REFFLY: Melody-Constrained Lyrics Editing Model

Songyan Zhao[*], Bingxuan Li[*], Yufei Tian, Nanyun Peng
Department of Computer Science, University of California, Los Angeles
zhaosongyan@cs.ucla.edu

Abstract

Automatic melody-to-lyric (M2L) generation aims to create lyrics that align with a given melody. While most previous approaches generate lyrics from scratch, ***revision***—editing plain text draft to fit it into the melody—offers a much more flexible and practical alternative. This enables broad applications, such as generating lyrics from flexible inputs (keywords, themes, or full text that needs refining to be singable), song translation (preserving meaning across languages while keeping the melody intact), or style transfer (adapting lyrics to different genres). This paper introduces REFFLY (**RE**vision **F**ramework **F**or **LY**rics), the first revision framework for editing and generating **melody-aligned** lyrics. We train the lyric revision module using our curated synthesized melody-aligned lyrics dataset, enabling it to transform plain text into lyrics that align with a given melody. To further enhance the revision ability, we propose training-free heuristics aimed at preserving both semantic meaning and musical consistency throughout the editing process. Experimental results demonstrate the effectiveness of REFFLY across various tasks (e.g. lyrics generation, song translation), showing that our model outperforms strong baselines, including Lyra (Tian et al., 2023) and GPT-4, by 25% in both musicality and text quality.

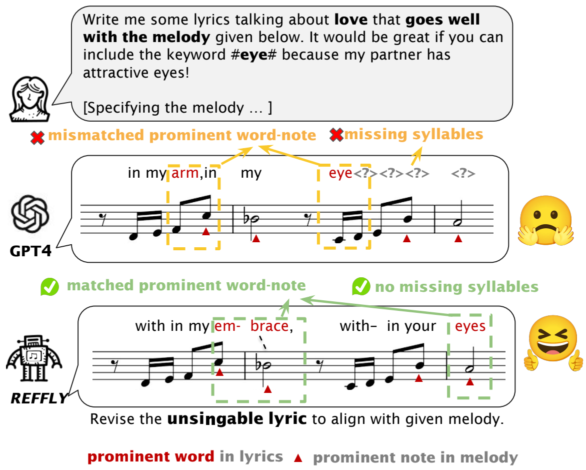


Figure 1: Human singers naturally emphasize certain words when singing, which align with prominent notes to ensure musical flow (details in § 3.1). However, LLMs like GPT-4 often misalign these prominent words (e.g."arm" with non-prominent notes) or omit syllables (e.g. no words for the last four notes), lowering lyric quality(§ 2). Our model, REFFLY, refines less singable drafts into melodically-aligned lyrics while preserving the meaning. Listen to the audios for an intuitive sense.

1 Introduction

Music acts as an important universal language that facilitates social connection and strengthens community involvement (Cross, 2009). Automatic melody-to-lyric (M2L), creating lyrics that are aligned with a given melody, has emerged as a promising task and received interest by the AI community, because it makes the process of music creation more accessible to a wider audience.(Sheng et al., 2021; Tian et al., 2023; Ding et al., 2024).

In practice, amateur songwriters may wish to craft lyrics that goes with their favorite melody with desired content (Tian et al., 2023; Qian et al., 2023), or translate songs into different languages for a wider audience or adapt existing lyrics to fit a different melody (Longshen et al., 2023; Nikolov et al., 2020). However, existing M2L approaches and AI-assisted songwriting frameworks fall short to support these use cases, due to insufficient control over sentence-level semantic. Most prior works generate lyrics with zero or limited user input such as keywords or topics (Sheng et al., 2021; Tian et al., 2023; Ding et al., 2024). Some prior works rely on in-filling text templates without providing sufficient automacy to the user (Zhang et al., 2020; Liu et al., 2020). In contrast, our revision framework refining plain text into melody-aligned lyrics offers greater flexibility and control.

In addition, both state-of-the-art LLMs and prior

[*]Equal contribution

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11295–11315
April 29 - May 4, 2025

works on lyrics generation struggle with producing singable lyrics that align well with a specific melody. For instance, as illustrated in Figure 1, ChatGPT-4 (OpenAI, 2023) generates coherent lyrics but fails to synchronize syllables with the last four music notes. Moreover, prominent words like 'arm' and 'eye' are paired with less prominent notes, disrupting the overall musical flow and resulting in low prosody. Similarly, prior works on lyrics generation generation either don't consider melody as a constraint (Zhang et al., 2022b; Ram et al., 2021; Zhang et al., 2020; Liu et al., 2020; Sun et al., 2022c), or overlooked the important relationship between prominent note in melody and prominent lyric words (Sheng et al., 2021; Tian et al., 2023; Qian et al., 2023), lowering the generated lyric quality.

Addressing these challenges, we propose a novel ***revision framework***, REFFLY, which transfers a draft prose to structured and singable lyrics align with a piece of melody. To enhance melody-lyric alignment, we develop a training-free heuristic for capturing prominent lyrical words and musical notes (§3.1). Since the melody-aligned lyric data is scarce due to copyright constraints, we design an instruction-based mechanism to guide LLMs towards highly singable lyrics by training on a synthetic dataset (§ 4.1). REFFLY can generate full-length songs with lyrical verses that develop the song's plot and message, and choruses that repeat a memorable musical motif.

Our contributions are summarized as follows:

- We propose the first melody-constrained lyric revision framework that, given a predefined melody, transfers an arbitrary text (also referred to as a ***draft*** or ***unsingable lyrics***) to a full-length, melody-aligned lyrics with high singability and prosody (also referred to as ***revised*** or ***singable lyrics***), with sentence-level semantic control.
- We introduce a training-free heuristic for capturing melody-lyrics alignment, semantically and musically, to improve both ***singability*** and ***prosody***. Correspondingly, we also contribute a expert labeled dataset with fine-grained annotations of music sheets. [1]
- In comprehensive experiments across two settings: **1)** generation of lyrics from user-specified inputs, and **2)** translation of lyrics from Chinese to English, REFFLY significantly enhances lyrics-melody alignment and text quality of the generated lyrics, resulting in a 25% and 34% improvements over strong baselines in terms of musicality and overall preference, respectively.[2]

2 Problem Setup and Background

2.1 What Makes a Good Lyric?

Great lyrics harmonize with the melody, blending musicality (*e.g.,* singability, prosody) with textual quality (*e.g.,* coherence, creativeness) (Perricone, 2018). Here, we elaborate the two terms related to musicality below:

- **Singability** is what makes a song easier to sing. For example, it is considered *not* singable when one single music note maps to a multi-syllable word (*e.g.,* beau-ti-ful) in the lyrics (Tian et al., 2023).
- **Prosody** measures whether melody and lyrics rise and/or fall together (Perricone, 2018). Lyrics with good prosody highlights prominent words by matching them with prominent notes. For example, in Figure 1, REFFLY enhances expression by stressing prominent words like 'embrace' and 'eye' by aligning them with prominent notes.

These concepts guided the development of heuristics to better align lyrics with the melody.(§ 3.1).

2.2 Task Formulation

Goal Given a predefined melody and a plain-text draft, our goal is to revise the unsingable draft into *full-length* lyrics that excel in both musicality and textual quality.

Formulation We consider full-length songs with the ***verse-chorus*** structure For example, the music in Figure 5 has the structure of <*verse 1, chorus 1, verse 2, chorus 2*>. Formally, the input melody $\mathbf{M}$ can be defined as a sequence of T substructures $\mathbf{M} = \{\mathcal{M}_{<tag_1>}, ..., \mathcal{M}_{<tag_T>}\}$, $tag_i \in \{\text{verse, chorus}\}$. Each $\mathcal{M}_{<tag_i>}$ consists of K_i music phrases (*i.e.,* $\mathcal{M}_{<tag_i>} = \{p_{i1}, p_{i2}, ..., p_{iK_i}\}$), where each music phrase further contains N_{ij} music notes (*i.e.,* $p_{ij} = \{n_{ij_1}, n_{ij_2}, ...n_{ij_{N_{ij}}}\}$). Here, each music note has three attributes: pitch (*i.e.,* how high or low it sounds), duration (*i.e.,* how long it lasts), and offset (*i.e.,* when it starts). The output is lyrics $\mathcal{L}$ that aligns with the input melody at the all granularities (*i.e.,* music notes, phrases, and substructures): $\mathcal{L} = \{w_{11_1}, w_{11_2}, ..., w_{ij_l}, ..., w_{TK_TN}\}$. Here, w_{ij_l} is a

[1] Dataset source: https://bit.ly/3X6nCqu

[2] Demo: https://bit.ly/4fGKWT3

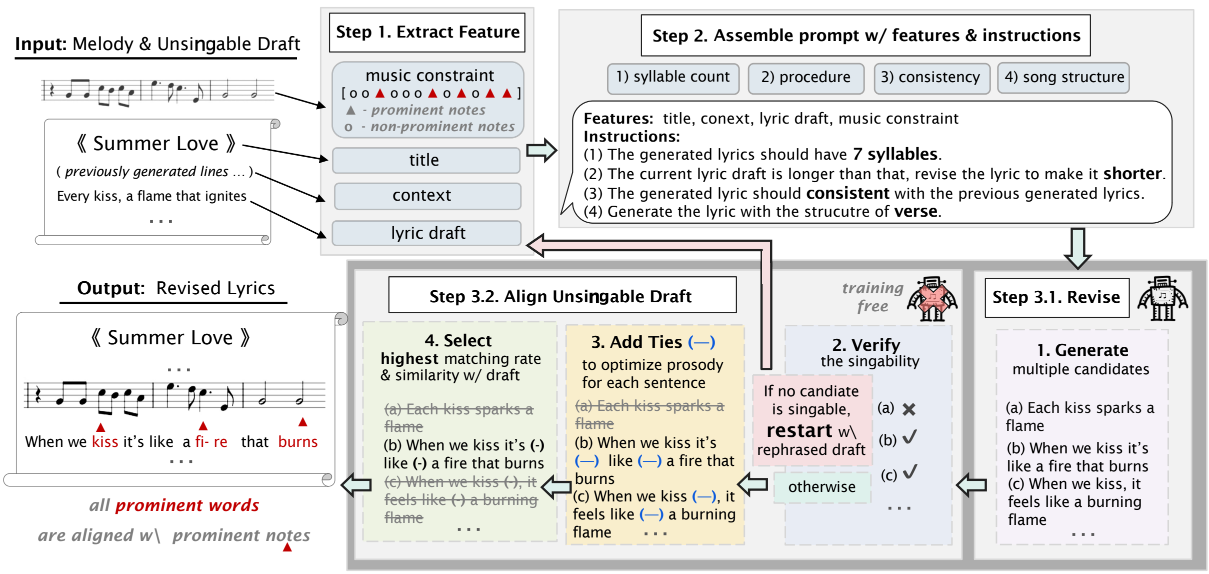


Figure 2: The overview of the inference process of REFFLY, an iterative approach to revise each sentence from the unsingable draft based on corresponding music constraint that is extracted from the music score. REFFLY begins by taking the melody and the unsingable draft as inputs. It then extracts features and constructs a prompt (Steps 1 and 2). Subsequently, it prompts a trained revision module to revise the unsingable draft (Step 3.1) and aligns the revised draft with the melody constraints using an alignment algorithm (Step 3.2). ***Note that only Step 3.1 requires training, and all other processes are training-free***.

word or a syllable of a word that aligns with the music note n_{ij_l}.

