</b>	0.5244( $\pm 0.1539$ )	0.4404( $\pm 0.1122$ )	
Average Score-A	0.5894( $\pm 0.1353$ )	0.5900( $\pm 0.1284$ )	<b>0.5923(<math>\pm 0.1284</math>)</b>	0.5770( $\pm 0.1417$ )	0.4602( $\pm 0.1449$ )	
Amateur Description	Perc. of Tempo	<b>0.4600(<math>\pm 0.1521</math>)</b>	0.4540( $\pm 0.1475$ )	0.4580( $\pm 0.1456$ )	0.4463( $\pm 0.1407$ )	0.4065( $\pm 0.0994$ )
Emo. Impact (L)	0.5977( $\pm 0.1780$ )	0.5894( $\pm 0.1798$ )	<b>0.6006(<math>\pm 0.1780</math>)</b>	0.5806( $\pm 0.1827$ )	0.4430( $\pm 0.1320$ )	
Cult.& Reg.	0.4565( $\pm 0.1013$ )	0.4539( $\pm 0.0975$ )	<b>0.4575(<math>\pm 0.0949</math>)</b>	0.4510( $\pm 0.1023$ )	0.4324( $\pm 0.0972$ )	
Vocal Components	<b>0.5195(<math>\pm 0.1208</math>)</b>	0.5190( $\pm 0.1216$ )	0.5186( $\pm 0.1227$ )	0.5117( $\pm 0.1200$ )	0.4795( $\pm 0.0950$ )	
Song Purp.	0.5240( $\pm 0.2377$ )	0.5210( $\pm 0.2356$ )	<b>0.5410(<math>\pm 0.2422</math>)</b>	0.5201( $\pm 0.2428$ )	0.3801( $\pm 0.1532$ )	
Perc. of Uniq.	0.5356( $\pm 0.2076$ )	0.5356( $\pm 0.2115$ )	<b>0.5547(<math>\pm 0.2085</math>)</b>	0.5060( $\pm 0.1942$ )	0.4130( $\pm 0.1191$ )	
Exp. Impact (S)	0.9404( $\pm 0.0328$ )	0.9385( $\pm 0.0315$ )	<b>0.9460(<math>\pm 0.0315</math>)</b>	0.9297( $\pm 0.0477$ )	0.7144( $\pm 0.2640$ )	
Tgt. Aud.	0.4417( $\pm 0.1041$ )	0.4448( $\pm 0.1114$ )	<b>0.4530(<math>\pm 0.0951</math>)</b>	0.4353( $\pm 0.1220$ )	0.3933( $\pm 0.1075$ )	
Instrum.	0.7144( $\pm 0.0737$ )	<b>0.7153(<math>\pm 0.0537</math>)</b>	0.6787( $\pm 0.0333$ )	0.6807( $\pm 0.1059$ )	0.4219( $\pm 0.2092$ )	
Audio Eff.	0.7056( $\pm 0.1448$ )	<b>0.7275(<math>\pm 0.1465</math>)</b>	0.7144( $\pm 0.1326$ )	0.7110( $\pm 0.1586$ )	0.5176( $\pm 0.1725$ )	

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**.

API<sup>2</sup> (MIR toolkit). POP909 [Wang *et al.*, 2020] presents 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<sup>3</sup> is a music website where users upload audio with their annotations such as melodies, chords, and beats. MTG [Bogdanov *et al.*, 2019] and Mōusai [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.

**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. GTZAN [Sturm, 2013] presents a test set for music genre classification. PMEMo [Zhang *et al.*, 2018] has collected music emotional annotations and simultaneous electrodermal activity signals for 794 songs, thereby providing a benchmark for music emotion recognition. 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.

<sup>2</sup><https://developer.spotify.com/>

<sup>3</sup><https://www.hooktheory.com/>

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 Cailion, 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. *ISMIR*, 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.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019*, volume 1, page 4171 – 4186, 2019.
- [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. *ISMIR*, 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, Zhiping Hu, Haoyu He, et al. Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. *ISMIR*, 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, Wenhui 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.
- [Mancio *et al.*, 2021] Ilaria Mancio, 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 Hermans, 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 Metzner, 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. *ISMIR*, 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. Song-driver: 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. *ISMIR*, 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. *(Advances in Neural Information Processing Systems*, 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.*, 2018] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. The pmemo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pages 135–142, 2018.
- [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.

