MusicRL: Aligning Music Generation to Human Preferences

Geoffrey Cideron [1] Sertan Girgin [1] Mauro Verzetti [1] Damien Vincent [1] Matej Kastelic [1] Zalán Borsos [1]
Brian McWilliams [1] Victor Ungureanu [1] Olivier Bachem [1] Olivier Pietquin [2] Matthieu Geist [2]
Léonard Hussenot [1] Neil Zeghidour [3] Andrea Agostinelli [1]

Abstract

We propose `MusicRL`, the first music generation system finetuned from human feedback. Appreciation of text-to-music models is particularly subjective since the concept of musicality as well as the specific intention behind a caption are user-dependent (e.g. a caption such as "upbeat work-out music" can map to a retro guitar solo or a technopop beat). Not only this makes supervised training of such models challenging, but it also calls for integrating continuous human feedback in their post-deployment finetuning. `MusicRL` is a pretrained autoregressive MusicLM (Agostinelli et al., 2023) model of discrete audio tokens finetuned with reinforcement learning to maximize sequence-level rewards. We design reward functions related specifically to text-adherence and audio quality with the help from selected raters, and use those to finetune MusicLM into `MusicRL-R`. We deploy MusicLM to users and collect a substantial dataset comprising 300,000 pairwise preferences. Using Reinforcement Learning from Human Feedback (RLHF), we train `MusicRL-U`, the first text-to-music model that incorporates human feedback at scale. Human evaluations show that both `MusicRL-R` and `MusicRL-U` are preferred to the baseline. Ultimately, `MusicRL-RU` combines the two approaches and results in the best model according to human raters. Ablation studies shed light on the musical attributes influencing human preferences, indicating that text adherence and quality only account for a part of it. This underscores the prevalence of subjectivity in musical appreciation and calls for further involvement of human listeners in the finetuning of music generation models. Website with samples.

1. Introduction

Generative modeling of music has experienced a leap forward: while it was until recently either limited to the fine modeling of individual instruments (Engel et al., 2017; Défossez et al., 2018; Engel et al., 2020) or the coarse generation of polyphonic music (Dhariwal et al., 2020), models can now handle open-ended, high-fidelity text-controlled music generation (Forsgren & Martiros, 2022; Agostinelli et al., 2023; Liu et al., 2023; Copet et al., 2023). In particular, text-to-music systems such as MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023) build on audio language models, as they cast the generative process as an autoregressive prediction task in the discrete representation space of a neural audio codec (Zeghidour et al., 2022; Défossez et al., 2022). While this approach has demonstrated its ability to generate realistic speech (Borsos et al., 2023a; Wang et al., 2023; Borsos et al., 2023b), sound events (Kreuk et al., 2022) and music, it suffers from a few shortcomings. First, the next-token prediction task used to train these systems — while generic enough to model arbitrary audio signals — lacks any prior knowledge about musicality that could bias those towards generating music that is more appealing to listeners. Second, while the temperature sampling used at inference allows for generating diverse audio from a single text caption, this diversity is only desirable along certain axes such as melody or performance, while musicality and adherence to the prompt should remain consistently high.

These fundamental issues of autoregressive generative models have been extensively observed and addressed in the context of language modeling. For example, several works have explored finetuning machine translation models to maximize the BLEU score (Ranzato et al., 2016; Wu et al., 2016) or summarization models to improve the relevant ROUGE metric (Ranzato et al., 2016; Wu & Hu, 2018; Roit et al., 2023). Such metrics are typically sequence-level, and evaluate the output of a non-differentiable sampling process (e.g., greedy decoding, temperature sampling). This is typically circumvented by using a reinforcement learning method which models the metric of interest of a reward function and the generative model as a policy. The underlying algorithmic similarity between such text gen-

[1] Google DeepMind [2] Now at Cohere [3] Now at Kyutai. Correspondence to: Geoffrey Cideron <gcideron@google.com>, Andrea Agostinelli <agostinelli@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

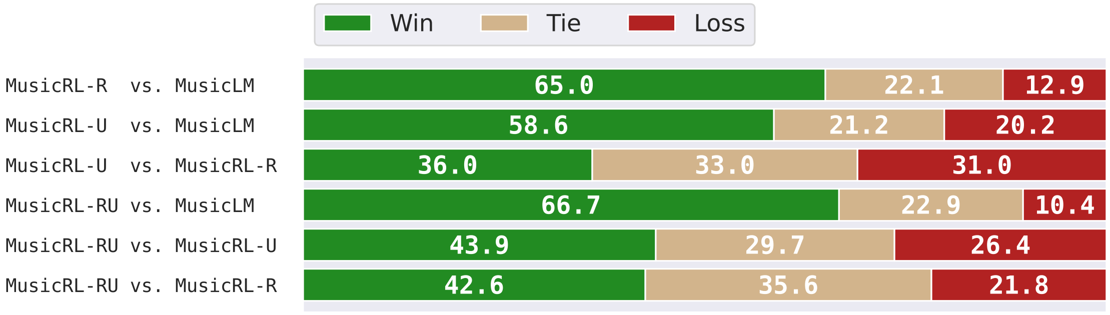


Figure 1. Results of the qualitative side-by-side evaluation for the RLHF finetuned models. In each X vs. Y comparison, the green bar corresponds to the percentage of times model X was preferred, the yellow bar to the percentage of ties and the red bar to the percentage of times model Y was preferred. `MusicRL-R` is the MusicLM model finetuned on quality and text adherence reward. `MusicRL-U` is finetuned on a reward model of user preferences. `MusicRL-RU` is finetuned sequentially on quality and adherence to text and then on a reward model of user preferences. While every RLHF finetuned version of MusicLM significantly outperforms MusicLM, `MusicRL-R` and `MusicRL-U` achieve comparable performance, while `MusicRL-RU` is overall the preferred model.

eration systems and autoregressive music models suggests that — given the proper reward functions— one could use reinforcement learning to improve music generation.

