

A Survey of Quality Evaluation on AI-Generated Lyrics

Weining Sun*

School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing, 100192

**Corresponding author: Weining Sun.*

Abstract

With the rapid development of generative artificial intelligence, AI lyric generation tools such as Suno, Baidu Wenxin Lyric Generator, and Doubao have been widely deployed. However, AI-generated lyrics commonly suffer from thematic deviation, emotional fragmentation, stylistic mismatch, grammatical inconsistency, and long-structure logical collapse, making objective quality evaluation a critical link for optimizing generation models and filtering outputs. This paper systematically reviews existing research on AI-generated lyric quality evaluation along three main lines: evaluation dimensions, datasets, and technical routes. Existing approaches are categorized into four technical routes: manual subjective evaluation, traditional statistical and rule-based modeling, deep learning and representation learning, and large language model and multimodal intelligent evaluation. A three-layer progressive evaluation indicator system is constructed, comprising the textual language layer, lyric-music layer, and artistic aesthetic layer. The findings reveal four key bottlenecks: the deep gap between automatic quantitative metrics and human subjective aesthetic experience, the modeling blind spot for Chinese phonological features and lyric-music coordination, the severe scarcity of dedicated annotated datasets with uneven cultural distribution, and the disconnect between static text quality evaluation and dynamic human-computer collaboration experience. Future research should advance in three directions: developing lyric-specific vertical large models integrating music theory, constructing audio-text cross-modal joint lyric-music evaluation systems, and building generation-evaluation-optimization closed-loop ecosystems for the entire creative workflow.

Keywords

AI-generated lyrics, lyric quality evaluation, natural language processing, pretrained models, text quality assessment

1. Introduction

1.1 Evolution and Current Deployment of AIGC in Music Creation

Artificial intelligence in music creation has experienced three phases: rule-based, deep learning-based, and large model-based [1,2]. Initial algorithmic composition was based on probabilistic structures like Markov chains and was limited by forgetfulness and style [1]. The advent of neural networks led to the development of architectures like RNN, LSTM, and Transformer and enhanced the representation of melodies and long-term dependencies [1]. In recent years, the concept of music creation has been extended to music understanding,

human-computer cooperation, and scenario-based tasks by large language models. AI lyrics writing, composition, arrangement, and others are all used as means [2].

AI songwriting is basically an implementation of large language models wherein the main goal is to produce text that satisfies requirements in a particular format [1]. Baidu Wenxin Lyric Generator can finish the creation in 10 seconds [1]. Suno is known as the “ChatGPT” of the music world and it can create complete pieces of work that include lyrics and melody when combined with prompts [1,3]. The Tencent Music Qimingxing AI platform has already created over 26 million AI works [4]. Jiang Ruijie studied mainstream music AI applications and found that they all have different benefits in creating melodies, selecting arrangements, controlling rhythm, and synthesizing voices, and can help musicians create complete works more efficiently, but even with high-quality creation, human-computer interaction and fine-grained tuning cannot be avoided [5]. AI music creation has changed its orientation to the demonstration of technology to the area of its industrial use and the quality control of what has been created has become the topic of concern in the industry.

1.2 Core Quality Defects of AI-Generated Lyrics and the Need for Objective Evaluation

Nevertheless, the quality issues of AI-based lyrics have gained a greater presence. Si Ruo et al. noted that AI-generated pieces differ significantly in terms of their quality, where there are clear distinctions between them in terms of thematic suitability and emotional correctness; the share of usable content that may be used directly in professional writing is relatively small, and authors have to sort through many options [6]. Yang Zhikai et al. mentioned that AI-generated music also has some defects, including shallow emotional expression, the absence of substance, an unpredictable emotional tone, and the lack of coherence between paragraphs of long musical compositions [4]. Wu Hongwei discussed this issue at the level of generation logic: AIGC is biased towards local maximum rationality, but as the sequence grows longer, it does not receive any top-level logical direction, resulting in chaos in the development logic and collapse of the macro-structure [7].

The aforementioned issues are especially noticeable when it comes to the lyrical text because lyrics do not only need semantic fluency but have many other requirements like emotional expression, adaptation to musical styles, rhyme and prosody, and coordination between lyrics and music [8,9,10]. Currently, the quality assessment of AI-generated lyrics in the industry is largely based on manual subjectivity scores and does not have standardized automated evaluation protocols. It not only limits iterative improvement of generation models but also raises the sifting price of creators. An objective, measurable, multidimensional lyric quality assessment framework has become the most significant prerequisite to developing AI music creation beyond the state of being able to create music to actually creating good music.