## A Tasks with Automatic Annotation

The performance of current algorithms designed for annotating textual description, lyrics, and musical section annotation is not satisfactory due to their reliance on subjective human evaluation. Except these, other kinds of data, such as phonetic alignment, vocal separation, and audio-to-MIDI conversion, do not significantly align with human perception. Annotating these elements manually is particularly challenging, requiring extensive effort and time. However, there are now numerous advanced algorithms that efficiently handle these tasks, as detailed in the **Appendix B**. As a result, we utilize data preprocessing algorithms for the automatic annotation of this content, eliminating the need for manual annotation or intervention, and seamlessly integrate this processed content into our dataset.

Classification	Task
A	Musical Section Annotation
	Lyric Correction
	Lyric Screening
	Rhyme Annotation
B	Professional Music Description
	Amateur Music Description

Table 3: Classification of Annotation Tasks.

## B Data Preprocessing

- **Music Genre Clustering** To mitigate subjective bias and ensure diverse descriptions across various music genres, it’s crucial to distribute a broad spectrum of music genres among annotators, thereby enriching the annotation’s diversity. To facilitate this, we utilize MERT [Li *et al.*, 2023], a pre-trained music audio encoder, to process the audio data. Following this, we cluster the encoded data, resulting in 1000 unique audio clusters. From these clusters, we evenly distribute music data, guaranteeing that annotators are presented with a balanced mix of music for labeling. This approach ensures that each music cluster is described by a range of annotators, significantly enhancing the diversity and richness of the annotated data.
- **Vocal & Track Separation** To make the dataset suitable for tasks such as accompaniment generation, melody generation, and vocal synthesis, we apply Demucs [Rouard *et al.*, 2023; Défossez, 2021] to perform vocal separation, separating the vocals from the musical accompaniment in the audio files. Furthermore, considering the requirements of a wider range of music-related tasks, we also separate individual instrument tracks, such as drums and bass.
- **Phonemic Level Alignment in Audio-Lyrics** To prepare audio-lyrics pairs for applications such as vocal synthesis, it’s necessary to align them at the phonemic level. We employ the Montreal Forced Aligner (MFA) [McAuliffe *et al.*, 2017] for this task, initially achieving a 67% accuracy rate. While the MFA demonstrates a commendable 95% accuracy for aligning monophonic phonemes to single characters, its performance

drops due to inaccuracies in marking the offsets for melismatic phonemes. These phonemes are characterized by multiple pitches sung within a single syllable or note, complicating the alignment process and diminishing the overall accuracy. To address this, we optimized the MFA algorithm with a focus on accurately identifying and aligning melismatic phonemes. Furthermore, we implemented features to recognize and annotate significant pauses and breaths during singing. These enhancements significantly improve our final alignment accuracy to 97%.

- **Automatic Pre-annotation** To improve the efficiency of future manual annotations, we implemented specific software for automatic pre-annotation of certain tasks related to lyric annotation. For annotating rhyme schemes in lyrics, we use a specialized program that pre-annotates the rhyme scheme for each line. For theme annotation in lyrics, we employ a fine-tuned version of Qwen to preliminarily identify the main theme of the lyrics for each piece of music. During the formal annotation phase, these pre-annotations serve as a basis for manual review. Annotators can assess the accuracy of these automatic annotations and adjust them as necessary, or use them as a guideline for their own annotation efforts.

- **Lead Sheet Transcription** To facilitate symbolic music-related tasks using MIDI, we transcribe the audio in the MuChin into lead sheets. These sheets, which are a simplified form of MIDI notation, are created using Sheet Sage [Donahue and Liang, 2021], software that utilizes the encoding model of Jukebox [Dhariwal *et al.*, 2020]. This conversion facilitates the application of MuChin to a wide range of tasks associated with symbolic music.