Music generated given a prompt should exhibit three properties: adherence to the input text, high acoustic quality (absence of artifacts), and "musicality" or general pleasantness. Automatic metrics have been proposed to quantify the text adherence like Classifier KLD (Yang et al., 2022) or MuLan Cycle Consistency (Agostinelli et al., 2023) as well as acoustic quality with Fréchet Audio Distance (Kilgour et al., 2019). Such metrics could be used as reward functions. Yet, designing automatic proxies to measure musicality is challenging. Most of the previous approaches (Jaques et al., 2017; Kotecha, 2018; Guimaraes et al., 2017; Latif et al., 2023) rely on complex music theory rules, are restricted to specific musical domains (e.g., classical piano) and only partially align with human preferences. This gap between automatic metrics and human preferences has again been extensively studied in language modeling, with RLHF (Reinforcement Learning from Human Preferences) becoming the *de facto* way of aligning conversational models (Achiam et al., 2023; Team et al., 2023) with human feedback.

Human preferences as referred in previous work (Ouyang et al., 2022; Stiennon et al., 2020) mainly refers to the preferences of raters. Raters may not be representative of the population interacting with the model (e.g. rating services such as Amazon Mechanical Turk[1] uses a global workforce). Especially in the context of music, this population gap can have a significant impact on the preferences (Trehub et al., 2015). Collecting large scale user preferences data allows to bridge the population gap and to collect considerably more interactions in constrast with raters.

In this work, we introduce `MusicRL`, a text-to-music generative model finetuned with reinforcement learning. Starting from a strong MusicLM baseline, we use an automatic measure of text adherence as well as a new acoustic fidelity metric as reward functions to perform RL finetuning. Human evaluations indicate that generations from the resulting `MusicRL-R` are preferred over those from MusicLM 83% of the time, as measured by $win/(win + loss)$. Secondly, to explicitly align the model with human judgment, we collect a dataset of pairwise preferences from users interacting with MusicLM to fit a reward model. Ablation studies on the reward model trained on user interaction data demonstrate that user preferences strongly correlate with musicality. Extensive human evaluations reveal that the music generations coming from the resulting `MusicRL-U` are preferred over the base model 74% of the time. Thirdly, we combine automatic rewards and human feedback to finetune `MusicRL-R` into `MusicRL-RU` and show that this models outperforms all alternatives more than 62% of the time. To the best of our knowledge, this work is the first attempt at leveraging human feedback at scale to improve an audio generative model.

2. Related Work

Music generation. While earlier approches to musical audio generation were limited in terms of producing high quality outputs (Dhariwal et al., 2020) or semantically consistent long audios (Hawthorne et al., 2022), recent research has achieved a level of quality that allows for an enjoyable listening experience. A first line of work casts the task of music generation as categorical prediction in the discrete token space provided by a neural audio codec (Zeghidour

[1] https://www.mturk.com/

et al., 2022; Défossez et al., 2022), and trains a Transformer-based (Vaswani et al., 2017) for next token prediction (Borsos et al., 2023a) or parallel token decoding (Borsos et al., 2023b; Garcia et al., 2023; Parker et al., 2024). Combining this generative backbone with text-conditioning either through a text encoder or text-audio embeddings (Elizalde et al., 2022; Huang et al., 2022) provides high-quality text-to-music models (Agostinelli et al., 2023; Copet et al., 2023). A parallel line of work relies on diffusion models and casts the task of music generation as denoising of audio waveforms and spectrograms (Huang et al., 2023) or learned latent representations (Schneider et al., 2023; Liu et al., 2023; Lam et al., 2023; Evans et al., 2024). In both cases, the models are trained offline on a collection of existing musical recordings and inference is run a stochastic fashion (e.g. diffusion or temperature sampling), which provides diversity but also uncertainty on the outputs (e.g. in terms of text-adherence or quality). Previous work (Kharitonov et al., 2023) has circumvented this issue by sampling many sequences, ranking them with a score function (e.g. a reference-free audio quality estimator) and returning the best candidate. This considerably increases inference cost and requires well-defined score functions.

MusicRL addresses these limitations by finetuning a MusicLM (Agostinelli et al., 2023) model with reinforcement learning, using reward functions derived from automatic metrics, small scale high-quality human ratings, and large scale user feedback. To the best of our knowledge, MusicRL is the first music generation system that shows the benefits from integrating feedback from hundreds of thousands of users.