1.3 Main Content and Organization of This Review

The present paper deals with qualitative assessment of texts of AI-generated lyrics, which separates it into the category of literary studies on human-written lyrics and qualitative evaluation of AI music as a whole, and examines the current state of the research, technical approaches, and development tendencies in the given area. The paper is structured like this: Chapter 2 discusses the central connotation and theoretical background of lyric quality evaluation; Chapter 3 builds a multi-level indicator system for lyric quality evaluation; Chapter 4 reviews the history of evaluation technical routes and offers a holistic comparison; Chapter 5 discusses the present situation of the construction of datasets and evaluation standards; Chapter 6 identifies the main bottlenecks and challenges; Chapter 7 projects the future development trends; Chapter 8 concludes the entire paper.

2. Core Connotations and Theoretical Foundations of Lyric Quality Evaluation

2.1 Multiple Constraint Characteristics of Lyrics

The main feature that separates lyrics and plain text is their several limitations. AI-generated lyrics are the lyrics created by big language models or any other natural language generation technologies, based on the constraints defined by the user, such as theme, emotion and musical style [1,3]. In comparison to human creation, they have these features: Firstly, fragmentation and local rationality. AIGC is probabilistic prediction-based; the choice of each word tends to be locally maximally rational, but as sequence length grows, no top-level logical control is present, and narrative clues are broken, theme drift occurs, and macro-structure

collapses [7]. Secondly, inadequate emotional expression depth. AI reproduces human expression by acquiring the paradigm of emotions in data but does not have a real-life experience; produced emotions tend to be shallow and cannot support deep emotional stories [4,6]. Thirdly, poor prosody and lyric-music synchronization. Rhyme design, tone matching, syllable rhythm, and melody adaptation are highly demanded in Chinese lyrics and the problem of “inverted characters” is a common one [6,10]. The current models are mainly concerned with the semantic level and cannot be used to model musical properties. Fifth, stylistic mediocrity. The underlying space of AIGC is influenced by a gravitational pull of the previous inertia, which drags every generation back to the same average level, which leads to massive stylization and lack of novelty [7].

2.2 Transformation Paths of Traditional Lyric Creation Theory in Machine Evaluation

The assessment of lyrics quality is based on the convergence of lyric creation theory and text quality theory. In the view of lyric creation theory, Lin Pufeng has summed up the creation methods into four aspects, namely thematic narrative, structural rhyme, narrative technique, and literary aesthetics [8]. Ni Yongdong also discussed it in the context of two facets of emotional expression and language art, noting that emotion can mean straightforward lyricism, indirect lyricism, or rhetorical usage, whereas language art may be expressed through rhythmic pattern, word selection, and sentence formation; the two are interdependent and mutually reinforcing [9]. Chen Zhitong, in terms of music analysis, suggested that there are four levels of lyric prosody, which are phonological elements, lyric-music coordination, genre-specific prosodic adaptation, and emotional prosodic sublimation [10]. Huang Zhiwen subsequently suggested three dimensions in which lyrical evaluation should take place, including: the emotion, which refers to an emotional style (architectural style), the conception, which means an expression of intentions (architectural technique); and the rhyme, which means rhyme forms (constructing materials). They have a hierarchical relationship of “emotion > conception > rhyme”; when needed, it is important to sacrifice the chariot to save the king- rhyme must be sacrificed to conceive the idea, and the idea must be sacrificed to the emotion [11]. These studies offer theoretical support to build a multidimensional evaluation system. The main path of transformation is to convert qualitative descriptions of creation techniques into quantifiable measures, such as translating the concept of structural rhyme into the rhyme detection rate and the rhyme scheme match level, and the concept of literary aesthetics into rhetorical density, imagery density, etc.

According to the text quality assessment theory, indicators of semantic similarity, emotional congruence, text fluency, and style match are generally used in evaluating generated texts [2]. Liu Jing stated that outcome evaluation of human-computer collaborative creation contains three types of indicators, which include acoustic quality, semantic alignment, and structural fidelity. The automatic indicators and human subjective judgment are separated, and it is hard to completely automate musicality and aesthetic preference [2]. That gap means that it would not be sufficient to use common text evaluation measures to rate a lyric, but rather it is necessary to modify them to suit a specific domain because its evaluation is based on its musical properties and aesthetic features.