Classification	Accuracy(%)
I	[90, 100]
II	[70, 90)
III	[60, 70)
IV	[0, 60)

Table 4: Classification of Annotators of Type A Tasks Based on Accuracy

Classification	Score
I	[90, 100]
II	[70, 90)
III	[60, 70)
IV	[0, 60)

Table 5: Classification of Annotators of Type B Tasks Based on Score

## C Quality Assurance Mechanisms

In this section, we will provide a detailed introduction to the quality assurance mechanism, including the classification of tasks, scoring guidelines and the classification of individuals.

Dimension	Score	Standard
Expressive Impact (S. & A.)	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Emotional Impact	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Textual Description	8	3 for Description Relevance; 5 for Word Counts and Innovation
Musical Genres	8	8 for Level of Detail
Tempo and Rhythm	5	5 for Label Relevance
Instrumentation	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Song Purpose	6	3 for Label Relevance; 3 for Innovation
Culture and Region	6	3 for Label Relevance; 3 for Innovation
Target Audience	6	3 for Label Relevance; 3 for Innovation
Vocal Components	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Audio Effects	5	5 for Label Relevance
Lyric Themes	6	3 for Label Relevance; 3 for Innovation
Total	100	-

Table 6: Scoring Guidelines of Professional Music Description

Dimension	Score	Standard
Perception of Uniqueness	8	4 for Label Relevance; 4 for Innovation
Perception of Tempo	5	3 for Label Relevance; 2 for Innovation
Expressive Impact (S.)	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Emotional Impact (L.)	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Textual Description	8	3 for Description Relevance; 5 for Word Counts and Innovation
Instrumentation	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Song Purpose	6	3 for Label Relevance; 3 for Innovation
Culture and Region	6	3 for Label Relevance; 3 for Innovation
Target Audience	6	3 for Label Relevance; 3 for Innovation
Vocal Components	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Audio Effects	5	5 for Label Relevance
Lyric Themes	6	3 for Label Relevance; 3 for Innovation
Total	100	-

Table 7: Scoring Guidelines of Amateur Music Description

### C.1 Classification of Annotation Tasks

We classify annotation tasks into two categories based on their potential for objective evaluation: **Type A**, which can be objectively assessed, and **Type B**, which are subject to subjective assessment. This section exemplifies the classification of each annotation task. To maximize the accuracy and comprehensiveness of each song’s annotations, we allocate two annotators to Type A tasks and one annotator to Type B tasks for each song. These tasks are carried out separately, not simultaneously. Additionally, apart from annotators, several quality assurance inspectors are needed to evaluate the annotators’ outputs. According to the division into Type A and B, we consolidate Type A tasks into one phase, denoted as the **Structure Annotation Phase**, and Type B tasks into the subsequent phase, denoted as the **Music Description Annotation Phase**. Data must sequentially pass through these two phases before inclusion in the dataset. That is, data must undergo structure annotation and pass quality assurance before proceeding to the music description annotation phase, after which, data that passes quality assurance following music description annotation can be added to the dataset. For Type A tasks, if both annotators provide identical annotations, we consider the annotation accurate. However, when there is a discrepancy, quality assurance inspectors must deliver their judgment to determine which result is correct, or if both are incorrect, provide their own accurate annotation. For Type B tasks, quality assurance inspectors are required to assign a

score ranging from 0 to 100 to the annotation results, with the scoring guidelines detailed in Table 6 and 7.

### C.2 Classification of Individuals

To ensure diligent performance from annotators, we have implemented a screening mechanism. During the structural annotation phase, the precision of Type A task annotations is assessed through the previously mentioned quality assurance system. In the music description annotation phase, given that Type B tasks involve subjective descriptions challenging to assess objectively, we randomly review 20% of the annotations from each annotator for quality control. Moreover, we evaluate behaviors indicated by backend analytics, such as interaction frequency with the progress bar and task skipping. Annotators showing superficial engagement will be warned. In both phases, annotators are categorized into four groups based on their weekly accuracy rates or average scores, as detailed in Table 4 and 5. Type IV annotators, and those receiving two or more warnings, will be excluded from future tasks, and their data for the current week will be disregarded. Type I annotators will be rewarded, while Type III annotators may incur penalties.