RL-finetuning of music generation models. Most previous works in RL-finetuning music generation models involve designing handmade reward signals based on principles of music theory (Jaques et al., 2017; Kotecha, 2018; Guimaraes et al., 2017; Latif et al., 2023) or simple patterns like repetitions (Karbasi et al., 2021). Jaques et al. (2017) use a set of rules inspired by a melodic composition theory (Gauldin, 1988) (e.g., stay in key, play motifs and repeat them, avoid excessively repeating notes) in combination with a KL regularization term. These approaches have several limitations: the rule sets can be incomplete or contradictory, practitioners must find the correct balance between different rewards, and the rules themselves derive from music theory, which is an imperfect approximation of human musical preferences. Jiang et al. (2020) finetune an online music accompaniment generation model with four reward models learned from data and rule-based reward that assign -1 when a note is excessively repeated. Each reward model corresponds to the probability of a chunk of the generation given a context (the context and the chunk to predict is different for each reward). These rewards are learned with a masked language model (Devlin et al., 2019) loss on a music dataset. Yet, such methods only apply to restricted musical domains (e.g. monophonic piano) or symbolic generation. In contrast with previous work, MusicRL learns human preferences from its own raw audio generations. This allows for improving music generation across the whole spectrum of musical genres and styles, from lo-fi hip-hop to orchestral symphonies and modal jazz.

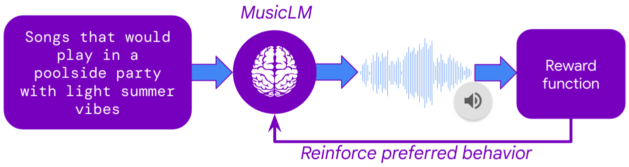

MusicLM
Songs that would play in a poolside party with light summer vibes
Reward function
Reinforce preferred behavior

Figure 2. Given a dataset of music captions, MusicLM generates audio samples that are scored with a reward function. The RL algorithm finetune the model to maximize the received reward.

RL from human feedback. RLHF recently became a critical step in the training of conversational models used in applications such as Bard (Gemini Team, 2023) or GPT-4 (OpenAI, 2023). RLHF has first been applied to solve Atari games (Christiano et al., 2017) before being used widely, for example in natural language tasks (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Jaques et al., 2019; Bai et al., 2022) or in image generation (Lee et al., 2023; Wallace et al., 2023). Wallace et al. (2023) uses Direct Optimization Algorithm (DPO) (Rafailov et al., 2023) to finetune a diffusion model on human preference data. To the best of our knowledge, we are the first to apply RLHF to music generation models.

3. Method

3.1. MusicLM

MusicLM (Agostinelli et al., 2023) is an autoregressive model for generating music from text descriptions. Following the design of AudioLM (Borsos et al., 2023a), MusicLM relies on two different types of audio representations for generation: *semantic* tokens, which are quantized representations of masked audio language models such as w2v-BERT (Chung et al., 2021) and *acoustic* tokens, the discrete representations produced by neural audio codecs such as SoundStream (Zeghidour et al., 2022). While the semantic tokens ensure the long-term structural coherence of the generation process, the acoustic tokens tokens allow for high-quality synthesis. To ensure high-bitrate reconstructions, SoundStream uses residual vector quantization (RVQ) — a stack of vector quantizers where each quantizer operates on the residual produced by the previous quantizers — to discretize the continuous audio representations, imposing a hierarchical structure on the acoustic tokens. Additionally, MusicLM relies on MuLan (Huang et al., 2022), a joint music-text contrastive model, for conditioning the audio generation task on descriptive text.

MusicLM was initially introduced as a 3-stage Transformer-based autoregressive model. The first stage learns the mapping between MuLan and semantic tokens. The second stage predicts the first levels from the output of the SoundStream RVQ (coarse acoustic tokens) from MuLan and semantic tokens. The last stage predicts the remaining SoundStream RVQ levels (fine acoustic tokens) from coarse acoustic tokens.

For the purpose of RL finetuning, we choose to optimize the semantic and coarse acoustic modeling stages, which are the most important contributors to acoustic quality, adherence to the text and overall appeal of the generated music. We address the challenges of jointly optimizing semantic and coarse acoustic modeling by using a single autoregressive stage that operates on frame-interleaved semantic and acoustic tokens. While simplifying the RL setup and problem formulation, this approach increases modeled token sequence length. We address this with a hierarchical transformer, similarly to Lee et al. (2022); Yu et al. (2023); Yang et al. (2023). Finally, instead of the original autoregressive fine acoustic modeling stage of MusicLM, we use Soundstorm (Borsos et al., 2023b) for achieving efficient parallel generation.

For simplicity, by referring to MusicLM in this work, we refer only to the autoregressive modeling stage of interleaved semantic and coarse acoustic tokens, which is the text conditioned modeling stage that can be finetuned with RL.

