# A Tasks with Automatic Annotation

The effectiveness of predetermined algorithms for textual description, lyrics, and musical section annotation is unsatisfactory, as they rely on subjective human judgment. However, other types of information, such as phonetic alignment, vocal separation, and audio-to-MIDI transcription, bear no significant correlation to human perception. Moreover, annotating these aspects is challenging for humans, necessitating substantial effort and time. Currently, there is an abundance of mature algorithms capable of accomplishing these tasks, which will be discussed in **Appendix B**. Consequently, we employ the data preprocessing algorithms for the automatic annotation of this content, foregoing manual annotation or intervention, and integrate it directly 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 prevent subjective bias and a lack of diverse descriptions for certain music genres, it is essential to allocate a wide range of music genres to each annotator, ensuring the diversity of the annotations. To achieve this, we employ MERT [Li *et al.*, 2023], a pre-trained music audio encoder, to encode the audio data. Subsequently, we perform clustering on the encoded data, resulting in 1000 distinct audio clusters. We then extract music data from these clusters evenly, ensuring that annotators receive a balanced selection of music for labeling. By implementing this method, we ensure that each music cluster receives a sufficient number of descriptions from different annotators, thereby enhancing the diversity 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** For audio-lyrics pairs, we need to perform phonemic level alignment to make them suitable for tasks like vocal synthesis. We utilize the Montreal Forced Aligner (MFA) [McAuliffe *et al.*, 2017] to align the audio-lyrics pairs, achieving an accuracy of 67%. While MFA achieves a high accuracy of 95% for aligning single-pitch phonemes with single characters, it tends to incorrectly mark the offsets for melismatic phonemes, which involve singing multiple pitches within a single syllable or note, resulting in lower overall accuracy. To address this, we optimize the MFA algorithm by prioritizing the identification and alignment of melismatic phonemes. Additionally, we recognize and mark long pauses and breaths that occur during singing. With these improvements, we achieve a final alignment accuracy of 97%.
- **Automatic Pre-annotation** In order to enhance the efficiency of subsequent manual annotation, we employ predetermined programs for automatic pre-annotation on certain lyric annotation tasks. For lyric rhyme annotation, a identification program is utilized to pre-annotate the rhyme scheme of each line, while in lyric theme annotation, the fine-tuned Qwen is utilized to extract the main theme of lyrics for each music piece in advance. During the formal annotation phase, all pre-annotation results serve as references for manual annotation. Annotators can verify the accuracy of the pre-annotation and make modifications based on it, or use the pre-annotation results as references for their own annotation tasks.
- **Lead Sheet Transcription** To facilitate symbolic music-related tasks using MIDI, we transcribe the audio in the MuChin into lead sheets, a simple form of MIDI notation, using Sheet Sage [Donahue and Liang, 2021], which is based on the encoding of Jukebox [Dhariwal *et al.*, 2020]. This enables its application to tasks related to 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 classfication of individuals.

## C.1 Classification of Annotation Tasks

Based on whether the annotation tasks can be objectively assessed, we categorize them into **Type A (yes)** and **Type B**

| 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

**(no)**. 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 that annotators perform their tasks diligently, we design a screening mechanism for the annotators. During the structure annotation phase, the accuracy of annotations for Type A tasks is evaluated using the quality assurance mechanism mentioned previously. During the music description annotation phase, since Type B tasks are subjective descriptions and difficult to objectively judge as correct or incorrect, we randomly select 20% of the data annotated by each annotator for quality assurance scoring. Additionally, behaviors detected by the backend, such as the frequency of the interactions with the progress bar or skipping through tasks, are also evaluated. Annotators deemed to be perfunctory will receive warnings. In both phases, annotators are classified into four types based on their weekly accuracy rate or average score, as shown in Table 4 and 5. Type IV annotators, as well as those who have accumulated two or more warnings, will no longer participate in the following annotation tasks, and their data for the current week will be invalidated. Type I annotators will receive rewards, while Type III annotators will face certain 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 (Type A), annotators are required to listen to a piece of music repeatedly. Therefore, we judge the seriousness of their annotation efforts based on the duration of time they spend on the annotation page, the frequency of their interactions with the progress bar, and how often they click the play/pause button.

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.