2.3 Formal Definition of the Lyric Evaluation Task

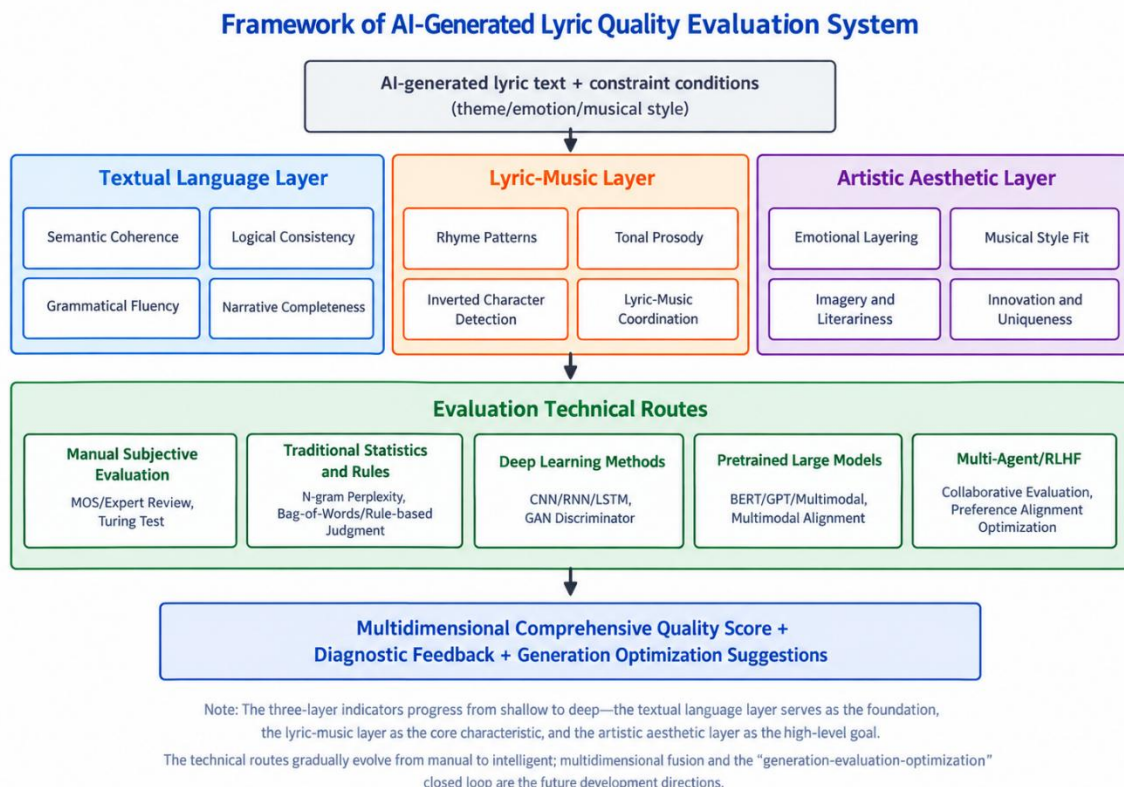
Theoretically, the problem of assessing the quality of AI-generated lyrics may be formulated as follows: provided with generated lyric text L , user constraints C (theme, emotion category, musical genre, intended audience, etc.) and optional reference melody M , an evaluation function $f(L, C, M)$ produces multidimensional quality score vectors $Q=(q_1, q_2, \dots, q_n)$ and the total quality score S . Dimensions q_i are related to particular evaluation indicators in the textual language layer, lyric-music layer and artistic aesthetic layer. Semantic ambiguity and polysemy of constraints C , the complexity of modeling the interaction between lyrics and melody M based solely on text, the subjectivity of artistic aesthetic layer indicators and trade-off between global structure and local quality in long-text evaluation are the central issues in the evaluation task. The formal definition above can serve as a common basis of the development of the indicator system and technical route.

3. Construction of a Multi-Level Lyric Quality Evaluation Indicator System

According to the theoretical discussion in Chapter 2, this paper builds a three-layer progressive evaluation indicator system of the lyric quality of the form of textual language layer-lyric-music layer-artistic aesthetic layer, as illustrated in Figure 1. The system starts with simple language features, followed by essential qualities

of music, and finally advanced aesthetic properties that can satisfy all the assessment requirements of AI-generated lyrics.

Figure 1: Framework of AI-generated lyric quality evaluation system



3.1 Textual Language Layer: Semantic Coherence, Logical Consistency, and Narrative Completeness

The textual language layer is the initial aspect of the evaluation system and analyzes the fundamental quality of lyrics as text. Regarding semantic coherence, Lin Pufeng noted that the initial stage of the process of writing lyrics is the definition of the theme and narrative outline [8]. Current approaches tend to rely on semantic similarity to compare embeddings of generated lyrics and theme descriptions however, themes are often written indirectly in lyrics via imagery and metaphor and so simple keyword matching is not sufficient to make a correct judgment. Logical consistency In terms of logical consistency, Wu Hongwei noted that AIGC tends to be logically confused when dealing with lengthy structures [7]. The logical reasonability of lyrics is reflected in the narrative sequence of verse and chorus, emotional bond between paragraphs, and the uniformity of person and perspective. Automated assessment of this dimension is still challenging. Narrative completeness evaluates whether lyrics contain a full narrative structure of introduction, development, transition and conclusion. AI-generated lyrics have structural issues like too much repetition of choruses, absence of bridge, and sudden end. Sentence fluency typically employs perplexity indicators, but these indicators can only measure statistical plausibility and do not find logical contradiction.