3.2. RL finetuning procedure

We use the standard formulation of RL in the context of finetuning large language models as done in previous work (Ziegler et al., 2019). Figure 2 illustrates the RL training loop. The agent acts according to its policy π_θ with θ the weights that parameterize the policy. The policy is an autoregressive model taking as input $a_0, \dots, a_{t-1}$, the sequence of previously generated tokens and outputs a probability distribution over the next action, i.e., the next token to pick : $a_t \sim \pi_\theta(.|a_0 \dots a_{t-1})$. The RL finetuning phase aims at maximizing $\mathbb{E}_{\pi_\theta}[\sum_t r(a_0 \dots a_t)]$ with r a given reward function. We use a KL regularized version of the REINFORCE algorithm (Williams, 1992; Jaques et al., 2017) to update the policy weights. Given a trajectory $(a_t)_{t=0}^T$ and denoting $s_t = (a_0 ... a_{t-1})$, the corresponding policy gradient objective to maximize is

$$\mathbb{J}(\theta) = (1-\alpha)[\sum_{t=0}^{T} \log \pi_\theta(a_t|s_t)(\sum_{i=t}^{T} r(s_i) - V_\phi(s_t))] \\ -\alpha \sum_{t=0}^{T} \sum_{a \in A} [\log(\pi_\theta(a|s_t)/\pi_{\theta_0}(a|s_t))],$$

with A the action space which here corresponds to the codebook, α the KL regularization strength, and V_ϕ the baseline. The baseline value function V_ϕ is used to decrease the variance in the policy gradient objective (Sutton & Barto, 2018) and it is trained to estimate the mean return of the current policy. The baseline is learned as follows:

$$\min_\phi \mathbb{E}_{\pi_\theta} \sum_t (\sum_{k=t}^{T} r(s_k) - V_\phi(s_t))^2.$$

Both the policy and the value function are initialized from the initial MusicLM checkpoint with weight θ_0.

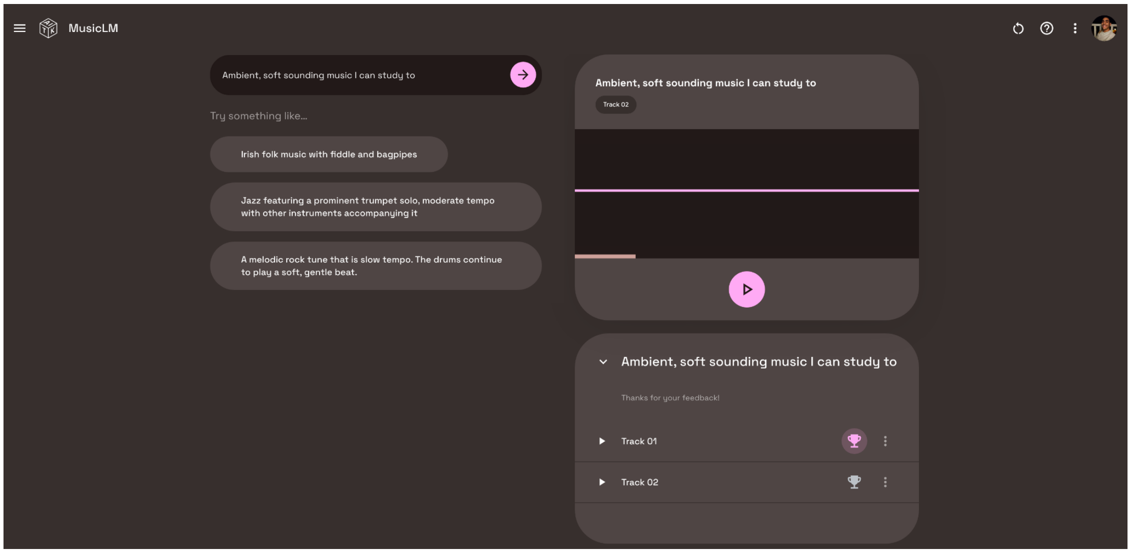
3.3. Reward Signals

Text adherence. We derive a reward model for text adherence from pretrained MuLan (Huang et al., 2022) embeddings. MuLan is a contrastive audio-text embedding model trained on music clips and weakly-associated, free-form text annotations. We compute the cosine similarity between the text embedding of the input prompt and the audio embedding of the generated music, resulting in a reward value in $[-1; 1]$. We refer to this metric as MuLan score. Because our models generate 30-second audios, while MuLan is trained on 10-second audio clips, we divide each audio into three segments, we calculate MuLan scores for each segment, and we average the results.

Acoustic quality. Another main attribute of musical generation is acoustic quality, e.g. whether a clip sounds like a professional recording or is contaminated with artifacts. We rely on a reference-free quality estimator trained to predict the human Mean Opinion Score (MOS - between 1 and 5) of a 20 second music clip. The model architecture is a Conformer (Gulati et al., 2020). We train the model on a dataset of $\approx$ 5000 samples that are a mix of human-created and MusicLM-generated music clips, where each clip was rated by 3 raters. The raters were tasked to judge only the acoustic quality, to avoid confounding factors such as musicality. We refer to this metric as the quality score. Because our models generate 30-second clips, we compute quality scores on the first 20 seconds and on the last 20 seconds, and average the two scores.