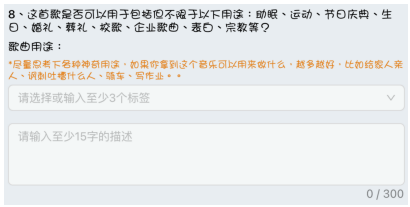


Figure 4: Supplementary actual screenshots from the main text. A screenshot of the 'Song Purpose' section during the Description Annotation Phase.

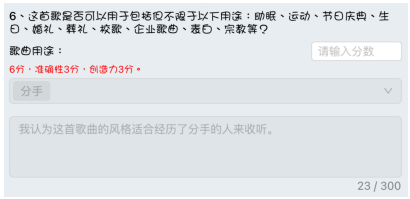


Figure 5: Supplementary actual screenshots from the main text. A screenshot of the 'Song Purpose' section during the Description Quality Assurance Phase.


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 introduced an array of annotation tasks and a sophisticated quality assurance mechanism. To actualize these intricate designs, we developed the Caichong Multitask Music Annotation Platform (CaiMAP), to integrate the series of tasks and mechanisms. The platform will be introduced in brief in this section.

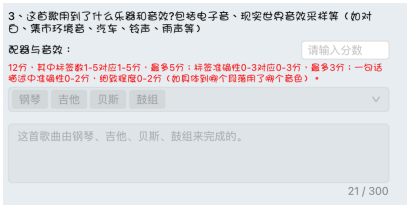


Figure 7: Supplementary actual screenshots from the main text. A screenshot of the 'Instrumentation' section during the Description Quality Assurance Phase.

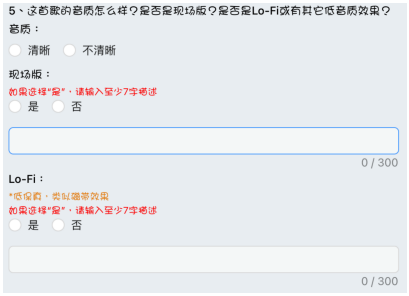


Figure 8: Supplementary actual screenshots from the main text. A screenshot of the 'Audio Effects' section during the Description Annotation Phase.

- **Account and Login.** The platform employs a username and password system for access, with accounts distributed to users based on the specific nature of their tasks. Each account is tethered to a distinct task, necessitating users to utilize the allocated account to accept, execute, and submit the pertinent task.
- **Annotation Interface.** Annotators, upon logging in and selecting a specific music piece, are directed to its dedicated annotation interface. This interface is equipped with a media player and a text box tailored to the task. Users can adjust the player's progress bar and the speed of playback. And the music description annotation page features an integrated lexicon and search utility, allowing users to choose appropriate descriptive terms from the lexicon or to search for the desired terms.
- **Quality Assurance Interface.** Quality assurance inspectors, upon logging in and selecting a specific music piece, are directed to its dedicated quality assurance interface. This interface adapts its presentation to the specific quality assurance task. For Type A tasks, inspectors are tasked with evaluating the annotations of two users concurrently. In such cases, the interface displays the annotations side by side, delineating the differences for review. Inspectors may determine one of the annotations as correct, make modifications to either, or opt to re-annotate correctly themselves by selecting the re-annotate option. For Type B tasks, the quality assurance interface exhibits a single complete annotation for the inspector's assurance and scoring. The inspectors simply review the annotations and submit their scores.
- **Administrator Interface.** Administrators have the ac-

cess to viewing the submissions of any designated user, including annotators and quality assurance inspectors. Both the annotation and quality assurance interfaces incorporate a feedback button for reporting platform issues, enabling annotators and quality assurance inspectors to communicate with administrators for issue 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

Given that each data require double annotations during the structure annotation phase, composed of Type A tasks, and only a single annotation during the music description annotation phase, composed of Type B tasks, fewer participants are involved in the music description annotation phase. The musical section annotation task of the lyric annotation phase necessitates a basic understanding of music theory; therefore, only the 104 professionals participate. From these, 11 individuals with a high level of expertise and a conscientious attitude are selected as quality assurance inspectors through resume screening and subsequent assessments, while the remaining 93 serve 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. Specific task assignments for each group are as follows.

## E.2 Training

Next, we provide training for the annotators and quality assurance inspectors on their respective tasks. On the one hand, each annotator logs in to Cai and pre-annotate a small clustered dataset of approximately 20 entries, including the tasks of Type A and B. During this period, the annotators should familiarize themselves with the platform's functionalities and understand how to properly execute the annotation tasks. We also provide targeted training for the annotators on common errors, such as removing irrelevant information from the lyric texts and clearly marking each interjection.