### C.3 Other Quality Assurance Measures

Annotators are responsible for screening the data (Type A & B). For songs that contain languages other than Chinese, have poor audio quality, or involve pornography or violence,

therefore unsuitable for inclusion in the dataset, annotators can mark these for exclusion and skip their annotation.

When annotating musical sections of Type A, annotators must repeatedly listen to a music piece. Consequently, the dedication to their annotation tasks is assessed by the amount of time they spend on the annotation page, their frequency of interactions with the progress bar, and the frequency of their play/pause button clicks.

In the textual description annotation (Type B), to ensure that annotators listen to each song attentively and provide thoughtful music descriptions, we stipulate that annotators must listen to the entire song in one sitting before adjusting the progress bar and playback speed. They must compose a textual description of no fewer than 50 words, and are prohibited from writing the description within the first 30 seconds of the song's playback, as well as from copying and pasting any content.



Figure 4: Supplementary actual screenshots from the main text. A screenshot of the 'Song Purpose' section during the Description Annotation Phase.



Figure 5: Supplementary actual screenshots from the main text. A screenshot of the 'Song Purpose' section during the Description Quality Assurance Phase.



Figure 6: Supplementary actual screenshots from the main text. A screenshot of the 'Instrumentation' section during the Description Annotation Phase.

## D CaiMAP: Caichong Multitask Music Annotation Platform

In Appendix C, we have launched a comprehensive suite of annotation tasks alongside an advanced quality assurance system. To bring these complex designs to life, we developed the Caichong Multitask Music Annotation Platform (CaiMAP), which harmonizes this series of tasks and systems. This section will provide a brief overview of the platform.



Figure 7: Supplementary actual screenshots from the main text. A screenshot of the 'Instrumentation' section during the Description Quality Assurance Phase.



Figure 8: Supplementary actual screenshots from the main text. A screenshot of the 'Audio Effects' section during the Description Annotation Phase.

- Account and Login.** The platform utilizes an access control system, assigning specific roles to each user account. Users can log into their accounts, review and complete assigned tasks, and submit their results.
- Annotation Interface.** Upon logging in and selecting a specific piece of music, annotators are directed to a dedicated annotation interface designed for the task. This interface includes a media player and a specialized text box. Users have the ability to control the progress bar and playback speed of the media player. Furthermore, the music description annotation interface incorporates a comprehensive lexicon and search tool, enabling users to select suitable descriptive terms directly from the lexicon or to search for specific terms as needed.
- Quality Assurance Interface.** Upon logging in and selecting a specific piece of music, quality assurance inspectors are taken to the quality assurance interface. For Type A tasks, inspectors are responsible for simultaneously evaluating the annotations provided by two users. The interface presents these annotations side-by-side, highlighting the differences for easy comparison. Inspectors can then decide which annotation is correct, make adjustments to either, or choose to re-annotate the piece. For Type B tasks, the interface displays a single, complete annotation for the inspector to verify and score. Inspectors simply review the annotation and submit their scores.
- Administrator Interface.** Administrators have the access to view the submissions of any designated user, including annotators and quality assurance inspectors. Both the annotation and quality assurance interfaces in-

corporate a feedback button for reporting platform issues, enabling annotators and quality assurance inspectors to communicate with administrators for resolution.

We have provided screenshots of several platform pages as examples, as shown in Figures 4–8.

## E Individual Grouping and Training

### E.1 Grouping

During the structure annotation phase, which consists of Type A tasks, each piece of data requires two annotations. In contrast, the music description annotation phase, made up of Type B tasks, necessitates only one annotation. As a result, the latter phase involves fewer participants. The task of annotating the musical sections in the lyric annotation phase demands a basic knowledge of music theory. Consequently, only 104 professionals are engaged in this task. Out of these, 11 individuals, distinguished by their high level of expertise and conscientious approach, are chosen as quality assurance inspectors. This selection process involves screening their resumes and conducting further assessments. The remaining 93 individuals function as annotators.