3 A Revision Framework for Lyric Generation

Figure 2 illustrates the inference process of REFFLY. To manage the complexity of lyric revision, the revision process is conducted at the sentence level. We iteratively revise each sentence from the unsingable draft to lyric that fits the melody, aligning the prominent words with prominent notes while maintaining the overall coherence.

In this section, we detail each component of REFFLY. We develop a training-free heuristic for capturing prominent lyrical words and musical notes (Section §3.1). Then, §3.2 introduces our lyrics revision module that refines unsingable drafts based on musical constraints. Last, §3.3 provides an overview of the inference process to achieve optimal lyric-melody alignment.

3.1 Aligning Melody with Lyrics

Building on the way experienced singers emphasize certain lyrics to enhance their connection with the melody for musical expressiveness (Robinson, 2005), we develop a heuristic to align prominent words with prominent musical notes.[3] This subsection outlines the process of identifying these prominent notes and words, for which we constructed an expert-annotated dataset to evaluate their effectiveness (see §5.1 for results). These heuristics are then used in lyrics generation (§3.3) and the construction of synthetic training data (§4.1).

Extracting Prominent Musical Notes We identify prominent musical notes that stand out in melodies based on three fundamental characteristics of music: *Time signature*, *Rhythm*, and *Pitch* (Caroline Palmer, 2006). A musical note is considered prominent if it appears on a stressed beat location as defined by its time signature, exhibits syncopation, or having a large pitch jump from the proceeding notes (More details in Appendix E.2).

Extracting Prominent words from Lyrics According to Reikofski (2015), nouns, verbs, adjectives are crucial for effectively conveying meaning. Therefore, we identify nouns, verbs, and adjectives that are non-stop words as prominent words.

Assessing the Accuracy To the best of our knowledge, we are the first to apply computational algorithms to identify prominent notes and words. Therefore, to evaluate the accuracy of our heuristic, we **collected a validation dataset** consisting of

[3] It is not feasible to use a neural network-based method for aligning prominent words with prominent musical notes due to the lack of such annotated data.

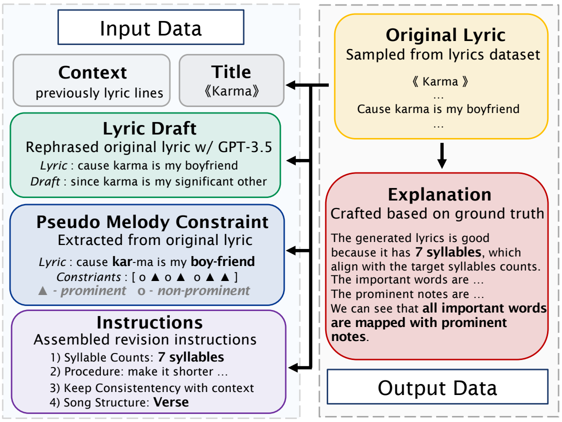


Figure 3: An exemplary data point in the fine tune dataset. The task is to use a rephrased input, title, music constraint, previously generated lyrics, and assambled instruction to generate the original lyrics, and some explanation. Rephrasing is done by GPT-3.5. During training, the revision model is guided by pseudo melody constraints derived from the original lyrics, enabling it to follow real melody constraints during inference.

100 song clips, each containing three to five musical phrases, annotated by professional musicians marking all the important notes in each melody. This dataset covers a diverse range of music styles, including Jazz, Country, Blues, Folk, Pop, and Comedy. Figure 6 shows an exemplary data point.

3.2 Lyrics Revision Module

To achieve high-quality lyrics revision, we fine-tuned a LlaMA2-13b-chat model (Touvron et al., 2023) to effectively transform an unsingable draft into lyrics that fit a given melody while preserving the original meaning. We address three main tasks: (1) Syllable planning: Generating sentences with the necessary syllable count to ensure singability. (2) Aligning prominent words with notes: Matching the stressed syllable of prominent words with prominent notes to enhance prosody. (3) Maintaining local and global coherence: Ensuring smooth transitions between sentences for local coherence and capturing global structures, such as verse and chorus, for structure-awareness. Addressing these tasks is challenging for LLMs, which are known to struggle with numerical planning (Sun et al., 2023c). Furthermore, we face a lack of labeled datasets to train a supervised model that learned to map the prominence in lyrics and melody.

To empower the model with the first two abilities, we propose a "pseudo music constraint" (blue box of Figure 3) to improve the syllable planning and word-note matching. The pseudo constraint, derived from generated lyrics, indicates prominent note positions and syllable counts. During training, the model follows pseudo constraints, while in inference, it applies melody constraints. This approach addresses both the lack of melody-aligned data and copyright issues associated with aligned data. We assume that the lyrics in the training data exhibit good prosody and are singable. we assign a special token, symbolizing a pseudo note, to each syllable in the sentence (singability assumption). If the syllable associated with the special token occupies the stress position of a prominent word, the token denotes an prominent note, otherwise it represents a less prominent note (good-prosody assumption). This way, we "back-translating" ((Longshen et al., 2023)) the pseudo music constraint from pure lyric-side data (refer to §D.1 for more details).

To maintain both local and global coherence, we introduce an instruction template (purple box of Figure 3). To ensure local, sentence-level, coherency, we provide the model with previously generated lyrics and the song's title as context for each training data point. To enable full-length song generation, our framework incorporates **structure-awareness** by embedding song structure into the fine-tuning phase. During this phase, we use song structure information (introduced in section 2.1) from the data source for each lyric line, allowing the model to recognize features of different song structures like verses and choruses. Structure tags are embedded within the instruction component and integrated into the prompt.

3.3 Generate Lyrics at Inference Time

As illustrated in figure 2, REFFLY takes each sentence-level melody and unsingable draft as input, to produce singable lyrics as the output, with three steps as following.

Step 1: Extract Features. Using heuristic in §3.1 we identify prominent musical notes from the melody, encoding them into a melody constraint. We then prepare this music constraint, title, previously generated lyrics as context, and lyrics draft as input features for further assembling prompt.

Step 2: Assemble Prompt. We assemble instructions that specify various features for the desired lyrics, such as matching the syllable count given melody, refining lyrics, providing previous context to ensure coherence, and guidance to maintain desired song structures. An example input prompt

Algorithm 1 Candidate Selection

input: List of candidates C, orignal draft o, melody constraint m, max number of ties K
output: Revised singable lyric

$c_{qualified}$ = empty list
for candidate c in C **do**
  c_{num} = calculate_num_ties(c, m)
  if $0 \leq c_{num} \leq K$ **then**
    $c_{tie} = add\_tie(c, m, c_{num})$
    $c_{qualified} += c_{tie}$
  end if
end for
for candidate c in $c_{qualified}$ **do**
  c_{best} =argmax(sim(c_{best}, o), sim(c, o))
end for
return c_{best}

can be found in Figure 9 in Appendix.

Step 3: Revise and Align. To enhance the model's ability to generating singable, prominence-aligned lyrics, we break down the process into two sub-steps, iteratively generating the lyrics.

Step 3.1 Revise: We adopt diverse beam search (Vijayakumar et al., 2016) to generate multiple candidate revisions of the unsingable draft, evaluating each for singability. A lyric is singable if: 1) each note corresponds to one or zero syllables; 2) each syllable in multi-syllable words matches a note no longer than a half-note; 3) multi-syllable words do not cross rests. If no candidates meet these criteria, we restart the process with a rephrased draft.

Step 3.2 Align Unsingable Draft: Algorithm 1 illustrates the alignment algorithm. It refines a list of lyric candidates to select the best match based on melody constraints, calculating ties and similarities to ensure a singable output. Note that there is no need for human to post process the lyrics. For each qualified candidate c and music constraint m, we determine the number of ties (a common musical notation that maps more than one notes to one syllable) to add using the following:

$$\#\text{Ties} = \#\text{Notes}(m) - \#\text{Syllables}(c) \quad (1)$$

Next, we define K, a tune-able hyper parameter, as maximum number of ties allowed within each musical phrase. We set $K = 2$ as a reasonable number in all of our experiments. If $\#\text{ties} < 0$ or $\#\text{ties} > K$, we reject the input. Otherwise, we explore all feasible positions to insert ties, aiming to maximize the number of prominent words mapped to prominent notes.

Finally, we select the candidate whose most important words align with prominent notes. If multiple candidates align perfectly, we choose the one most similar to the original sentence based on BERTScore (Zhang et al., 2019).

4 Experiments setup

In this section, we introduce dataset setup, experiment setting, baseline models, and evaluation setup.

4.1 Synthetic Training Dataset

As shown in Figure 3, the objective of our training dataset is to instruct the model to generate original lyrics by revising draft lyrics, following music constraints. We construct this dataset using 3,500 song-lyrics collected from the internet. Notably, our revision model only requires lyrics during training, alleviating the lack of aligned melody-lyrics data and potential copy-right issue.[4]

4.2 Tasks Setup

Our model's versatility is demonstrated through its performance across three distinct tasks. We prompt LLaMA2-13b in a few-shot manner to generate lyrics drafts based on user thoughts[5]:

1. Lyrics generation from arbitrary content This task generates song lyrics from scratch, starting with a draft based on scattered user thoughts. The lyrics' quality and melody alignment are evaluated using automated metrics and human judgment.