User preferences. We deploy the pretrained text-to-music MusicLM model through the AITK web-based interface[2] (Figure 3) to a large scale userbase. We choose to collect feedback through pairwise (Christiano et al., 2017): when a user seizes a prompt, we generate two 20s candidate clips and let the user optionally assign a trophy to one of them. An important design choice implied by this process is the absence of specific instructions, which is intended not to bias users towards precise musical attributes and rather communicate their overall subjective taste. We only con-

[2] https://aitestkitchen.withgoogle.com/



Ambient, soft sounding music I can study to

Irish folk music with fiddle and bagpipes

Jazz featuring a prominent trumpet solo, moderate tempo with other instruments accompanying it

A melodic rock tune that is slow tempo. The drums continue to play a soft, gentle beat.

Ambient, soft sounding music I can study to

Ambient, soft sounding music I can study to

Figure 3. The AI Test Kitchen MusicLM interface. The user can write a prompt or choose from suggestions. Each prompt generates two 20s clips, and the user can label their favorite clip among the two with a trophy.

sider preferences from users that listen to both generations. After filtering, we obtain a dataset of pairwise user data of size 300,000. This dataset minimizes the biases that often arise from human raters (as detailed in Appendix D).

Our reward model takes as input the caption's text and corresponding audio tokens and outputs a scalar score. This model is trained with a Bradley-Terry Model (Bradley & Terry, 1952) as in Christiano et al. (2017), which enables learning a pointwise ELO score from pairwise preferences. It is initialized with the MusicLM checkpoint, as first results demonstrated that, starting from scratch, the reward model was not able to do better than chance at predicting human preferences. We split the user preference dataset into a train split of size 285,000 and an evaluation split of size 15,000. After training for 10,000 steps on batches of 32 pairs, the reward model achieves 60% of accuracy on the evaluation set (see Figure 6).

To pre-assess the performance of the reward model, we conduct an internal small-scale human evaluation on 156 audio comparisons from the user preference dataset. In 60% of cases, our team's preferences aligned with the established preferences in the dataset. This result is comparable to the performance of the reward model. Furthermore, this low agreement rate highlights the inherent subjectivity in judging music preferences, compared to domains such as summarization where Stiennon et al. (2020) estimated at 73-77% the agreement rate for the OpenAI human preference dataset. When finetuning MusicLM on the user preference reward model, since our models generate 30-second audios, we average the scores computed from the first and last 20 seconds of audio.

4. Experimental Setup

4.1. Datasets

Given the pretrained reward signals as described in Section 3.3, the RL finetuning step uses a dataset exclusively composed of captions, used for prompting all MusicLM-based models. Consequently, no ground-truth audio is involved in the finetuning process. We follow the same procedure as Huang et al. (2023) for synthetically generating captions from three sources. We use the LaMDA model (Thoppilan et al., 2022) to generate descriptions of 150,000 popular songs. After providing song titles and artists, LaMDA's responses are processed into 4 million descriptive sentences about the music. We split 10,028 captions from MusicCaps (Agostinelli et al., 2023) into 35,333 single sentences describing music. Furthermore, we collect 23,906 short-form music tags from MusicCaps. Additionally, we extend the previous captions with the 300,000 prompts collected from users, as described in Section 3.3. We randomly split the data, using 90% for training and 10% for evaluation.

4.2. Training procedure

In the following experiments, we RL-finetune the MusicLM model with the same RL algorithm and the same hyperparameters. The common decoding scheme is temperature sampling with temperature $T = 0.99$. The temperature was chosen with subjective inspection to have a good quality-diversity tradeoff for the generations of MusicLM. The RL-finetuned models differs only with the reward function employed during their training process.

`MusicRL-R`. We RL-finetune MusicLM for 20,000 training steps (1) with the MuLan reward, (2) with the quality reward, and (3) with a linear combination of the MuLan and the quality reward: the resulting models are respectively called `MusicRL-MuLan`, `MusicRL-Quality`, and `MusicRL-R`. Throughout our experiments, we normalize the quality reward from $[1; 5]$ to $[0; 1]$ as preliminary experiments have shown that the combination of the MuLan and the quality reward gives the best results when both rewards are on the same scale. We still display in figures the un-normalized scores.

`MusicRL-U`. We RL-finetune MusicLM for 5000 training steps with the user preference reward model to obtain a model that we call `MusicRL-U`.

`MusicRL-RU`. To combine all the reward signals, we RL-finetune `MusicRL-R` for 1000 training steps on the user preference reward model. For this experiment, the KL regularization is computed between the model being finetuned and `MusicRL-R`. The resulting model is called `MusicRL-RU`. We find that the sequential approach of first finetuning on MuLan and quality and then finetuning on the user preference reward outperforms learning from the three rewards at the same time. We hypothesize this comes from the fact that it takes a small number of gradient steps (under 2000) before over optimizing on the user preference reward while it takes around 10,000 steps to optimize the other rewards. Moreover, using the user preference reward model in a final stage in this matter may allow the model to align better on the human preferences.

4.3. Evaluation

The main metrics we report in our experiments are the quality reward, the MuLan reward, and the user preference reward model. We report the metrics either against the training step to show progress along the training, or against the KL divergence to the base model. This is typically used as a proxy to measure the distance to the base checkpoint and thus the retention of the original capabilities of the model (Christiano et al., 2017; Roit et al., 2023).