During the music description annotation phase, the 109 amateurs form the amateur group, and the 93 professionals from the previous phase form the professional group. Additionally, the 11 inspectors from the previous phase continue to serve as inspectors in this phase. Beyond the roles of annotators and quality assurance inspectors, we also select a member from our research team who is adept at using the platform, with a high level of expertise, and with strong communication skills to act as the platform administrator.

### E.2 Training

Next, we offer training for both the annotators and quality assurance inspectors, focusing on their specific roles. Initially, each annotator accesses CaiMAP to pre-annotate a compact dataset of around 20 entries, which encompasses tasks of both Type A and B. This phase allows annotators to acquaint themselves with the platform's features and learn the correct procedures for completing annotation tasks. Additionally, we provide specialized training to address common mistakes, such as the elimination of extraneous information from lyric texts and the accurate identification of each interjection.

On the other hand, training for inspectors entails a more intricate process. They must not only master the platform's use but also develop a set of consistent evaluation standards. We gather data annotated by the annotators during the pre-annotation phase and distribute the same dataset to all inspectors. For the lyric annotation phase, inspectors must choose the annotation they consider correct based on the guidelines outlined in **Section 2.2**, or provide an alternative correct annotation if they find the existing ones inaccurate. During the music description annotation phase, inspectors evaluate each annotation independently. Once the inspectors have completed their tasks, we compile all the scores for the music descriptions and organize a meeting with the inspectors. At this meeting, we identify instances where scores from different inspectors significantly vary, with a maximum discrepancy exceeding 10 points, and encourage inspectors to discuss and

agree on a unified evaluation criterion. This training process is repeated until the inspectors' scores for the same dataset show substantial consistency.

```
(verse1)
晚上来临了ccccR
游戏通关了ccccR
我们的爱情早已结束了ccccccccR
(verse2)
我的心碎了ccccR
你也解脱了ccccR
你就像只船开走了ccccccccR
(chorus1)
船要起航了ccccR
你要出发了ccccR
到处漂泊的你也许会累了ccccR
```

Figure 9: A Fragment from an Illustrative Example of Structure Annotation

<i>Main Question:</i> 首歌带给你的感受? <i>Main Question:</i> How does this song make you feel?	<i>Label Selection</i> Q1: 特色感受 <b>Q1: Perception of Uniqueness</b> A1: "情歌", "青春", "积极面对", "动听" <b>A1: "Love song," "Youth," "Face positively," "Melodious."</b> Q2: 快慢感受 <b>Q2: Perception of Tempo</b> A2: "欢快", "踩点", "跟着哼唱" <b>A2: "Cheerful," "On beat," "Hum along."</b> Q3: 表现力感受 (歌手) <b>Q3: Expressive Impact (Singer)</b> A3: "感悟", "情深意切", "余音袅袅", "动情" <b>A3: "Insight," "Deep emotion," "Lingering sound," "Moving."</b> Q4: 情绪感受 (歌词) <b>Q4: Emotional Impact (Lyrics)</b> A4: "成长", "追忆", "愉悦", "释然" <b>A4: "Growth," "Reminiscence," "Joy," "Relief."</b>
<i>Compose Description</i> A: 这是一首正能量的歌曲，在成长中难免会遇到困难挫折，克服它们继续向前，向着人生的目标奔跑，要无所畏惧。 <b>A: This is a positive song that acknowledges the inevitable difficulties and setbacks encountered during growth. It encourages overcoming these obstacles and continuing to move forward fearlessly towards the goals of life.</b>	

Figure 10: A Fragment from an Illustrative Example of Amateur Description Annotation

## F Caichong Music Dataset

### F.1 Annotated Data Processing

On one hand, we seamlessly incorporate annotations of musical sections into the lyrics by marking the start of each musical section with a section label, positioned before the lyrics

*Main Question:* 首歌带给你的感受?  
*Main Question:* How does this song make you feel?