2. Full-Length generation with song structures (Structure-Aware Generation) In this task, we generate lyrics with specific structural requirements, starting from scattered user feedback. Domain experts then review these generated lyrics for coherence and clarity.

3. Song Translation This task focuses on translating lyrics from Chinese to English. The initial draft is a straightforward text translation produced by a translation model. We recruit bilingual evaluators to assess the translated lyrics quality.

4.3 Compared Models

We compare our framework with two **baselines** and introduce two **ablation variations** of REFFLY to validate each component.

[4] More details about the different components of input and output of training dataset can be found in Appendix D.1.

[5] Details on generating the lyrics draft are in Appendix C.2

Baselines. We compared REFFLY with three baselines. **(1). Lyra** is an unsupervised, hierarchical melody-conditioned lyric generator that can generate high-quality lyrics with content control without training on melody-lyric data (Tian et al., 2023). **(2). SongMass** is an LLM design that leveraging masked sequence to sequence (MASS) pretraining and attention based alignment modeling for lyric-to-melody and melody-to-lyric generation (Sheng et al., 2021). **(3). GPT-4** is a strong versatile LLM (OpenAI, 2023) to compare with. We utilize few-shot prompt to provide a template and instruct the model to follow it.

The comparison between REFFLY and the baselines is fair. We used ChatGPT-4-turbo as a baseline, prompted in a 2-shot manner with exemplary revisions and provided lyrics with serialized scores (via music21) to match REFFLY's input. For Lyra (Tian et al., 2023), we re-implemented it with LLaMA2-13b (replacing GPT-2), using the same lyric drafts as REFFLY and extracting three keywords per sentence with Yake (Campos et al., 2020), following the original setup.

Variations. We also conducted an ablation study to compare REFFLY with two variations. **(1). REFFLY w/o S.** is a variant of our proposed framework without the candidate selection algorithm (shown in the green and yellow boxes of Step 3.2 in Figure 2). **(2). REFFLY w/o I.** excludes the instruction component during training (purple and red boxes in Figure 3).[6]

4.4 Evaluation Setup

We conduct both automatic and human evaluations to assess our framework. While human evaluation is more reliable, it is difficult to scale and reproduce. Therefore, we use widely-adopted metrics like diversity, perplexity, and BERTScore to evaluate creativity, smoothness, and semantic similarity between generated lyrics and initial drafts (Sheng et al., 2021; Tian et al., 2023).

4.4.1 Automatic Evaluation

We evaluate the generated lyrics on text quality and melody alignment. For text quality, we assess several aspects: 1) **Diversity**, measured by calculating the number of unique n-grams in the text; 2) **Perplexity**, using GPT-2 to evaluate fluency and predictability; 3) **Similarity**, evaluated with BERTScore, to measure the similarity between our model-generated lyrics and the lyrics draft. For melody alignment, we proposed the **prominent word-note matching rate**, as explained in § 2.1, to measure how well prominent words are aligned with prominent musical notes.

4.4.2 Human Evaluation

Annotation Task We conducted a qualification task to select annotators with expertise in song and lyric annotation on Mechanical Turk. Additional details on the qualification process are provided in Appendix B.1. Our annotation process is comparative, with annotators reviewing groups of songs produced by various systems that share the same melody and title. All baseline models were assessed. At least three workers annotated each song, rating the lyrics' quality on a 1-5 Likert scale across five categories. For musicality, the workers assessed **prosody** (whether prominent words were exaggerated by melody), **intelligibility** (whether the lyric content was easy to understand when listen to it), and **singability** (how clearly the lyrics could be understood). In terms of text quality, they evaluated **coherence** and **creativity**. Annotators evaluated **structural clarity** (whether the verse-chorus structure is clear) structural-aware generation and assessed **translation quality** in song translation. The average inter-annotator agreement in terms of Pearson correlation was 0.69.

5 Results

In this section, we present results for validation of proposed heuristic, automatic evaluation, human evaluation, and ablation studies.

5.1 Effectiveness of the proposed heuristic

Our heuristic for identifying prominent notes and words is validated against baselines using a musician-annotated dataset of 100 song clips (§ further details in 3.1). The first baseline (Duration-only) relies solely on note duration to determine prominence, similar to the decoding constraints used in Lyra (Tian et al., 2023), and pairs this with our word extraction heuristic. The second baseline (Comprehensive w/o adj.) utilizes our note extraction heuristic but restricts prominent words to nouns and verbs.

As shown in Table 2, our prominent note extraction heuristic achieves an accuracy of 96%, substantially outperforming both baselines. Furthermore, our comprehensive heuristic for extracting

[6] We describe details of baselines in Appendix C.1.

Model	Automatic Evaluation					Human Evaluation				
	Diversity (Unigram)↑	Diversity (Bigram)↑	Similarity↑	Perplexity↓	Match Rate↑	Prosody↑	Coherence↑	Intelligibility↑	Singability↑	Creativity↑
Lyra	0.52	0.86	0.72	3305	0.48	1.97	1.66	2.02	1.83	1.70
SongMASS	0.50	0.76	—	3759	0.40	1.35	1.11	1.65	1.46	1.07
GPT-4	0.51	0.81	0.83	635	0.35	1.63	1.96	1.59	1.45	1.92
REFFLY w/o S.	0.54	**0.88**	0.78	1226	0.51	2.12	2.27	2.24	2.29	2.06
REFFLY w/o I.	0.51	0.81	0.74	635	0.59	1.98	2.01	1.87	1.93	1.74
REFFLY	**0.59**	0.87	**0.84**	**310**	**0.82**	**2.27**	**2.46**	**2.35**	**2.32**	**2.22**

Table 1: Evaluation Results for the Arbitrary Generation task. REFFLY and its variants (REFFLY w/o S. and REFFLY w/o I.) consistently outperform other models across most metrics, both in automatic and human evaluations.

	Note Extraction Success Rate	Alignment Success Rate
Duration-Only	74%	43%
Comprehensive w/o adj.	96%	65%
Comprehensive	**96%**	**91%**

Table 2: Comparison of three extraction and alignment strategies. The highest performance in each category is highlighted in bold, illustrating the superior effectiveness of our strategy in both note extraction (96%) and alignment (91%).

prominent words and notes yields a 91% alignment success rate[7], surpassing the best baseline by 26%. These results underscore the effectiveness and non-trivial nature of our approach in capturing the alignment between prominent words and prominent notes (more details in Appendix E.3).

5.2 Result of Lyrics Generation from Arbitrary Content

The results of automatic evaluation (mainly assesses fluency, topic relevance, and melody-lyric alignment) and human evaluations (assesses overall quality across multiple aspects such as musicality, creativity, etc.) are reported in Table 1.

Automatic Results The similarity scores in Table 1 indicate that REFFLY and GPT-4 excel in preserving the meaning of unsingable drafts. In contrast, SongMASS and Lyra surpass GPT-4 in terms of musicality, but at the cost of fluency. The qualitative example (shown in Section 6) shows that SongMASS and Lyra tend to generate cropped sentences to fit the music, leading to higher perplexity. Although GPT-4 matches REFFLY in retaining lyrical meaning, it falls short in diversity and melody alignment, as reflected in its lowest Match Rate. Overall, REFFLY surpasses all baselines, producing lyrics with superior textual quality, optimal melody alignment, and faithful preservation of the original draft's meaning.

Human Evaluation Results For melody-alignment quality, REFFLY achieves the highest scores in prosody, singability, and intelligibility. Lyra performs adequately but falls short compared to REFFLY, as it does not align prominent words and notes during generation. SongMASS and GPT-4 have much lower scores, suggesting that their lyrics may not fit well with the melody. This indicates that REFFLY excels in generating lyrics that align well with the melody are easy to sing.

For text quality, REFFLY scores the highest in both creativity and coherence, indicating its ability to generate lyrics that are both creative and contextually consistent. While GPT-4 performs reasonably well in text quality, its musicality remains poor. The other models score low in coherence, suggesting their lyrics may lack logical progression and contextual consistency.

Note that given the current limitations of open-source AI singing voice generation models, the quality of the singing in the generated songs may not meet human standards, which can lead to lower scores. Despite this, REFFLY outperforms all baselines under the same test settings by a large margin.

5.3 Ablation Study

We conduct an ablation study to validate each component in REFFLY. A qualitative example in Figure 7 (Appendix) compares variations of the model. Candidate selection algorithm select optimal candidate generated from revision model (see Figure 2). Instruction component in training simplifies the task for the revision module and enabling better context awareness. As shown in Table 1, removing the candidate selection (REFFLY w/o S.) or instruction mechanisms (REFFLY w/o I.) results in

[7]Alignment success rate is the accuracy of prominent words correctly mapped to prominent notes; note extraction success rate is the accuracy of extracting prominent notes

Model	Automatic Evaluation: Diversity↑ (Unigram)	Diversity↑ (Bigram)	Perplexity↓	Match Rate↑	Human Evaluation: Prosody↑	Coherence↑	Intelligibility↑	Singability↑	Creativity↑	Translate Quality↑
GPT-4	0.50	0.76	522	0.35	1.59	2.24	1.83	1.54	2.26	2.33
REFFLY	**0.59**	**0.87**	**310**	**0.83**	**3.28**	**3.08**	**3.25**	**3.08**	**2.69**	**3.04**

Table 3: Song translation task result. REFFLY scores the highest for all metrics.

	GPT-4	REFFLY
Prosody	1.31	**3.10**
Sinability	1.28	**3.11**
Coherence	1.84	**2.70**
Creativity	1.85	**2.62**
Intelligibility	1.26	**2.99**
Structural Clarity	1.15	**3.36**

Table 4: Structure-aware generation results: REFFLY outperforms GPT-4 by producing lyrics with a clearer song structure while maintaining lyric quality and melody alignment.

a noticeable performance decline in almost every metrics compared with REFFLY.