For the qualitative evaluation, we use 101 diverse, internally-collected prompts, representing a balanced range of musical genres (see Appendix A for the full list). We use these prompts to generate audio samples from each evaluated model. We select raters for their experience listening to varied musical styles (>6 years) and fluency in written English. During the qualitative evaluation, raters are presented with two audio clips generated by different models using the same text prompt. We ask raters to rate each clip on a scale of 1 to 5, considering adherence to the text prompt, acoustic quality and overall appeal to the audio clip. Each comparison is performed by three different raters, totaling 303 ratings per model comparison. From these ratings, we compute a win rate metric which is defined as $win/(win + loss)$.

4.4. Checkpoint selection

For all RL-finetuned models, we manually select the best checkpoint by inspecting the quantitative results and listening to the music generations. For `MusicRL-R`, `MusicRL-U`, and `MusicRL-RU` we respectively choose the checkpoint after 10,000 training steps, 2000 training steps, and 1000 training steps.

5. Results

We aim to answer the following questions: (1) Can RL-finetuning on MuLan and quality rewards improve the generation quality of text-to-music models such as MusicLM? (2) Can RLHF improve the alignment of the generated music to generic preferences from users? (3) Is it possible to combine all reward signals to further improve performance?

5.1. Quantitative Results

In all quantitative evaluations, we analyze model progress during RL finetuning by tracking scores of rewards against the KL divergence from the initial model. Regardless of whether we train with a single reward model or a combination of both as in `MusicRL-R`, we evaluate model performance on all reward signals.

Figure 4 shows that RL-finetuning successfully optimizes both quality and MuLan scores. Specifically, finetuning on the quality reward alone leads to the greatest increase in quality score (from 3.5 MOS to 4.6 MOS), and a smaller increase in the MuLan score (from 0.58 to 0.61). Conversely, finetuning on only the MuLan reward maximizes the MuLan score (from 0.58 to 0.71), with a less pronounced quality score improvement (from 3.5 MOS to 4.1 MOS). Leveraging both quality and MuLan rewards significantly improves both scores (quality: 3.5 MOS to 4.4 MOS; MuLan: 0.58 to 0.71), while marginally increasing KL divergence. Given the promising and stable performance in simultaneously optimizing MuLan and quality scores, we perform qualitative evaluations only on `MusicRL-R`.

Figure 8 (in Appendix B) shows that after 10,000 finetuning steps on the quality reward, the reward model trained on user preference begins assigning lower scores to music samples. This suggests that finetuning solely on the quality reward is prone to reward over-optimization (Coste et al., 2023; Ramé et al., 2024; Jiang et al., 2020).

Figure 5 demonstrates that finetuning with the user preference reward model significantly improves generation scores, increasing them from -1.5 to over 1.5. Figure 4 shows that despite not training on the quality reward, the quality score increases from 3.5 MOS to 4 MOS. The MuLan score

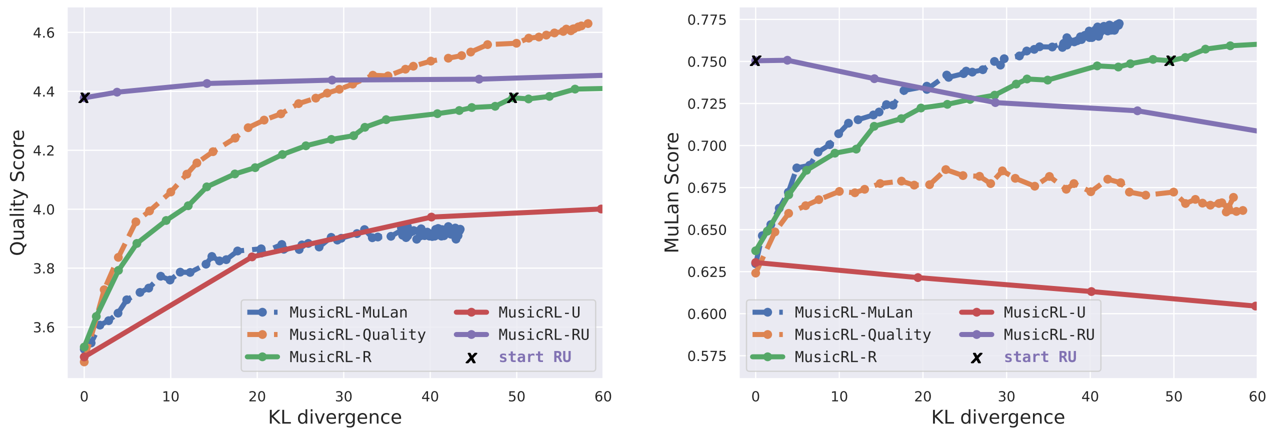


Figure 4. Quality (left) or MuLan score (right) vs KL divergence for the RL-finetuned models. The KL divergence is computed between the RL-finetuned models and MusicLM except for `MusicRL-RU` where the KL divergence is computed against `MusicRL-R`. The black cross corresponds to the checkpoint used to start the training of `MusicRL-RU`. RL-finetuning successfully optimizes the quality and the MuLan scores (`MusicRL-R`). Additionally, optimizing the user preference reward (`MusicRL-RU`, `MusicRL-RU`) improves the quality score while marginally decreasing the MuLan score.