3.2 Lyric-Music Layer: Rhyme Patterns, Tonal Prosody, Character-Tone Matching, and Lyric-Music Coordination

The lyric-music layer is the fundamental distinguishing feature that distinguishes lyrics from plain text. The soul of lyrics is prosody. The analytical model suggested by Chen Zhitong involves the breakdown of phonological elements (type of rhyme, arrangement of syllables, alignment of tones), regulation of the coordination of lyrics and music (alignment of character tone and pitch, syllable rhythm and beat, articulation aperture and vocalization fit), genre-based prosodic adaptation, and emotional prosodic sublimation [10]. Rhyme evaluation may be further divided into perfect rhyme and near rhyme, stanza rhyme variation, and bar-

level rhyme schemes [8]. Syllable evaluation concerns monosyllabic balance and inter-line rhythmic difference [10]. Tone pitch should identify the inverted character phenomenon, which means awkward singing due to the mismatch of the character tone and the direction of melody pitch [6,10]. It is important to mention that, according to Huang Zhiwen, rhyming cannot be regarded as an absolute standard of lyric measurement; what can be set to music is a lyric. Identical rhyme is the best choice, close rhyme is a good option, and different rhymes are also able to generate great works (e.g., Tian Mi Mi and Yue Ban Xiao Ye Qu). Rhyme evaluation should be evaluated in its entirety, considering emotional expression and expression of conception and not just based on rhyming [11].

The best demand on the lyric-music layer is lyric-music coordination, which checks the level of match between the rhythm of lyric syllables and the beat, the appropriateness of the tone of the characters and pitch, and the correspondence between articulation aperture and vocalization [10]. Nevertheless, automatic measurement of prosody and literariness is a recent weak area. Detection of rhyme and syllable statistics may be programmed, but assessment of rhyme quality and subtle analysis of lyric-music coordination do not have a developed methodology. Current AI-based evaluation of lyrics does not pay much attention to the aspects of Chinese prosodic characteristics and lyric-music coordination, which is a severe gap in the literature [10].

3.3 Artistic Aesthetic Layer: Emotional Layering, Genre Fit, Classical Imagery, and Literary Depth

The artistic aesthetic layer is the highest level dimension of the evaluation system and considers the artistic value and creative level of lyrics. With regard to the emotional layering, Ni Yongdong classified the emotions evoked by lyrics into groups like love, family love, friendship, and patriotism, and the ways of expressing them were divided into two categories, i.e., direct lyricism and indirect lyricism [9]. But emotion assessment has several issues: the various studies make use of different emotional spaces that cannot be compared due to the combination of the basic emotions of Ekman, the valence-arousal model of Russell, and the emotion categories of music, among others [2]; AI-generated emotions tend to be shallow in nature [4]; assessing emotional coherence in extended lyrics is not done well; and quantifying emotional depth is very challenging, and the border between the concepts of sincere and formulaic cannot be differentiated by automatic measures.

Genre fit is concerned with whether the language style, prosodic characteristics, and imagery selections of lyrics are compatible with the given genre. Chen Zhitong noted that guofeng, lyrical, folk, and rock all have their own prosodic aesthetics [10]. Current assessments generally rely on text classification models to derive the genre and then compare it with the target, however style fitting is not a straightforward category comparison, it also considers more profound attributes like imagery choice, language elegance level, and sentence rhythm, and AI generation is prone to stylistic mediocrity [7]. Regarding literary depth, Lin Pufeng observed that good quality lyrics need to be characterized by ingenious text polishing, proper application of rhetoric, and imagery expression [8]. Ni Yongdong stressed that rhetoric should be measured because too much piling has negative effects on natural fluency [9]. Huang Zhiwen further develops his hierarchy theory of emotion, conception, and rhyme to clarify that emotional expression (emotion) is the main commander of the lyrics, conception conveyance (conception) is the backbone support, and rhyme form (rhyme) is the external decoration. The assessment of the literary depth of AI-generated lyrics must focus on the emotional sincerity and novelty of conception first, instead of concentrating on formal techniques [11]. Automated evaluation of literary measures, including imagery richness, rhetorical appropriateness, and recognition of memorable lines, is still an unsolved problem.

4. Evolution and Comparison of Technical Routes for AI-Generated Lyric Quality Evaluation

4.1 Manual Subjective Evaluation and Auditory Turing Test

Manual evaluation has been the most traditional way of doing things and is the nearest to a genuine aesthetic experience, which includes expert ratings, average opinion score (AOS), user surveys, and music Turing tests. Its strength is its ability to measure aesthetic qualities that cannot be measured by automatic indicators. As Liu Jing noted, human judgment is more realistic in terms of aesthetic experience, however, it is limited by work labels, task settings, listener background, and experimental conditions [2]. In random pairing, listeners cannot

tell whether AI or human-generated music has been played in music Turing tests; however, with highly similar styles, AI can be identified using technical properties [2], and creator identity labels also have a significant effect on aesthetic rating [2]. The disadvantages of manual evaluation are that it is expensive, not easy to apply on a massive scale, and very subjective. It is typically applied as a calibration standard and additional validation of automated techniques.