5.4 Result of Structure-Aware Generation

Among all baselines, only GPT-4 has the capability of generating full-length structured lyrics. Table 4 shows the results for the structure-aware generation task, where REFFLY achieved a 44% improvement in structural clarity. Figure 10 in Appendix shows the generated lyrics's clear verse-chorus-verse-chorus structure.

5.5 Result of Song Translation

Similar to the previous task, only GPT-4 has the capability of song translation, so it is our only baseline. Table 3 presents the evaluation result. REFFLY demonstrates a remarkable on average 23% increase in all human evaluation metrics. This results suggest REFFLY significantly enhances the quality, coherence, and melody-alignment of generated lyrics compared to GPT-4, making it more suitable for practical applications in song translation.

Interestingly, although GPT-4 generally has stronger translation abilities compared to LLaMA2-13b, REFFLY outperforms GPT-4 in translation quality by 14%. This suggests that successful translation in the lyrics-writing context requires not only high text quality, but also an emphasis on how well the lyrics sound when singing, as shown in Figure 4

6 Case Study

We conducted a case study to better understand the advantages of Reffly compared to baselines. An exemplary generated output is shown in Figure 4.

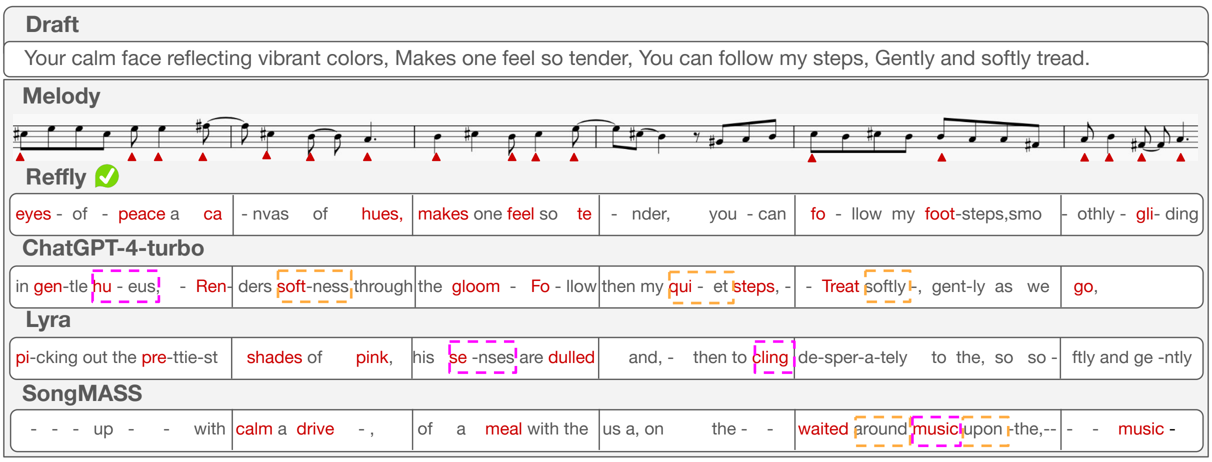


Figure 4: The output of different models given the same input draft. REFFLY is the only model that aligns lyrics with the melody while preserving the original meaning. Other models produce unsingable or low-prosody lyrics (introduced in Section 2.1). The orange box highlights the lyrics that is **not *singable***. For example, in SongMASS generated lyrics, 'a-round', a two-syllable word, is mapped to one musical note, making it hard to sing. The purple box highlights important words that **failed** to map to a prominent musical note (**low *prosody***), which would disrupt the expressiveness of lyrics. Listen to the audios in the demo page for an intuitive sense.

Musicality: Our model generates melody-aligned lyrics while preserving the original draft's meaning. Unlike the baselines, which overlook the importance of mapping prominent words to prominent notes, our approach ensures that the melody emphasizes these words. In Figure 4, the purple boxes highlight important words that are *failed* to map with prominent notes, and only REFFLY generate melody-aligned lyrics. In addition, SongMASS and ChatGPT-4 generate *unsingable* lyrics, indicated by yellow box.

Revision capability: REFFLY rephrases sentence structures or modifies words, adding ties to ensure prominent word-note alignment. For example, in Figure 4, it rephrased 'Your calm face reflecting vibrant colors' into 'eyes of peace, a canvas of hues', and 'You can follow my steps' into 'You can follow my footsteps'. In both cases, the original meaning is retained. Although both REFFLY and Lyra generate lyrics line-by-line, REFFLY produces coherent lyrics due to our training strategy that considers the context of previously generated lines.

7 Related work

Creative generation Melody-to-lyrics (**M2L**) generation is a form of creative generation, alongside tasks like pun (Tian et al., 2022; Sun et al., 2022a,b), poetry (Tian and Peng, 2022), hyperbole (Tian et al., 2021), and story generation (Yang et al., 2023, 2022; Chen et al., 2022; Han et al., 2022). M2L poses unique challenges as it requires consideration of musical characteristics (e.g., pitch, rhythm, time signature) and precise lyrics–melody alignment. Our framework addresses these challenges by being the first to model the relationship between prominent words and notes, ensuring better singability and prosody. Furthermore, to address the challenge of automatically evaluate creative generation (Ghazarian et al., 2022; Tian et al., 2024; Sun et al., 2023a), we propose the Prominent notes-words matching rate for assessing melody–lyrics alignment.

Controllable Lyrics Generation Recent works investigate efficient approaches for controllable generation tasks (Sun et al., 2023b; Li et al., 2024), yet controllable lyric generation requires a more specialized method to ensure both semantic coherence and musical alignment. While many lyric generators and AI-assisted lyrics writing systems have been developed to follow control signals like themes, rhyme, specific text format, or in-filling text template (Ma et al., 2020; Li et al., 2020; Liu et al., 2020; Zhang et al., 2020; Sun et al., 2022c), none provide *full control* over the sentence-level semantics of generated content and follow melody constraint; Fan et al. (2019) employed control mechanisms to generate lyrics based on specific topics, Tian et al. (2023) used keywords and genre to control the content, and Saeed et al. (2019), used music audio to condition the generation process. Other approaches have incorporated stylistic elements, such as rhyme schemes and meter or text format to influence the lyrical output (Potash et al., 2015; Zhang et al., 2020). Despite these efforts, they have insufficient sentence-level semantics and note-level music alignment. REFFLY can generate coherent, full-fledged high quality melody-aligned lyrics with sentence-level semantics control.

Melody-Lyrics Alignment LLMs have proven effective in the M2L generation task, with various attempts to integrate music representation (Lee et al., 2019; Qian et al., 2022). For example, Sheng et al. (2021) applied two transformers for cross-attention between lyrics and melody; Tian et al. (2023); Qian et al. (2023) considered duration of musical note and stressed syllables mapping, beat, or song structures during lyrics generation. ReLyMe (Zhang et al., 2022a), a lyrics-to-melody (L2M) generation model, considers the mapping of keywords to stressed beat locations but does not discuss how pitch influences the determination of prominent notes or how keywords are identified. While revision approaches have been explored in poetry generation (Andrea et al., 2021), REFFLY is the first framework for melody-constrained lyrics revision with sentence-level semantic control. Furthermore, existing M2L models fail to align prominent words with prominent notes, resulting in poor prosody. We propose a novel heuristic for melody-lyrics alignment, achieving a 26% improvement. REFFLY enhances lyric generation and emotional expression.

8 Conclusion

We introduced REFFLY, the first melody constrained revision framework to generate high-quality lyrics from plain text drafts while retaining original meaning. To enhance the lyrics-melody alignment, we designed a heuristic to identify and align prominent notes and words. Finally, we show REFFLY excel across diverse applicability.

Limitation

The limitation of our work include: 1) In this work, we use a rule-based method to identify important words in lyrics, specifically, nouns, verbs, and adjectives. Future work could investigate more nuanced definitions of important words. 2) Similarly, Our method for extracting prominent notes considers only two levels: prominent notes and other notes. While this simple approach has yielded satisfying results, exploring more fine-grained categories could potentially enhance performance. 3) Although our approach preserves the original meaning of the lyrics, the genre of the resulting lyrics largely depends on the training dataset. Future research could aim to provide more control over the genre.

Acknowledgments

The authors would like to thank the members of PLUSLab and the anonymous reviewers for helpful comments. This work is supported in part by NSF CAREER Award #2339766.

References

Zugarini Andrea, Pasqualini Luca, Melacci Stefano, and Maggini Marco. 2021. Generate and revise: Reinforcement learning in neural poetry. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Sean Hutchins Caroline Palmer. 2006. What is musical prosody? *Psychology of Learning and Motivation*, 46:245–278.

Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. 2022. Character-centric story visualization via visual planning and token alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ian Cross. 2009. The evolutionary nature of musical meaning. *Musicae scientiae*, 13(2_suppl):179–200.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

Haoshen Fan, Jie Wang, Bojin Zhuang, Shaojun Wang, and Jing Xiao. 2019. A hierarchical attention based seq2seq model for chinese lyrics generation. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part III 16*, pages 279–288. Springer.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. Deam: Dialogue coherence evaluation using amr-based semantic manipulations. In *Proceedings of the Conference of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. 2022. Go back in time: Generating flashbacks in stories with event temporal prompts. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. icomposer: An automatic songwriting system for chinese popular music. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 84–88.

Bingxuan Li, Yiwei Wang, Tao Meng, Kai-Wei Chang, and Nanyun Peng. 2024. Control large language models via divide and conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15240–15256, Miami, Florida, USA. Association for Computational Linguistics.

Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. Ai-lyricist: Generating music and vocabulary constrained lyrics. In *In Proceedings of the 58th annual meeting of the association for computational linguistics*, pages pp. 742–751. IEEE.

Nayu Liu, Wenjing Han, Guangcan Liu, Da Peng, Ran Zhang, Xiaorui Wang, and Huabin Ruan. 2020. Chipsong: A controllable lyric generation system for chinese popular song. In *In Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants*, pages pp. 742–751. IEEE.

Ou Longshen, Ma Xichu, Kan Min-Yen, and Wang Ye. 2023. Songs across borders: Singable and controllable neural lyric translation. In *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages pp. 447–467.