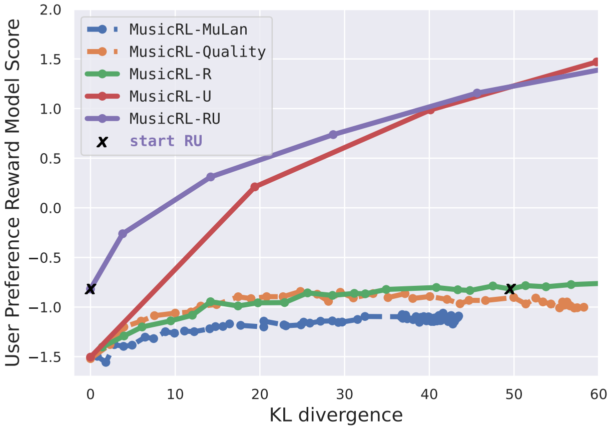


Figure 5. User Preference Reward Model Score for the different RL-finetuned models. The KL divergence is computed between the RL-finetuned models and MusicLM except for `MusicRL-RU` where the KL divergence is computed against `MusicRL-R`. The black cross corresponds to the checkpoint used to start the training of `MusicRL-RU`. RL-finetuning successfully improves the user preference reward model score of the generations (see `MusicRL-U` and `MusicRL-RU` curves). When trained on other rewards (MuLan and/or quality) the user preference reward model score slightly improves.

slightly decreases from 0.58 to 0.55. Yet, Figure 7 highlights that over-optimizing the user preference reward model can drastically reduce the MuLan score. Overall, this suggests that user preference feedback particularly enhances audio quality while having minimal impact on text adherence.

Model	MOS	# wins
MusicLM	3.07	133
`MusicRL-R`	3.54	362
`MusicRL-U`	3.54	372
`MusicRL-RU`	3.82	460

Table 1. Average mean opinion score (MOS) and number of wins across all rating tasks, for each model. The music generated from the RL-finetuned models are significantly scored higher in average than the ones from MusicLM. The best performing model both in term of MOS and number of wins is `MusicRL-RU`.

Figure 5 shows that optimizing the user preference reward model on a model finetuned for 10,000 steps on quality and MuLan improves the user preference reward model score significantly. Figure 4 shows that the quality score slightly increases while the MuLan score slightly decreases, which confirms the impact of the user preference reward model observed in the previous paragraph.

5.2. Qualitative Results

Figure 1 presents human rater evaluations of pairwise comparisons between all possible model combinations across MusicLM, `MusicRL-R`, `MusicRL-U` and `MusicRL-RU`. When compared to MusicLM, `MusicRL-R` wins 65% of the time, ties 22.1% and loses 12.9%. This translates into a 83% win rate in favor of `MusicRL-R`. `MusicRL-U` is also strongly preferred over MusicLM as it achieves a 74% win rate against MusicLM. The best performing model overall is `MusicRL-RU`. When compared to MusicLM, MusicRL-RU is strongly preferred by the raters with a 87% win rate. When compared to the other RL-finetuned models, MusicRL-RU achieves a win rate of 66% against MusicRL-R, and 62% against MusicRL-U. All results described above are statistically significant according to a post-hoc analysis using the Wilcoxon signed-rank test (Rey & Neuhäuser, 2011).

Table 1 summarizes results from all qualitative evaluations by showing average mean opinion score (MOS) and number of wins across all rating tasks, for each model. On both metrics, all RL-finetuned models outperform MusicLM, with MusicRL-RU being the best performing model.

Lastly, MusicRL-R and MusicRL-U perform comparably according to raters, as shown from Figure 1 and Table 1.

5.3. Takeaway

Our results demonstrate several key findings: (1) MusicRL-R shows that RL-finetuning on text adherence and quality rewards improves the generation quality of MusicLM; (2) MusicRL-U confirms the ability to leverage generic user preferences data to improve MusicLM; (3) MusicRL-RU outperforms all other models, demonstrating that the above reward signals are complementary and can be constructively combined for the highest performance.

6. Understanding Human Feedback Through the Lens of the Reward Model

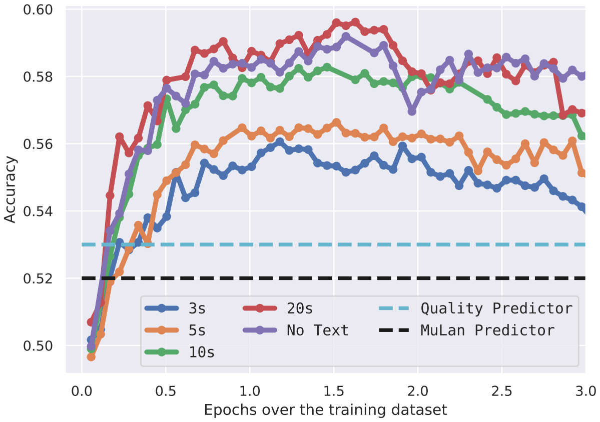


Figure 6. Ablations on the user preference reward model. The reward model is learned either with no text tokens (No Text) or with a cropped version of the input audio (i.e. 10s, 5s, 3s). While dropping the text tokens does not significantly impact the accuracy of the reward model, cropping the audio substantially degrades performance. This suggests that text adherence and audio quality are not the primary factors influencing user audio preferences, as additionally shown by the low accuracy when using text adherence based (MuLan) or audio quality based predictors for user preference.