On the other hand, training for the inspectors is somewhat more complex: not only do they need to become proficient in using the platform, but they must also establish a set of unified evaluation criteria. We collect data annotated by the annotators during the pre-annotation phase and distribute this identical dataset among all the inspectors. For the lyric annotation phase, inspectors are required to select the annotation they deem correct based on the mechanisms mentioned in **Section 2.2** or provide their own correct annotation if they believe neither of the existing options is accurate. During the music description annotation phase, inspectors independently score each annotation. After the completion of the inspectors' tasks, we collect all the scores for the music descriptions and convene a meeting of the inspectors. We identify data where different inspectors' scores significantly diverge with a maximum difference of more than 10 points, and ask the inspectors to discuss and establish a unified evaluation standard. This training is repeated until the inspectors' scores for the same dataset are roughly consistent.

(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
cccccccR
(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

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

*Label Selection*
Q1: 特色感受
Q1: Perception of Uniqueness
A1: "情歌","青春","积极面对","动听"
A1: "Love song," "Youth," "Face positively," "Melodious."
Q2: 快慢感受
Q2: Perception of Tempo
A2: "欢快","踩点","跟着哼唱"
A2: "Cheerful," "On beat," "Hum along."
Q3: 表现力感受（歌手）
Q3: Expressive Impact (Singer)
A3: "感悟","情深意切","余音袅袅","动情"
A3: "Insight," "Deep emotion," "Lingering sound," "Moving."
Q4: 情绪感受（歌词）
Q4: Emotional Impact (Lyrics)
A4: "成长","追忆","愉悦","释然"
A4: "Growth," "Reminiscence," "Joy," "Relief."

*Compose Description*
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。
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.

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

*Main Question:* 这首歌带给你的感受?
*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 & Melody)
A2: "感性的","放松和安慰","浪漫情感"
A2: "Sentimental," "Relaxing and comforting," "Romantic emotions."

*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

# F Caichong Music Dataset

## F.1 Annotated Data Processing

On the one hand, we integrate the annotated information of music sections directly into the lyrics by marking the beginning of each music section with a section label placed before the start of the lyrics for that section. Rhyming information is indicated with strings containing 'c' and 'R' markers: at the end of each sentence that rhymes with the preceding one, we mark an 'R', while non-rhyming parts are marked with a 'c'. All the annotated lyric information, including the theme of the lyrics, music sections, and rhyming information, is consolidated in a JSON file using the aforementioned method.

On the other hand, during the music description annotation phase, we get textual descriptions of each music piece across multiple dimensions, with each annotation including both several descriptive terms and a complete descriptive text. To better enhance the completeness of the descriptions, we concatenate the terms into textual descriptions and merged them into the texts. Subsequently, descriptions from different dimensions are also concatenated, forming a unified and comprehensive annotation that describes the various dimensions 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.

## 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.

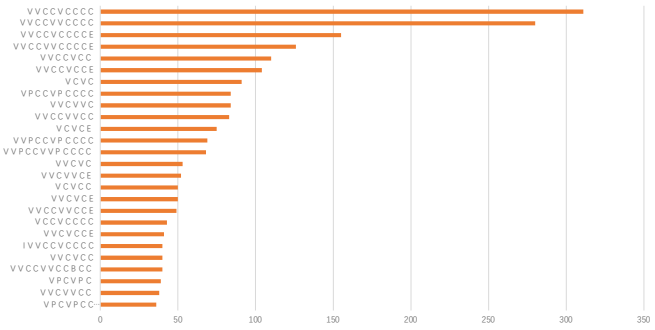


Figure 12: Distribution of Song Structures. The bin labels on the left side of the histogram represent the various musical sections of a song. Specifically, 'i' stands for "Introduction," 'v' corresponds to "Verse," 'c' denotes "Chorus," 'p' indicates "Pre-chorus," 'b' signifies "Bridge," and 'e' represents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