4.2 Traditional Statistical and Rule-Based Modeling Methods

The classical statistical and rule-based approaches to modeling are the evaluation of fluency using N-gram, style categorization using the TF-IDF model, and quality scoring using the logistic regression or support vector machine model and rhyme detection engines based on rules. The concept of N-gram is founded on the Markov chain principle and measures fluency by means of perplexity [1]. It is very computationally efficient and highly interpretable but has “forgetfulness” and is difficult to capture long-range semantic dependencies [1]. Genre classification or theme matching can be done with bag-of-words features including TF-IDF, which do not take into consideration word order or contextual semantics. Pinyin comparison is a basic tool in rhyme detection and rhyme scheme recognition, and rhyme rule engines can be used to assess lyrics music layers but cannot evaluate the aesthetic quality of rhymes. On the whole, traditional techniques prove useful in the shallow task of fluency, crude style separation, and rhyme recognition, but fail to capture deep semantics, emotional complexity, and literariness, and eventually give way to deep learning.

4.3 Deep Learning and Representation Learning Methods

CNN-based models, RNN-based models, LSTM-based models, and GAN-based models are some of the latest mainstream avenues. The convolutional kernels enable CNN to extract local n-gram features, and they are commonly applied in genre classification, emotion classification, and fluency discrimination. RNN maintains time information by using hidden-layer recurrence, addressing the forgetfulness issue of Markov chains [1]. LSTM adds gating controls that are more effective when dealing with vanishing gradients in long sequences and performs well in capturing long-range dependencies and is frequently used in evaluating emotional consistency and thematic coherence [1]. The discriminator of GAN is able to tell apart real lyrics and AI-generated lyrics, but training is hard and there are cases of mode collapse [1]. Simple deep learning models are able to represent more complex semantics, but are unable to represent music-specific characteristics of the lyrics, such as prosody, tone, and lyric-music synchronization. Also, they lack interpretability and use large-scale annotated data.

4.4 Large Language Models and Multimodal Intelligent Evaluation

The assessment of pretrained models is a hot area of research, such as BERT-type understanding models, GPT-type generative models, multimodal large models, multi-agent collaboration, and RLHF evaluation. The BERT-type bidirectional encoding models have powerful semantic understanding, and they are capable of emotion classification, theme matching, and semantic coherence analysis. They are better at semantic understanding tasks than unidirectional RNN/LSTM because of their bidirectional attention. Autoregressive models like GPT can score fluency on perplexity and style and emotion fit on conditional generation probability, and they have instruction-following properties such that direct multidimensional scores can be achieved by prompt engineering [1]. Evaluation of multimodal large models is multimodal cross-modal audio-text alignment. MuLan projects music audio and natural language into a shared embedding space using a dual-tower encoder and contrastive learning [2]. In the case of lyrics, it can also take into consideration the suitability of the text to the melody and thus evaluate the lyric-music coordination together.

Multitask decomposition into subtasks is used in multi-agent collaborative evaluation. ComposerX splits music creation into five links, which are: 1) requirement analysis, 2) composition, 3) arrangement, 4) lyric writing, and 5) quality evaluation, performed in a chain by several large language model agents. The quality evaluation agent is particularly tasked with the multidimensionality of judgment [2], which enhances interpretability and systematicity of evaluation; however, issues like agent communication protocols and failure attribution of collaboration have not been addressed yet. RLHF uses human preference data to train reward models. MusicRL consists of three kinds of reward signals: MuLan text adherence, MOS acoustic quality predictor, and Bradley-Terry user preference, and fine-tunes MusicLM using REINFORCE. Experimental results demonstrate that human preference signals can enhance subjective quality, but there is

no guarantee that the improvement will be completely captured by automatic measures, including FAD [2]. This concept can be extended to the optimization of lyrics generation. Regarding joint lyric-melody generation and assessment, Liu Ruixue presented two models, MelodyTransformer and StructureTransformer, which are both oriented to ordinary users and professional musicians to address the lyrics-to-melody (L2M) problem. They search and splice melody fragments to extract lyric structure information and add a bidirectional self-attention module to maximize melody coherence, offering a generation-side technical reference to lyric-music coordination evaluation [12]. On the whole, semantic understanding, multidimensional evaluation and human-computer cooperation have great benefits of the pretrained model method, however, the challenge remains in lack of sufficient domain fine-tuning, absence of Chinese prosodic feature modeling and poor interpretability of evaluation.

4.5 Comprehensive Performance Comparison of Various Evaluation Schemes and Technical Selection Guide

In order to systematically compare the features and the conditions under which they can be applied of different evaluation systems, this paper performs an extensive comparison between five aspects, i.e., evaluation dimension coverage, computational efficiency, interpretability, data dependence, and aesthetic capture ability, as indicated in Table 1.