## *Label Selection*

Q1: 表现力感受（歌手和伴奏）

### **Q1: Expressive Impact (Singer & Accompaniment)**

A1: "美妙的声音","温暖磁性的男声","完美","柔和的男声"

A1: "Exquisite voice," "Warm and magnetic male voice," "Perfection," "Gentle male voice."

|Q2: 情绪感受（歌词和旋律）

#### Q2: Emotional Impact (Lyrics)

A2: "感性的","放松和安慰","浪漫情感"

A2: "Sentimental." "Relaxing and comfort

### *Compose Description*

A：这首歌整体表现了较为感性的情感主题，描述了两人的陪伴和旅行方面的故事。歌手的表现力很棒，情感充沛，技法高超，很快能把人带入到一个舒缓平静的氛围中。

A: The song overall conveys a sentimentally charged emotional theme, depicting the story of companionship and travel between two individuals. The singer's performance is impressive, brimming with emotion and technical proficiency, swiftly drawing the listener into a soothing and tranquil atmosphere.

Figure 11: A Fragment from an Illustrative Example of Professional Description Annotation

of that section begin. We denote rhyming information using strings that include ‘c’ and ‘R’ markers: an ‘R’ is added at the end of any sentence that rhymes with the one before it, while ‘c’ indicates words that do not rhyme. This method is used to compile all annotated lyric information—encompassing the lyrics’ theme, musical sections, and rhyming details—into a JSON file.

On the other hand, during the phase dedicated to annotating music descriptions, we collect textual descriptions of each music piece from various perspectives. Each annotation consists of several descriptive terms along with a comprehensive descriptive text. To enhance the richness of these descriptions, we integrate these terms into the textual descriptions, which are then combined with the texts. Furthermore, we concatenate descriptions from different aspects to create a single, detailed annotation that captures the multifaceted nature of the music.

## F.2 Overview

This section provides an overview of the descriptive tag distribution and song structure distribution in CaiMD, as illustrated in Figure 12–14. Song structure is the arrangement of musical sections.

### F.3 Examples

This section presents a range of annotation examples, encompassing both professional and colloquial musical descriptions, along with the musical sections and rhymes featured in CaiMD, as depicted in Figures 9–11.

## G Evaluation Metrics of Structured Lyric Generation

## G.1 Formula

The similarity of the overall structure and musical section structure is calculated according to Equation 1, where  $K_m$  represents the number of matching characters in the longest

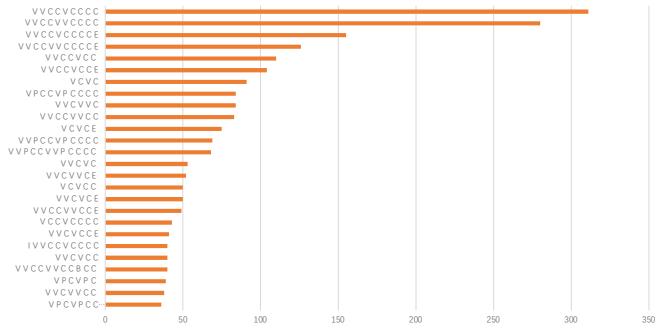


Figure 12: Distribution of Song Structures. The bin labels on the left side of the histogram represent the various musical sections of a song. Specifically, 'i' stands for "Introduction," 'v' corresponds to "Verse," 'c' denotes "Chorus," 'p' indicates "Pre-chorus," 'b' signifies "Bridge," and 'e' represents the "Ending."