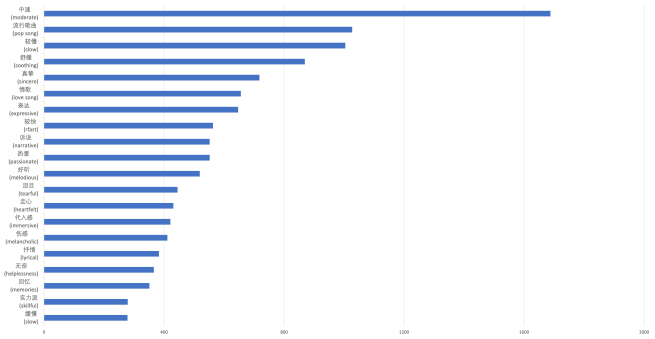

Figure 13: Distribution of Colloquial Descriptive Tags

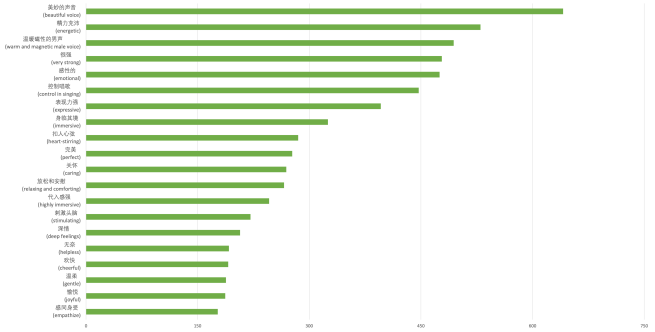

Figure 14: Distribution of Professional Descriptive Tags

# 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 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.

$$p = \frac{2K_m}{L_A + L_B} \tag{1}$$

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.

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

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.

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

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

### G.2 Award 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.

### 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.

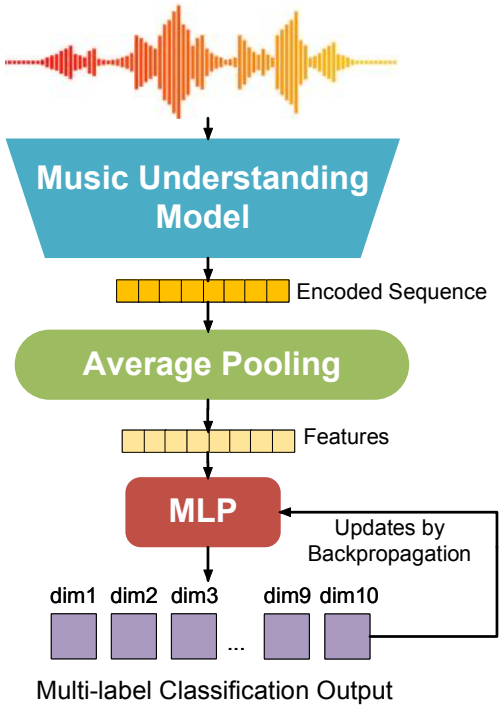


Figure 15: The pipeline of evaluating music understanding models

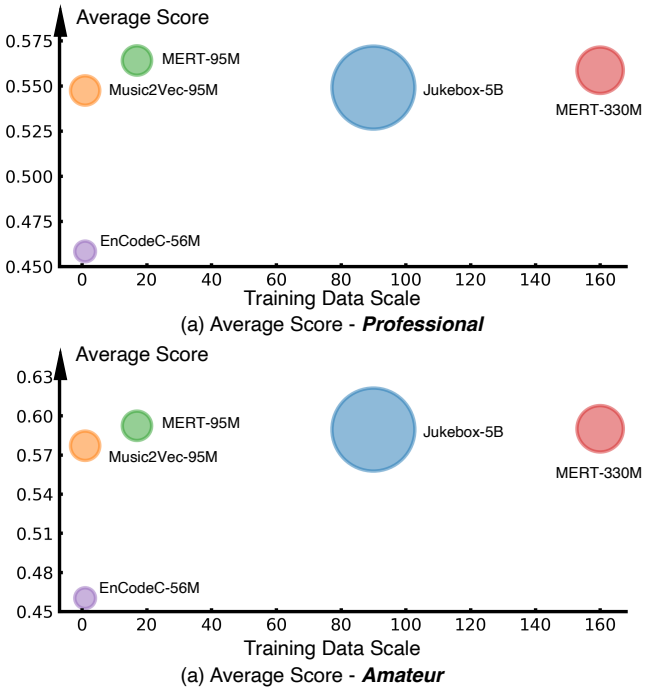


Figure 16: Evaluation of selected music understanding models on the benchmark as represented in a scatter plot.