Table 1: Comparison of different evaluation methods

Evaluation Technical Route	Dimension Coverage	Computational Efficiency	Interpretability	Data Dependence	Aesthetic Capture	Typical Application Scenarios
Manual Subjective Evaluation	Full-dimension	Low	High	None	Strongest	Benchmark calibration, final evaluation verification
Traditional Statistics and Rules	Shallow + Prosody	High	High	Low	Weak	Fluency initial screening, rhyme detection
Deep Learning Methods	Semantics + Emotion	Medium	Low	High	Medium	Offline model evaluation, batch scoring
Large Language Model Evaluation	Multidimensional Synthesis	Medium	Medium	Medium	Relatively Strong	Online interactive assistance, multidimensional diagnosis
Multi- Agent/RLHF	Full-dimension + Preference	Low	Medium	High	Strong	Generation optimization closed loop, personalized evaluation

According to the above comparison, technical selection needs to be classified depending upon the application contexts: in offline model evaluation, deep learning techniques may be applied to evaluate large batches of scores, with manual sampling calibration; in online interactive lyric-writing support, large language models are more appropriate since they have multiple dimensional diagnostics and real-time response benefits, and may be integrated with conventional rule engines to detect real-time prosody; in generation model training and optimization, RLHF and multi-agent collaborative evaluation can incorporate human preferences to reward signals to achieve a generation-evaluation-optimization cycle. Fusion schemes are multidimensional because they attempt to combine the benefits of different paths. The current approaches usually rely on weighted sums or multi-task learning, however the calculation of weights per dimension does not have a consistent foundation. Text adherence, acoustic quality, and stylistic innovation are also directions where tensions occur and breakthroughs are needed in the future.

5. Current Status of Dataset Design and Evaluation Benchmarks

5.1 Inventory of Existing Public Lyric Datasets and Their Quality Limitations

The data form the basis of training evaluation models. The existing public datasets consist primarily of human-written lyric data and exclude examples generated by AI and high-quality annotations. MQAD is based on the Million Song Dataset and includes around 270K songs and almost 3M question-answer pairs, however it is mostly used to perform music understanding activities and does not have lyrics quality annotations [2]. Available lyric datasets are mostly aimed at the task of lyric generation and offer extensive lyric texts and metadata (songwriter, album, genre, year, etc.) but they do not contain manual annotations of the quality dimensions that makes them hard to apply directly in the process of training an evaluation model. Very few studies use automated annotation (like quality level division according to perplexity thresholds), but the difference between automated annotation and actual aesthetic evaluation is quite large.

5.2 Difficulties and Protocol Specifications in Constructing Dedicated Annotated Datasets for AI-Generated Lyrics

Annotated datasets specifically designed to be used with AI-generated lyrics are exceptionally rare. Most of the current research employs small-scale datasets developed by them with issues like lack of uniform annotation criteria and poor genre representation. There are three significant challenges that arise during the construction of dedicated datasets: Firstly, the distribution of training data is not balanced. The music of the global north takes up approximately 86 percent of the length of AI music datasets, and non-Western genres only around 5.7 percent [2]. There is not enough coverage of Chinese lyric data, and it is necessary to add conscious samples of subdivided genres, including guofeng, folk, and rap. Secondly, the cost of annotations is expensive. Annotation of lyric quality needs annotators with musical and literary literacy, and multidimensional annotation consistency cannot be easily assured; specific annotation instructions and control measures of quality should be designed. Thirdly, samples of AI-generated samples should vary in the samples of different generation models, various prompt conditions, and varying quality levels so as to prevent bias in the dataset. Unified annotation specifications will be needed in the future, and more refined datasets with prosody annotation, tone annotation, and lyric-music coordination annotation will be required to offer an infrastructure to support training of evaluation models and benchmark evaluation.

6. Current Key Bottlenecks and Challenges

6.1 Deep Gap Between Automatic Quantitative Metrics and Human Subjective Aesthetic Experience

The disconnect between automatic indicators and human assessment of aesthetics is the most fundamental issue in this area. Experiments with MusicRL indicate that human preference signals may enhance subjective quality, yet automatic indicators like FAD do not capture this. Indicators derived by machine translation [2] Perplexity and BLEU are not entirely appropriate to evaluate lyrics. Various evaluations have various dimensions of evaluation, emotional space, and quality levels, hence comparison of these results is challenging to make across them [2,4]. The majority of research tends to be limited to a single dimension and does not have a holistic model that considers both semantics and emotions, musical style and style, and literariness. Moreover, the assessment of products and the assessment of processes are also detached. Available assessments tend to be based more on static quality and do not take into account interaction round, time cost, and sense of control in human-computer cooperation. It is essential to move away from the concept of scenario effect verification and build up an evaluation chain involving the indicators of generation, the responses of users, and the outcomes of applications to bridge the gap between automatic indicators and human judgment [2].