Xichu Ma, Ye Wang, Min-Yen Kan, and Wee Sun Lee. 2020. Rigid formats controlled text generation. In *In Proceedings of the 29th ACM International Conference on Multimedia*.

Nikola I Nikolov, Eric Malmi, Curtis Northcutt, and Loreto Parisi. 2020. Rapformer: Conditional rap lyrics generation with denoising autoencoders. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 360–373.

OpenAI. 2023. Gpt-4 technical report. *https://arxiv.org/abs/2303.08774*.

Jack Perricone. 2018. *Great Songwriting Techniques*, new edition edition. Oxford University Press.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924.

Tao Qian, Fan Lou, Jiatong Shi, Yuning Wu, Shuai Guo, Xiang Yin, and Qin Jin. 2023. Unilg: A unified structure-aware framework for lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 983–1001.

Tao Qian, Jiatong Shi, Shuai Guo, Peter Wu, and Qin Jin. 2022. Training strategies for automatic song writing: A unified framework perspective. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4738–4742. IEEE.

Naveen Ram, Tanay Gummadi, Rahul Bhethanabotla, Richard J Savery, and Gil Weinberg. 2021. Say what? collaborative pop lyric generation using multitask transfer learning. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 165–173, New York, NY, USA. Association for Computing Machinery.

Helen Dewey Reikofski. 2015. *Singing in English in the 21st century: A study comparing and applying the tenets of Madeleine Marshall and Kathryn Labouff*. University of North Texas.

Jenefer Robinson. 2005. *10 Emotional Expression in Music*, page 295–321. Oxford University PressOxford.

Asir Saeed, Suzana Ilic, and Eva Zangerle. 2019. Creative gans for generating poems, lyrics, and metaphors. *arXiv preprint arXiv:1909.09534*.

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13798–13805.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022a. Expunations: Augmenting puns with keywords and explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022b. Context-situated pun generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023a. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023b. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng, and Xuezhe Ma. 2023c. Evaluating large language models on controlled generation tasks. *arXiv preprint arXiv:2310.14542*.

Yusen Sun, Liangyou Li, Qun Liu, and Dit-Yan Yeung. 2022c. Songrewriter: A chinese song rewriting system with controllable content and rhyme scheme. *arXiv preprint arXiv:2211.15037*.

Yufei Tian, Divyanshu Arun Sheth, and Nanyun Peng. 2022. A unified framework for pun generation with humor principles. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP-findings)*.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. Unsupervised melody-to-lyric generation. *arXiv preprint arXiv:2305.19228*.

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2024. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yufei Tian, Arvind Krishna Sridhar, and Nanyun Peng. 2021. Hypogen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. Doc: Improving long story coherence with detailed outline control. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen Zhang, Luchin Chang, Songruoyao Wu, Xu Tan, Tao Qin, Tie-Yan Liu, and Kejun Zhang. 2022a. Relyme: improving lyric-to-melody generation by incorporating lyric-melody relationships. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1047–1056.

Le Zhang, Rongsheng Zhang, Xiaoxi Mao, and Yongzhu Chang. 2022b. QiuNiu: A Chinese lyrics generation system with passage-level input. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 76–82, Dublin, Ireland. Association for Computational Linguistics.

Rongsheng Zhang, Xiaoxi Mao, Le Li, Lin Jiang, Lin Chen, Zhiwei Hu, Yadong Xi, Changjie Fan, and Minlie Huang. 2020. Youling: an ai-assisted lyrics creation system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–91.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Exemplary Data

A.1 Example of full-length Song

[Verse 1]
Would you go with me if we rolled down streets of fire
Would you hold on to me tighter as the summer sun got higher
If we roll from town to town and never shut it down
Would you go with me if we were lost in fields of clover
Would we walk even closer until the trip was over
And would it be okay if I didn't know the way

[Chorus 1]
If I gave you my hand
Would you take it and make me the happiest man in the world
If I told you my heart couldn't beat one more minute without you girl
Would you accompany me to the edge of the sea
Let me know if you're really a dream
I love you so
So would you go with me

[Verse 2]
Would you go with me if we rode the clouds together
Could you not look down forever
If you were lighter than a feather
Oh, and if I set you free, would you go with me

[Chorus 2]
If I gave you my hand
Would you take it and make me the happiest man in the world
If I told you my heart couldn't beat one more minute without you girl
Would you accompany me to the edge of the sea
Help me tie up the ends of a dream
I gotta know, would you go with me
I love you so, so would you go with me

Figure 5: Example song with verse-chorus-verse-chorus structure

A.2 Example of validation dataset

Figure 6: Exemplary data points in validation datasets, where experts annotate the ground truth prominent notes, actual data points have at three to five musical phrases.

B Human Evaluation Details

Human annotators are paid with $ 20 per hours.

B.1 Qualification Task

To evaluate the Turkers' expertise in the field, we designed a task that included the initial verse from 9 different songs, each with ground-truth labels. These songs were chosen with care to avoid unclear cases, allowing for a clear assessment of quality. The selected songs were those whose scores showed a strong correlation with the ground-truth labels. We select 49 qualified annotators out of 87 annotators, based on Pearson correlation metric. The average inter-rater agreement interms of

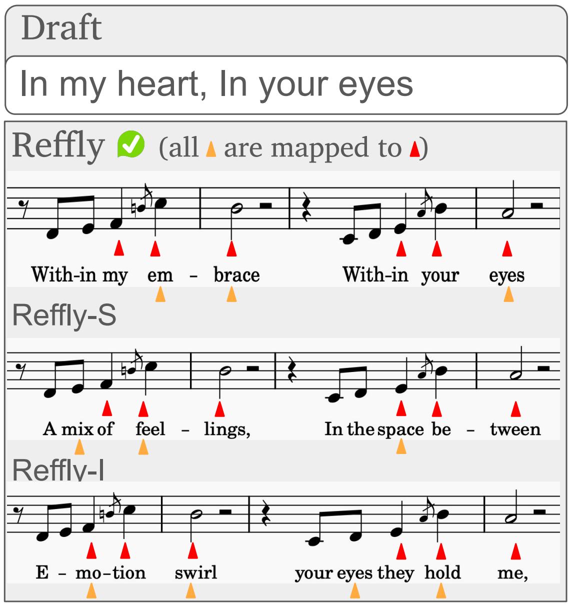


Figure 7: Example song

Pearson correlation among qualified annotators in qualification task was 0.43.

B.2 Annotation Task

We present the original survey, including evaluation instructions and the annotation task, in Figure 11 through Figure 14. Figure 11, Figure 12, and Figure 13 outline task instructions, defining each metric—intelligibility, singability, prosody, coherence, creativity—and accompanied by examples of good and bad lyrics for each criterion. Figure 14 display the actual annotation task.

C Experiments Details

C.1 Details regarding to baseline construction

ChatGPT We used ChatGPT-4-turbo as the base model to construct this baseline. In order to make this baseline to be fair, we tried our best to prompt ChatGPT-4. Firstly, we use 2-shot manner to prompt it: we provide two golden exemplary revision example every time. To make sure that ChatGPT have the same information that REFFLY has, we provided lyrics and the corresponding serialized score using music21, a format that zero-shot ChatGPT could understand. This score encompasses every detail about the music, including rhythm, pitch, and time signature. Note that extracting additional details, such as the position of prominent notes, would require the prominent note extractor from Reffly's framework. Our objective is to use

ChatGPT-4 as a baseline, not to replace Llama2-13b as a revision module.

Lyra Since the original Lyra paper (Tian et al., 2023) used GPT-2 as the base model, in order to make the comparison fair, we re-implemented the Lyra using Llama-2-13b. When doing the experiments, we use the exact same lyrics drafts as REFFLY, which is generated from a collected user prompt. Since Lyra requires keywords as inputs to generate each sentence, we use Yake (Campos et al., 2020) to extract three keywords from the lyrics draft, as the same setting as the original paper.

C.2 Example of the interface used to collect scattered user input

The Figure 8 illustrates the interface of our input-to-draft model. Initially, the user's requirements are extracted from the prompt using few-shot LLaMA2-13b with intent extraction examples. The extracted requirements are then presented to the user for confirmation, after which a draft is generated based on the confirmed requirements using LLaMA2-13b.

Note that we use the same revision model in all of arbitrary generation, full-length generation, and song translation. Only the input lyrics draft is different, which are generated by LLaMA2-13b in few-shot manner.

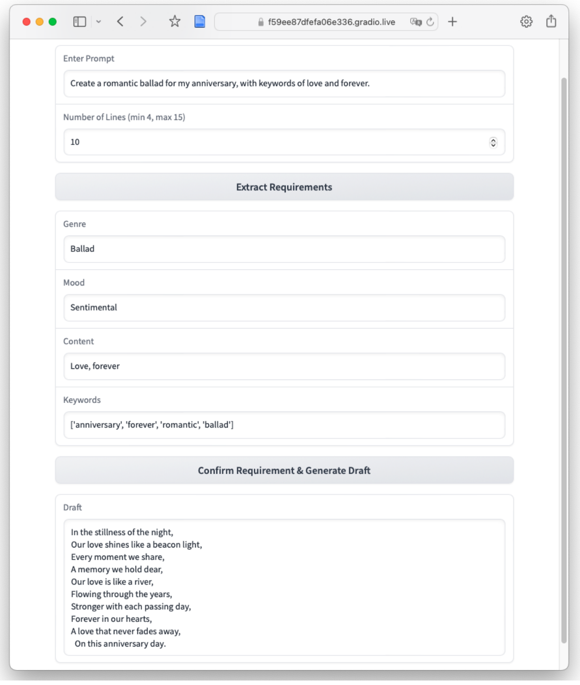


Figure 8: Interface used for input to draft

C.3 Computation cost

As illustrated in Figure 2, we "restart" if the all of the outcome of Step 3.1 is *unsigable*. Restart happens when the length of lyrics draft is too different from the length of music constraint. When rephrasing, we rephrase the original draft so that the length of draft is closer to the length of music constraint. When doing the experiment, all of the lyrics are generated within 3 iterations.

C.4 More details regarding to experiment metrics

We utilized the GPT-2 Large model, which has 774 million parameters, to calculate perplexity. Prominent word-note matching rate, or matching rate in Table 1, is the accuracy of stressed syllables of prominent words being correctly mapped to prominent notes.