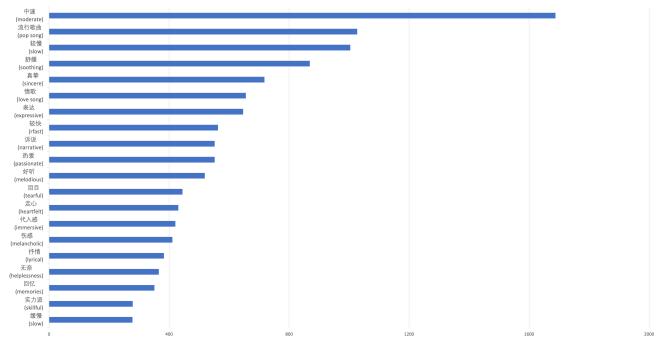


Figure 13: Distribution of Colloquial Descriptive Tags

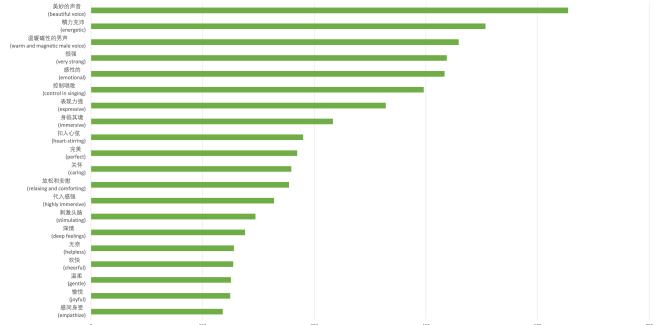


Figure 14: Distribution of Professional Descriptive Tags

common subsequence between strings  $A$  and  $B$ .  $L_A$  denotes the length of string  $A$ , and  $L_B$  denotes the length of string  $B$ . In the context of the overall structure,  $A$  and  $B$  represent the entire set of lyrics. In the context of musical section structure,  $A$  and  $B$  refer to the sequence of musical section labels.

$$p = \frac{2K_m}{L_A + L_B} \quad (1)$$

The within-section structure similarity is calculated according to Equation 2. In this equation, each element of *ListA* and *ListB* represents the number of sentences contained in each matching musical section of song *A* and *B*, respectively, e.g., [4, 8, 4] indicates that the three matching musical sections contain 4, 8, and 4 sentences, respectively.

---

**Algorithm 1** Reward Score Algorithm

---

**Input:** max\_equ\_slc\_sum, rc\_ing, acmp\_sr, rc\_ino  
**Parameter:** EXTRA\_POINTS  
**Output:** extscore

```

1: if max_equ_slc_sum == 0 then
2:   r_ratio = 0
3: else
4:   r_ratio = rc_ing / max_equ_slc_sum
5: end if
6: extscore = EXTRA_POINTS * acmp_sr
7: if 0.6 <= r_ratio and r_ratio <= 0.8 then
8:   extscore *= 1.0
9: else if rc_ino == rc_ing and rc_ino > 0 then
10:  extscore *= 0.7
11: else
12:   r_delta = |r_ratio - 0.7|
13:   if r_delta <= 0.3 then
14:     extscore *= 0.4 * (1 - r_delta)
15:   else
16:     extscore *= 0.0
17:   end if
18: end if
19: return extscore

```

---

$$p = \frac{2 \sum \min(ListA, ListB)}{\sum ListA + \sum ListB} \quad (2)$$

Similarly, the within-sentence structure similarity can also be calculated using Equation 2. In this calculation, each element of  $ListA$  and  $ListB$  represents the number of words in each matching sentence of songs  $A$  and  $B$ .

The calculation of rhyming similarity follows Equation 1, where  $K_m$  represents the number of sentences that contain rhyming markers in the lyrics, and  $L_A$  and  $L_B$  respectively represent the total number of sentences in songs  $A$  and  $B$ .

Since each more detailed structure depends on the match of the preceding structure, cumulative similarity is used when calculating similarity, to take into account the influence of more macroscopic structures on the similarity of more microscopic structures. With the similarities of the overall structure, musical section structure, within-section structure, within-sentence structure, and rhyming structure calculated as  $p_1$  to  $p_5$  respectively, and their corresponding weights in the overall scoring as  $w_1$  to  $w_5$ , the overall similarity can be calculated using Equation 3.

$$p = \sum_{i=1}^5 w_i \prod_{j=1}^i p_j \quad (3)$$

Multiplying the overall similarity by 100 gives the overall score. Additionally, the extra reward score based on the proportion of rhyming sentences within the overall lyrics is also incorporated into the overall score.