6.2 Modeling Blind Spots in Chinese Phonological Features and Lyric-Music Coordination

Chinese lyrics also have their own phonological properties, such as the tonal system (Mandarin has four tones), the rhyme category system (13 rhyme categories), and level and oblique tone regulations. These characteristics are closely linked to the melodic pitch and the rhythmic beat, which constitute the relationships

of lyric-music coordination [10]. Nonetheless, the current evaluation systems pay much attention to the semantic aspect, and they do not include modeling the quality of Chinese rhymes, matching the tones, lyric-music coordination, and literary images at all [10]. The pretrained models are not sufficiently optimized in the lyric space, and large and general models are hard to apply directly to structural constraints, musical features, and aesthetics of lyrics. The diagnostic assessment indicates that music reasoning of large language models cannot be considered stable [2]. Long-sequence modeling is problematic. AI-written lyrics are susceptible to theme drifting, abrupt emotional shifts, and disorganized paragraph flows in longer-length lyrics [1,4,7]. The ability to culturally adapt is poor, and the understanding of classical imagery, classical Chinese language, and traditional rhyme in guofeng lyrics is low [4]. Blind spots in the modeling of Chinese phonological properties and lyric-music coordination are critical bottlenecks that hinder the quality of Chinese AI lyrics assessment.

6.3 Disconnect Between Static Text Quality Evaluation and Dynamic Human-Computer Collaboration Experience

Current evaluation models are largely in the experimental phase and do not have an engineering solution to AI lyric writing tools [6]. The assessment is generally an offline single evaluation, and there is no generation-evaluation-feedback closed loop [2]. When it comes to the human-computer collaboration context, there are still many areas where screening depends greatly on the aesthetic judgment of the creators, and the AI-assisted evaluation capability is inadequate [3, 6]. Jiang Ruijie stated that despite the potential of music AI software to produce full compositions, the best quality in creation cannot be achieved without interaction and cooperation among people and machines. The human creators should make minor adjustments to the generated melody, arrangement, and lyrics to fulfill particular emotional expression requirements. Not only does this interaction enhance the quality of work, but it also enriches the auditory experience [5]. The measurement of the quality of static text is based on the measurement of the final quality of the generated results, whereas the measurement of the dynamic experience of human-computer collaboration is based on the measurement of process indicators such as interaction efficiency, sense of control, and the feeling of satisfaction during creation. There is a considerable gap between the two. An evaluation system highly prone to statistical reasonableness might contribute to the mediocrity of produced lyrics. To address the needs of future evaluation systems, they should include the static result quality as well as the dynamic process experience, incorporate evaluation within the whole creative process, and achieve real-time quality detection, intelligent filtering and decision-making. Besides, Luo Hao mentioned that the advanced fusion of technical innovation in digital music devices like audio plugins with music art would offer technical resources to implement AI-assisted creation and quality evaluation in engineering [13].

7. Future Development Trends and Prospects

7.1 Development of Lyric-Specific Vertical Large Models Integrating Music Theory

Large vertical models specific to lyrics with the integration of music theory should be created in the future. The knowledge of the domain (structural constraints, prosodic rules, tone matching, and coordination of lyrics and music) should be infused into the pretrained models to come up with lyric-specific understanding and evaluation models [2,10]. There is a possibility to use a general model + culture adaptation layer approach [2]. In particular, pretraining tasks, including prediction of lyric prosody, matching of tone-pitch, and identification of musical-form structure, can be added during the pretraining step to have the model learn about music theory knowledge; then multidimensionally annotated data can be utilized at the fine-tuning stage to perform domain adaptation. As far as long-range structure modeling is concerned, hierarchical modeling may differentiate between levels (phrase, paragraph, and overall musical form), long-term memory may retain themes and emotional cues, and structured control may consider musical form, rhythm, and emotional arc as explicit constraints to address theme drift and structural collapse of long lyrics [2,7].

7.2 Audio-Text Cross-Modal Joint Lyric-Music Evaluation System

The assessment of the lyric quality should not be restricted to the textual level, as it is necessary to develop an audio-text cross-modal joint lyric-music evaluation system. Through integrating melody audio information in order to construct a joint lyric-music evaluation model, automatic evaluation of lyric-music coordination could be achieved [2]. The dual-tower encoder and contrastive learning structure of MuLan may also be used

[2] to project lyric text and associated melody audio to a shared embedding space and learn joint representations of lyric-music coordination. The lyric-to-melody generation model suggested by Liu Ruixue motivates melody creation through the extraction of lyric structure and emotion, and its lyric-melody alignment mechanism has the potential to offer technical insights on reverse lyric-music coordination evaluation [12]. Based on this, additional fine-grained indicators of lyric-music coordination are able to be built on such a basis, such as the character tone-pitch matching degree, syllable-beat adaptation degree, and the pronunciation-vocalization fit degree. In addition to being able to evaluate lyric quality more holistically, multimodal fusion assessment is capable of generating more diverse feedback signals to generation models that facilitate AI music creation beyond text generation to lyric-music coordinated generation.