D REFFLY Details

Input

Lyric that needed to be revised based on the music constraint: 'When you're tuning in during the nighttime'. **Previously generated lyrics** are: '(Verse1) If you're listening to this song, But they're not, We just wrote it like that.' It is in **Verse** section. Title is 'Only a Northern Song'. The **music constraint**: S_0: /UNSTRESSED/ S_1: /UNSTRESSED/ S_2: /STRESSED/ S_3: /UNSTRESSED/ S_4: /UNSTRESSED/ S_5: /UNSTRESSED/ S_6: /STRESSED/ S_7: /STRESSED/. The goal is to firstly, match the number of syllables in the music constraint, and secondly, match the important word to the /STRESSED/ syllables.The music constraint indicates that there **should be 8 syllables** in the generated lyrics. The original sentence has 10 syllables. Therefore, you should rephrasing the original sentence so that **generated lyrics have less syllables**.The **important words** in the original lyric is ['tuning', 'nighttime'], and the **syllables for each word** is When(/STRESSED/) you're(/STRESSED/) tuning(/STRESSED/-/UNSTRESSED/) in(/UNSTRESSED/) during(/STRESSED/-/UNSTRESSED/) the(/UNSTRESSED/) nighttime(/STRESSED/-/UNSTRESSED/).Therefore, we want to rephrase the sentence, so that 1, the number of syllables in the generatedlyric is 8 by rephrasing the original sentence so that generated lyrics have less syllables, 2, the stress of each of the important word in the generated lyric matches with the music constraint,and 3, it is fluent, singable, and coherent with the previously generated lyrics."

Output

We want to generate a lyric with 8 syllables, and **the generated lyric** is 'When you're listening late at night'. The corresponding syllables for each word is When(/STRESSED/) you're(/STRESSED/) listening(/STRESSED/-/UNSTRESSED/-/UNSTRESSED/) late(/STRESSED/) at(/STRESSED/) night(/STRESSED/). It has 1+1+3+1+1+1 = 8 syllables and matches with the total number of syllables in the music constraint (8 syllables). The important words in the generated lyric is ['listening', 'night'].The position of the stressed syllables of these important words are [2, 7], and S_2, S_7 are all '/STRESSED/'. The position of stressed syllable of important words in the generated lyric matches the music constraint.

Figure 9: Training data example.

D.1 Training Data Construction

Construct Input Data We use ChatGPT 3.5 to rephrase sentences either trivially or non-trivially (50% vs. 50%). Trivial rephrasing changes only a few words without altering the sentence structure. The **rephrased lyric** is then used as the input for revision module training. Additionally, we provide **instructions** to better guide LLMs revise lyric that better align with melody. The instructions contain information about 1) the phoneme of each word sourced from the CMU pronunciation dictionary, 2) the number of syllables in both the original and rephrased sentences, and 3) guidelines for modifying the lyric drafts, such as making the sentence shorter or longer.

Assemble Output Data The **original lyric**, prior to rephrasing in the first step, is used as the output for revision module training. Additionally, we craft an "**explanation**" paragraph help LLMs revise the lyric by breaking down the revision task into multiple simpler sub-tasks (more detail see Figure 3).

After processing, the inputs consist of the rephrased sentence, the song title, pseudo music constraints, the original lyrics, the song structure, and specific instructions. The output includes the original lyrics accompanied by an explanation.

We used the default LoRA implementation from the official LLaMA2-13b GitHub repository to fine-tune the model. The revision module was trained for 3 epochs using a dataset of 3,500 data points.

Exemplary training data point Figure 9 shows an example training data point constructed using the pipeline introduced in Figure 3. The original lyric is "When you're listening late at night." We generate a "pseudo melody constraint" based on this lyric, then use ChatGPT to create a rephrased lyric draft, "When you're tuning in during the night-time." The model is trained to generate the correct original lyric using the title, lyric draft, pseudo melody constraint, and instruction as input.

E Music Theory

E.1 Representation in Melody

The representation for a melody is hierarchical. A melody M consists of a series of musical phrase $M = (p_0, p_2, ..p_x)$, where x is the total number of musical phrase; Each musical phrase consists of a series of measures $p_i = (m_1, m_2, ...m_y), i \in [0, x]$, where y is the total number of measures in i'th musical phrase. Note that $|p_j \cap p_{j+1}| \leq 1$. The intersection equals to 1 when a musical phrase end in the middle of a certain measure, so the next musical phrase starts from the same measure. Each measure consists of a series of notes and a corresponding time signature. $m_k = (n_1, n_2, ..., n_z)$, where z is the total number of notes in measure m_k. Each note has four component: pitch, duration, offset, and tie. Pitch represents the highness/lowness of a note; duration is the length of the note; offset is the beat when this note starts in its measure; and tie can be start (a tie starts from this note), or continue (in between of a tie), or end (a tie ends at this note).

E.2 Prominent note extraction heuristic details

Inspired by prior research in music theory (Caroline Palmer, 2006), we develop a more comprehensive heuristics to identify prominent musical notes based on three fundamental characteristics of music:

1. *Time Signature*: This characteristic provides a structured framework that dictates how beats are grouped and accented within each measure. We identify notes that fall on strong beats or downbeats as prominent notes.

2. *Rhythm*: For this characteristic, we specifically examine *syncopation*, a musical technique that shifts emphasis to beats or parts of a beat where it is not usually expected. This technique breaks the conventional rhythmic pattern by highlighting off-beats or weaker beats within a measure. Notes that are accentuated using this technique are identified as prominent notes.

3. *Pitch*: From this characteristic, we particularly focus on *pitch jump*. Large pitch jumps contribute to contrast and variety in the melody line, thereby making notes with significant pitch jumps more conspicuous. We classify notes that exhibit significant pitch jumps as prominent notes.

Melody Melody is a sequence of musical tones, consisting of multiple musical phrases that can be further decomposed into timed musical notes. Each musical note has two independent aspect: pitch and duration. Pitch refers to the perceived highness or lowness of a sound; duration refers to the length of time that a musical note is held or sustained.

Time signature time signature organizes the rhythm and provides a framework for how the beats are grouped and accented within each measure. A time signature is represented by two numbers, one stacked on top of the other: the top number indicates the number of beats in each measure; the bottom number indicates the duration value that represents one beat. For example, 4/4 means a quarter note as one beat, 4 beats in a measure. Table 1 shows the stressed location for some commonly-seen time signature. The elements in the list is the number of beat that is stressed. For example, [0,2] means the first and third beats are stressed.

Table 5: Time signatures and their stressed locations

Time Signature	Stressed Location
4/4	[0, 2]
3/4	[0]
2/4	[0]
3/8	[0, 2]
6/8	[0, 2]
9/8	[0, 2, 5]
12/8	[0, 2, 5, 8]

Syncopation Syncopation refers to the displacement or shifting of accents or emphasis to unstressed beats. If a note is in unstressed beat with a longer duration than its previous note, then this note, although in unstressed beat, is stressed, or syncopated.

Pitch jump Pitch jump for two consecutive notes is the absolute difference of their pitch value. Larger pitch jumps create contrast and variety within the melody line.

If a note is in a metrical stressed position (indicated by time signature), or it is a syncopation, or it has a pitch jump (larger than average interval), we consider it as a prominent note, otherwise, it is a non-prominent note.

We also provide the mathmatical formulations for prominent note extraction:

1. Time signature:

As shown in table 5, a function can determine if a note is in an important location in terms of time signature. Suppose the time signature for this melody is T, and the corresponding list for stressed location is SL_T

$$\text{stressed}(n_i) = \begin{cases} 1 & \text{if if offset}(n_i) \text{ in } SL_T \\ 0 & \text{, otherwise} \end{cases} \quad (2)$$

2. Rhythmic type:

We implemented two simple rules:

1) for a given k consecutive notes $n_i, ...n_{i+k}$ that are connected by one tie, we replace $n_i, ...n_{i+k}$ as a new note n_i', and the duration for n_i' is $duration(n_i') = \sum_{j=0}^{k} duration(m_{i+j})$, and offset (beginning of n_i') is: $offset(n_i') = offset(n_i)$

2) After combining all notes that connected by tie, we check syncopation. If a note is in a weak location but its duration is longer than previous note, it is a syncopation.

3. Pitch:

A note is more important if there as a dramatic change in pitch compared to previous note. Based on this assumption, we have

$$\text{jump}(n_i) = \begin{cases} 1 & \text{if } \Delta\text{pitch}(n_i) > AIJ, \\ 0 & \text{otherwise}, \end{cases} \quad (3)$$

where

$$\Delta\text{pitch}(n_i) = |\text{pitch}(n_i) - \text{pitch}(n_{i-1})|,$$

$$AIJ = \frac{\sum_{i=1}^{x} \Delta\text{pitch}(n_i)}{x},$$

and x is the number of notes.

Collectively, the importance of note m_i is defined by

$$\text{M}(m_i) = \begin{cases} 1 & \text{if stressed, jump, or syn.}(m_i) \\ 0 & \text{, otherwise} \end{cases} \quad (4)$$

, where (0)1 means the note is an (un)important note and syn. stands for syncopated.

E.3 Results for heuristics

We provide more details for § 5.1 at here. Because our validation dataset only contains ground truth prominent note, we use Yake (Campos et al., 2020) algorithm to extract up to 3 keywords from one lyrics sentence, and treat the extracted keywords that correspond to ground truth prominent notes as ground truth prominent words.

Note that in the heuristic, we are consider the stressed syllable position by aligning stressed syllables of prominent words to prominent notes.

E.4 An illustrative example of how the tie is added

In figure 4, we need to add two ties to the first sentence "eyes of peace a canvas of hues", because there are two more notes in the melody than number of syllables. Here, we discuss why there is a tie added to “Do” instead of “Mi”. This is because at here, “Mi” would be identified as a prominent note (because it is in a stressed beat in 4/4) and “peace” as a prominent word (because it is a noun). Note that here “Do” is not a prominent note, because it is neither in a stressed location, nor a syncopated note, nor a note with big pitch jump from the previous note. To maximize the number of prominent words mapped to prominent notes, a tie would be added at “Do”, and the important word “peace” is corresponding to the prominent note “Mi”. This process is detailed at Step 3.2 (§3.3) and Algorithm 1, candidate selection algorithm. This entire process is handled algorithmically, without requiring human inspection.