## G.2 Reward Score

The calculation method of the reward score is shown as Algorithm 1, by which generated lyrics are assigned a certain

amount of reward points based on the proportion of rhyming. In this algorithm,  $\text{max\_equ\_slc\_sum}$  denotes the maximum number of phrases that match;  $\text{rc\_ing}$  denotes the number of rhyming phrases that match;  $\text{acmp\_sr}$  denotes the cumulative product of similarities across the first 5 dimensions;  $\text{rc\_ino}$  denotes the proportion of rhyming within the given rhyme scheme. And  $\text{EXTRA\_POINTS}$  denotes the total score of the reward score.

## H Details of Evaluating Music Understanding Models

### H.1 Pipeline of MLP

To assess the effectiveness of music understanding models, we feed music audio into them and obtain their respective encoded sequences. Subsequently, for each model, we utilize an MLP comprising an average pooling layer and 5 linear layers to extract 10 sets of descriptive music tags corresponding to the dimensions of its output encoded sequences. The pipeline of this process can be found in Figure 15.

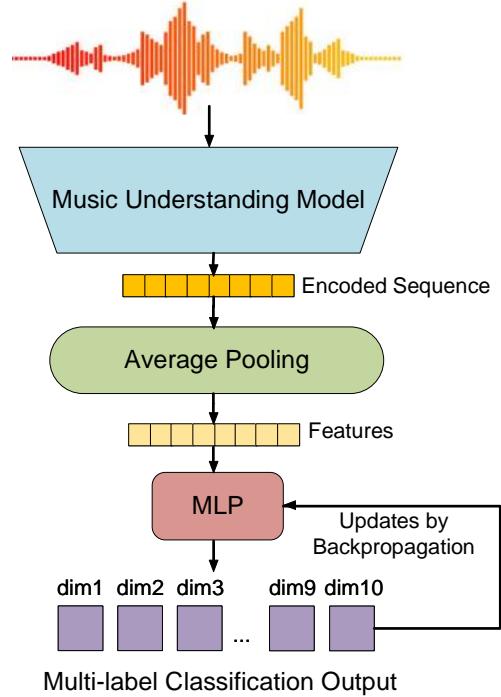


Figure 15: The pipeline of evaluating music understanding models

## H.2 Result Analysis

Figure 16 shows, despite having fewer parameters and a smaller amount of training data, MERT-95M performs best overall in the task of professional and colloquial music description.

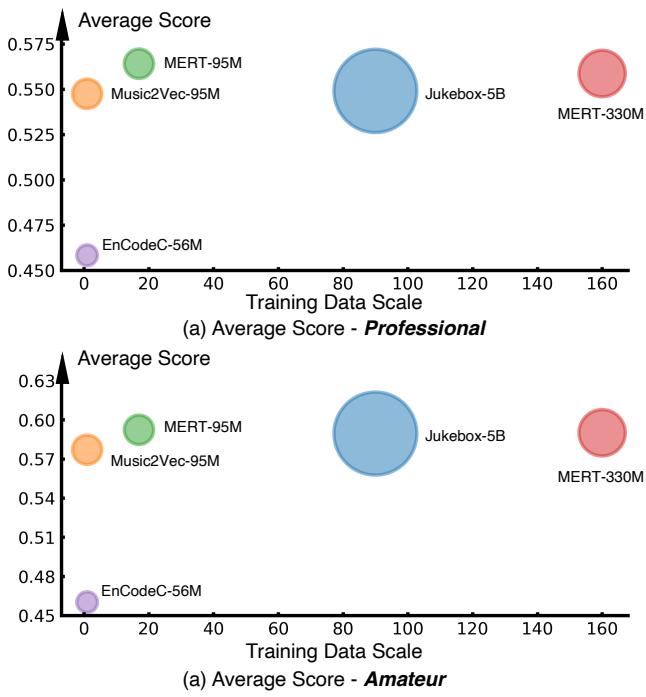


Figure 16: Evaluation of selected music understanding models on the benchmark as represented in a scatter plot.