7.3 “Generation-Evaluation-Optimization” Closed-Loop Ecosystem for the Entire Creative Workflow

The generation-evaluation-optimization closed-loop ecosystem of the creative process should be created in the future. The lyric evaluation agents with functions of task planning, tool invocation, and memory management, as well as feedback revision must be created to integrate evaluation in the overall creative process to achieve a generation-evaluation-revision closed-loop [2]. At the application layer, the engineering implementation of the evaluation system should be encouraged: providing real-time quality detection and feedback to the AI tools that write lyrics, giving reward signals and training preferences to generation models, and offering intelligent screening and supporting decision-making to human-computer cooperation [2,3,6]. When it comes to evaluation system and dataset construction, a multidimensional comprehensive evaluation system of indicators should be constructed that would be able to strike the balance between semantic fluency, thematic consistency, emotional correspondence, musical style correspondence, prosody norms, and literariness. One may refer to the creation technique framework formulated by Lin Pufeng [8], the two-dimensional framework of emotion and language art proposed by Ni Yongdong [9] and the major points of prosodic music analysis by Chen Zhitong [10], which turns the well-developed theories of human lyric creation into quantifiable indicators. The construction of dataset is a pioneer project. Common annotation standards are needed, and AI-created lyrical data sets with various styles, different qualities, and various emotions need to be increased so that efforts are made to reduce the cultural imbalance [2].

8. Conclusion

The present paper contains a systematic literature review of the current studies in the area of AI-generated lyric quality assessment. There are four technical directions in the field: manual subjective measurement, conventional statistics and rule-based models, deep learning and representation learning, and big language models and multimodal intelligent measurement. The evaluation dimensions have progressively broadened to multiple-dimension comprehensive evaluation, including theme, emotion, music style, prosody, and literariness. The paper builds a three-level incremental evaluation index system, which is textual language layer-lyric-music layer-artistic aesthetic layer, and performs a comparative analysis of overall performance and technical route selection.

Nevertheless, the modern-day field has several challenges: the gap between automatic quantitative indicators and human subjective aesthetic experience is great; there are gaps in Chinese phonology features and music-lyric coordination models; annotated datasets are few and their distribution across cultures is uneven; and the assessment of static text quality is not related to the experience of dynamic human-computer cooperation.

Future research must develop concurrently in three areas: creation of lyric-specific vertical large models incorporating music theory, creation of cross-modal audio-text joint lyric-music assessment systems, and creation of a generation-evaluation-optimization closed-loop system of the whole creative process. Not only will the enhancement of AI-created lyrics quality assessment improve the quality of output and user experience of AI-based lyric writers, but it will also facilitate AIGC music going beyond generating to generating well, which will be more useful to human artistic expression under the paradigm of human-computer collaborative creation.

References

- [1] LI Zhengqi, WANG Xin. Application of artificial intelligence technology in music creation and production[J]. Journal of Nanjing University of the Arts (Music & Performance), 2025(1): 183-188.
- [2] LIU Jing. Review of music artificial intelligence driven by large language models[J]. Computer Science, 2026, 53(8): 229-244. DOI:10.11896/jskx.260700167.
- [3] PI Jian. Application of artificial intelligence technology in music creation[J]. Musical Instrument Magazine, 2025(9): 68-71.
- [4] YANG Zhikai, WANG Shufang. Creative ecology, aesthetic value and emotional experience of AI-generated music[J/OL]. Journal of Yangtze Normal University, 2026.
- [5] JIANG Ruijie. Research on the application of music artificial intelligence in popular song creation[D]. Nanjing: Nanjing University of the Arts, 2024.
- [6] SI Ru, TIAN Xin. Research on application of artificial intelligence (AI) technology in film song creation[J]. Modern Motion Picture Technology, 2026(7). DOI:10.3969/j.issn.1673-3215.2026.07.007.
- [7] WU Hongwei. Confrontation between existing induction and avant-garde breakthrough: Aesthetic challenges and alienation turn of generative AI music and serious music creation[J]. China & World Cultural Exchange, 2026(13): 191-193.
- [8] LIN Pufeng. Exploration on lyric creation techniques of popular songs[J]. Journal of Fujian Preschool Education College, 2025(3): 56-59.
- [9] NI Yongdong. Analysis of emotional expression and language art in popular song lyric creation[J]. Yellow River Voice, 2025(15): 54-56. DOI:10.26938/j.cnki.CN42-1794/Z.2025.15.017.
- [10] CHEN Zhitong. Music analysis research on lyric prosody of Chinese popular music[J]. Popular Songs, 2026(9): 57-60.
- [11] HUANG Zhiwen. On the emotion, conception and rhyme of lyric creation[J]. Lingnan Music, 2015(3): 52-53.
- [12] LIU Ruixue. Music melody composition based on generative AI[D]. Wuhan: Wuhan Textile University, 2025.
- [13] LUO Hao. Integration of technology application and music art: Innovation of audio plugins in music creation[J]. Art Panorama, 2024(35): 55-57.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).