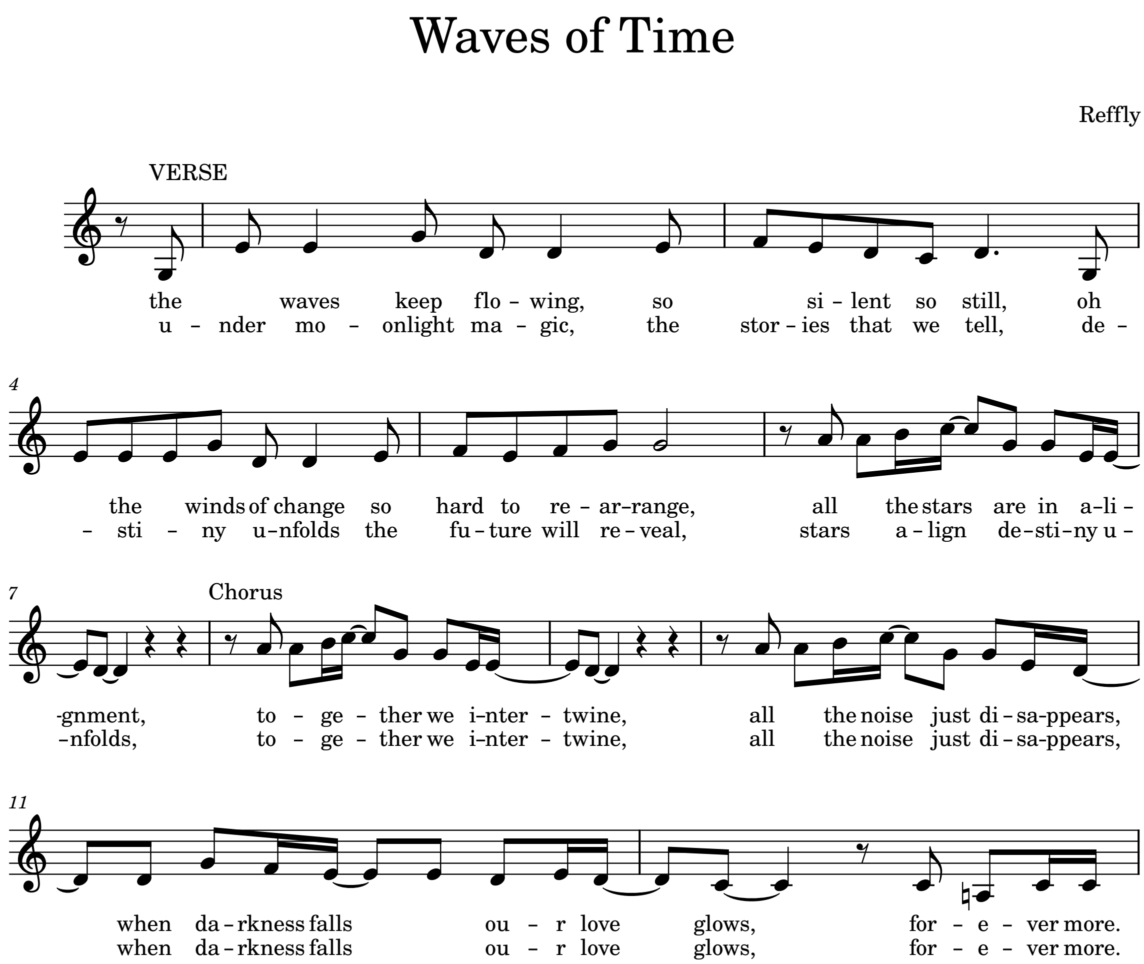
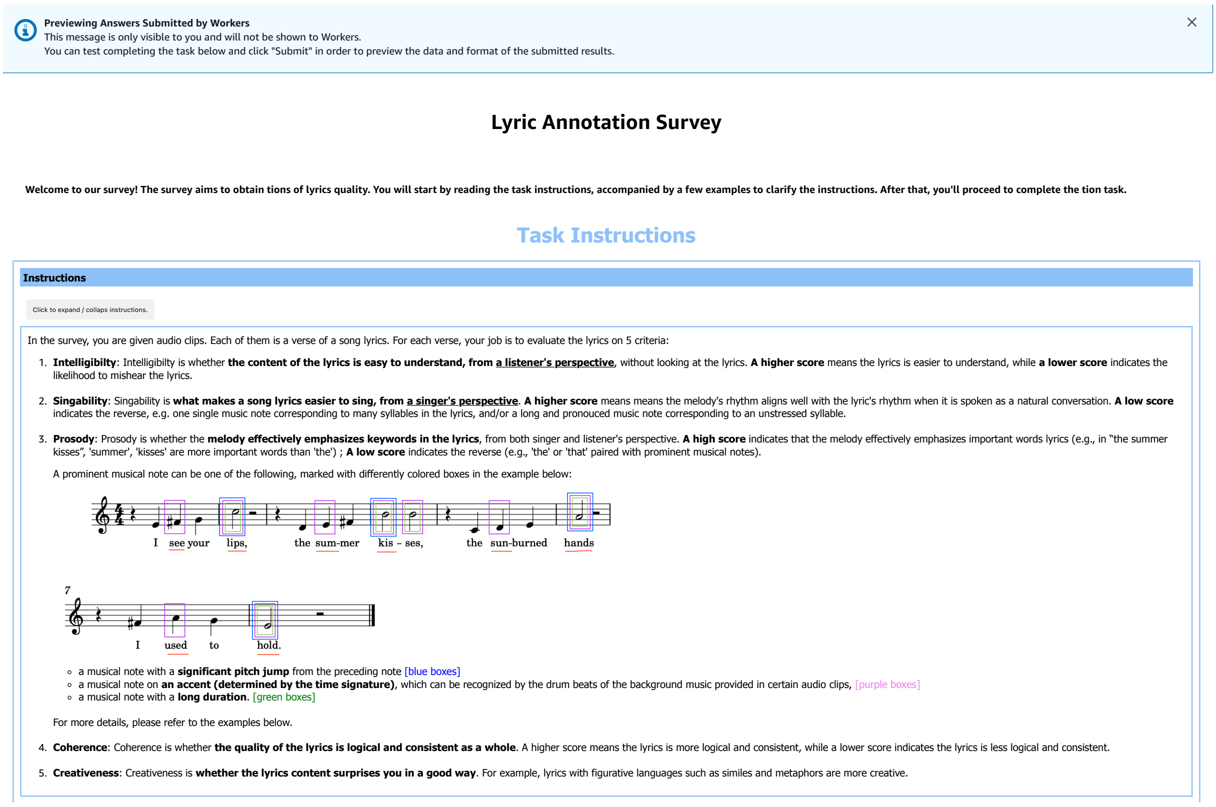


Figure 10: Exemplary generated song with verse-chorus-verse-chorus structure



Previewing Answers Submitted by Workers
This message is only visible to you and will not be shown to Workers.
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Lyric Annotation Survey

Welcome to our survey! The survey aims to obtain tions of lyrics quality. You will start by reading the task instructions, accompanied by a few examples to clarify the instructions. After that, you'll proceed to complete the tion task.

Task Instructions

Instructions

Click to expand / collaps instructions.

In the survey, you are given audio clips. Each of them is a verse of a song lyrics. For each verse, your job is to evaluate the lyrics on 5 criteria:

1. **Intelligibilty**: Intelligibilty is whether **the content of the lyrics is easy to understand, from a listener's perspective**, without looking at the lyrics. **A higher score** means the lyrics is easier to understand, while **a lower score** indicates the likelihood to mishear the lyrics.
2. **Singability**: Singability is **what makes a song lyrics easier to sing, from a singer's perspective**. **A higher score** means means the melody's rhythm aligns well with the lyric's rhythm when it is spoken as a natural conversation. **A low score** indicates the reverse, e.g. one single music note corresponding to many syllables in the lyrics, and/or a long and pronouced music note corresponding to an unstressed syllable.
3. **Prosody**: Prosody is whether the **melody effectively emphasizes keywords in the lyrics**, from both singer and listener's perspective. **A high score** indicates that the melody effectively emphasizes important words lyrics (e.g., in "the summer kisses", 'summer', 'kisses' are more important words than 'the') ; **A low score** indicates the reverse (e.g., 'the' or 'that' paired with prominent musical notes).

   A prominent musical note can be one of the following, marked with differently colored boxes in the example below:

   - a musical note with a **significant pitch jump** from the preceding note [blue boxes]
   - a musical note on **an accent (determined by the time signature)**, which can be recognized by the drum beats of the background music provided in certain audio clips, [purple boxes]
   - a musical note with a **long duration**. [green boxes]

   For more details, please refer to the examples below.
4. **Coherence**: Coherence is whether **the quality of the lyrics is logical and consistent as a whole**. A higher score means the lyrics is more logical and consistent, while a lower score indicates the lyrics is less logical and consistent.
5. **Creativeness**: Creativeness is **whether the lyrics content surprises you in a good way**. For example, lyrics with figurative languages such as similes and metaphors are more creative.

Figure 11: Human evaluation survey: task instruction

Examples

Click to expand / collaps examples.

We would like to show you a few examples of good and bad lyrics in each criterion.

Example 1: Lyric with good singability and intelligibility

Lyrics: take me back to the window, take me back to the door.

The rhythm of the lyrics is close to our natural conversation, which makes the lyrics easy to sing.

The lyrics is easy to understand when we first hear it, which makes the lyrics with good intelligibility.

Lyrics and melody are matched, which means the lyrics is in Lyrical-Melodic Alignment.

Therefore, the lyrics achieves **5** in singability, intelligibility and Lyrical-Melodic Alignment.

The lyrics makes sense to us, so it achieves **5** in coherence as well.

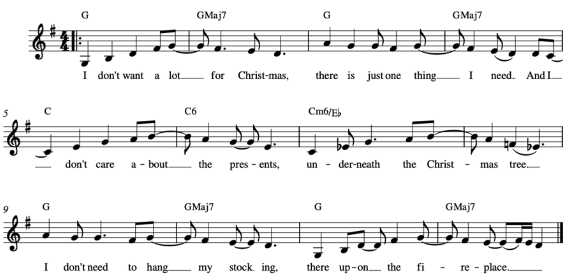
The content does not surprises us, so it achieves **3** in creativeness.

Example 2: Lyrics with good creativeness

Lyrics: you should bet on me, like I'm Apple in the '90s, you should bet on me, gonna wanna get behind me, like I'm 23, before Mikey was on Nikes, you should bet, bet, bet, bet on me. sunlight

The lyrics is creative because it uses similes (the comparison of one thing with another thing of a different kind). "I" am compared to Apple and Mikey, which shows my potential, therefore the lyrics achieves **5** in creativeness.

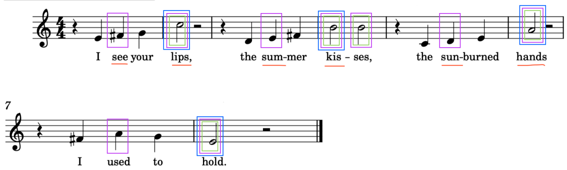
Example 3: Lyrics with good coherence



Lyrics: I don't want a lot for Christmas, there is just one thing I need. And I don't care about the presents, underneath the Christmas tree. I dont' need to hang my stocking, there upon the fireplace

The lyrics is coherent as the its content focus on a story of Christmas, therefore it achieves **5** in coherence.

Example 4: Lyrics with good prosody



Lyrics:I see your lips, the summer kisses, the sun-burned hands, I used to hold.

The rhythm of the lyrics is close to our natural conversation, which makes the lyrics easy to sing.

The lyrics is easy to understand when we first hear it, which makes the lyrics with good intelligibility.

The stress syllable of important nouns and verbs like "see", "lips", "summer", and etc. (marked by red lines) are mapped to prominent notes (marked by boxes), which makes the melody effectively emphasizes the keywords.

Therefore, the lyrics achieves **5** in singability and intelligibility.

The lyrics the lyrics achieves **5** in Prosody.

The lyrics makes sense to us, so it achieves **5** in coherence.

The content is novel, so it achieves **5** in creativeness.

Figure 12: Human evaluation survey: explanation of different metrics 1

Example 5: Lyrics with bad singability

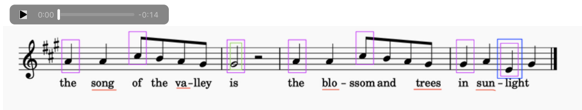
Lyrics: Challenges your think you know about, but are not fully ready.

The lyrics is difficult to sing with it as its rhythm strongly violates the rhythm when the sentence is spoken as a natural conversation.

For example, the word *'challenges'* is spoken with an accent on the first part (*'cha'*) and an unstress on the last part (*'ges'*), yet the melody has a long and strong note corresponding to unstressed part (*'ges'*), making it akward to sing.

In addition, there are cases where two syllables (*'cha-llenge'* and *'a-bout'*) that correspond with one music note, so the singer has to sing them in a hurry. Therefore the whole piece achieves **1** in singability.

Example 6: Lyric with bad prosody



Lyrics: The song of the valley is the blossom and trees in sunlight

The rhythm of the lyrics is close to our natural conversation, which makes the lyrics easy to sing.

The lyrics is easy to understand when we first hear it, which makes the lyrics with good intelligibility.

The unimportant words in lyrics (e.g. "The" in first and third measures; "in" in the last measure) are mapped to the prominent notes, but important nouns and verbs like "song", "valley", and "trees" are not being mapped to prominent notes, which makes the melody unable to emphasize the keywords.

Therefore, the lyrics achieves **5** in singability and intelligibility.

The lyrics get **1** in prosody.

The lyrics makes sense to us, so it achieves **5** in coherence.

The content is novel, so it achieves **5** in creativeness.

Example 7: Lyric with bad intelligibility

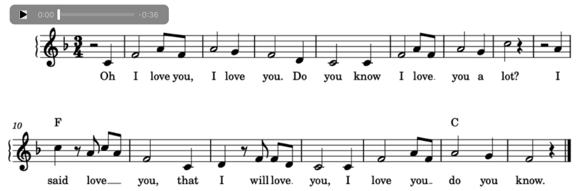
Correct lyrics: *got a long list of ex-lovers.*

Misheard lyrics: *all the lonely Starbucks lovers.*

The misunderstanding of the lyrics content when we hear the song means the lyrics is bad in intelligibility,

therefore, it achieves **1** in intelligibility.

Example 8: Lyric with bad creativeness



Lyrics: *Oh I love you, I love you. Do you know I love you a lot? I said love you, that I will love you, I love you do you know.*

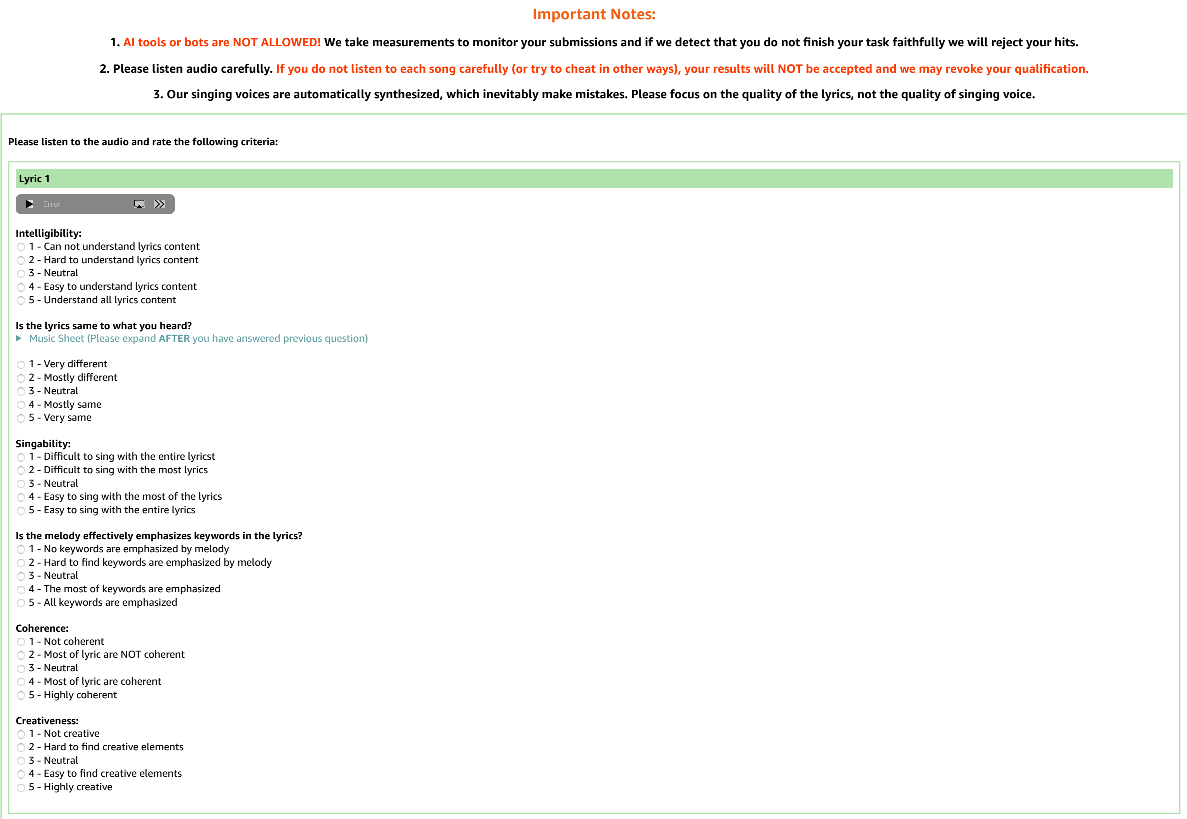
The lyrics keeps repeated, which does not bring us much new information, therefore, it achieves **1** in creativeness.

Example 9: Lyric with bad coherence

Lyrics: that there is the one thing complete, is when you find new places, the names of places that you once knew, trying to find your children, I'm trying to find my eyes, I deny that it brings me. sunlight

The lyrics is difficult to understand what it means because of grammar errors and the lack of main plot, therefore it achieves **1** in coherence.

Figure 13: Human evaluation survey: explanation of different metrics 2



Important Notes:

1. **AI tools or bots are NOT ALLOWED! We take measurements to monitor your submissions and if we detect that you do not finish your task faithfully we will reject your hits.**

2. **Please listen audio carefully. If you do not listen to each song carefully (or try to cheat in other ways), your results will NOT be accepted and we may revoke your qualification.**

3. **Our singing voices are automatically synthesized, which inevitably make mistakes. Please focus on the quality of the lyrics, not the quality of singing voice.**

Please listen to the audio and rate the following criteria:

Lyric 1

Error

Intelligibility:
- 1 - Can not understand lyrics content
- 2 - Hard to understand lyrics content
- 3 - Neutral
- 4 - Easy to understand lyrics content
- 5 - Understand all lyrics content

Is the lyrics same to what you heard?

▸ Music Sheet (Please expand **AFTER** you have answered previous question)

- 1 - Very different
- 2 - Mostly different
- 3 - Neutral
- 4 - Mostly same
- 5 - Very same

Singability:
- 1 - Difficult to sing with the entire lyricst
- 2 - Difficult to sing with the most lyrics
- 3 - Neutral
- 4 - Easy to sing with the most of the lyrics
- 5 - Easy to sing with the entire lyrics

Is the melody effectively emphasizes keywords in the lyrics?
- 1 - No keywords are emphasized by melody
- 2 - Hard to find keywords are emphasized by melody
- 3 - Neutral
- 4 - The most of keywords are emphasized
- 5 - All keywords are emphasized

Coherence:
- 1 - Not coherent
- 2 - Most of lyric are NOT coherent
- 3 - Neutral
- 4 - Most of lyric are coherent
- 5 - Highly coherent

Creativeness:
- 1 - Not creative
- 2 - Hard to find creative elements
- 3 - Neutral
- 4 - Easy to find creative elements
- 5 - Highly creative

Figure 14: Human evaluation survey: the annotation task