In this section, we analyze reward model accuracy to uncover the specific music elements that influence user preference. This analysis directly addresses our research question: What is the user paying attention to when rating the audio?

We categorize generated music into three components which might drive the users choice on their audio preferences: (1) text adherence, (2) audio quality, and (3) musicality. In particular, defining and modeling musicality is a complex task, which underscores our focus on human feedback as a solution, moving beyond rule-based limitations.

6.1. Importance of the text input

To isolate the impact of text on pairwise preference evaluation, we drop text tokens while training the reward model. Accuracy remains stable as shown by Figure 6. Additionally, we measure how often the clip with the highest MuLan score corresponds to the preferred one. On the evaluation set, these indicators only match 51.6% of the time, which is very close to random accuracy. Overall, these findings indicate that adherence to the text prompt was not a primary driver of human preference in our experiment. This aligns with our quantitative results in Section 5.1, which show no significant improvement in text adherence as measured by MuLan, when training MusicRL-U.

6.2. Importance of the audio quality

Since audio quality remains relatively consistent within a generated sequence, a few seconds of audio should provide sufficient information to evaluate this aspect. We train reward models on different input audio tokens length corresponding to 10, 5, and 3 seconds. As shown in Figure 6 the evaluation accuracy on pairwise preference decreases as we reduce the length of the input tokens, dropping from 60 to 56% when using 3-5 seconds of input audio. The significant accuracy decrease suggests that other musical components play a complementary role in user preference. Additionally, we replicate the analysis done in 6.1 and measure how often the clip with the highest quality score is preferred. As shown in Figure 6 the quality predictor achieves 53.3% accuracy on the evaluation dataset. These findings indicate that audio quality is not the only driver of human preference, while being a better signal than text adherence. This is consistent with our quantitative results in Section 5.1, where training MusicRL-U improves marginally on the quality score.

Overall, this analysis shows that user preference is influenced by music elements which go beyond text adherence and audio quality.

7. Limitations and Future Work

Aligning feedback and evaluation. When training on user preference data, a limitation of our current setup is the *population gap* between those who provide feedback to improve the model (general users) and those who assess the results (selected raters). A direction for future work is to directly measure the perceived improvements from the user's perspective.

Using on-policy data. For the reasons explained in Section 3.1, in this work we collected user preferences on a different version of MusicLM compared to the one used for RL finetuning. A clear path for improvement is to iteratively collect on-policy data (data generated by the model that is being finetuned) and use it to update the model. Eventually, this would allow for real integrated feedback where finetuned models are continuously deployed to collect new feedback while improving the user experience.

Refining the user preference dataset. Several interesting research directions involve refining the large user interaction dataset. For instance identifying and retaining examples where users express a confident and clear preference could reduce noise and improve the overall dataset quality. Furthermore, focusing on techniques to train robust reward models on smaller, but highly relevant datasets could facilitate research directions such as model personalization for specific users.

8. Conclusion

In this work, we introduce MusicRL, the first text-to-music generative model aligned with human preferences. In a first set of experiments, we derive sequence-level reward functions that inform on the adherence to the text and acoustic quality, and we finetune a pretrained MusicLMmodel to optimize these rewards with RL. The quantitative and qualitative results show consistent improvements over the pretrained baseline. We then show for the first time that we can align music generation with generic preferences from users. We collect 300,000 user generated captions and audios through a web interface to create a model of the user preferences. This allows improving our model through RLHF, again consistently outperforming the baseline. Lastly, we combine all reward signals to produce the highest performing model. Additional analysis indicates that the signal extracted from user preferences contains information beyond text adherence and audio quality. This highlights the subjective and complex nature of musical appeal, emphasizing the value of integrating user feedback when improving music generation models.

Impact Statement

A considerable concern in our research is the potential for biases from users to be integrated into music generation models. User biases can perpetuate through: 1) Unconscious bias, where users may have biased opinion on what constitutes good music; 2) Explicit bias, where some users might provide feedback driven by explicit biases related to race, gender, culture, or other factors; 3) Majority bias, where a particular demographic dominates the user feedback potentially marginalizing other musical styles and perspectives. Overall such biases might limit model creativity, discriminate certain groups and generally provide a negative user experience. Mitigation strategies include data awareness, actively seeking user feedback from diverse demographics and develop tools for detecting and correcting such biases.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C. Musiclm: Generating music from text, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023a.

Borsos, Z., Sharifi, M., Vincent, D., Kharitonov, E., Zeghidour, N., and Tagliasacchi, M. Soundstorm: Efficient parallel audio generation, 2023b.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39, 1952.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Chung, Y., Zhang, Y., Han, W., Chiu, C., Qin, J., Pang, R., and Wu, Y. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *arXiv:2108.06209*, 2021.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation, 2023.

Coste, T., Anwar, U., Kirk, R., and Krueger, D. Reward model ensembles help mitigate overoptimization. *arXiv preprint*, 2023.

Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., and Bach, F. R. SING: symbol-to-instrument neural generator. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9055–9065, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/56dc0997d871e9177069bb472574eb29-Abstract.html.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv:2210.13438*, 2022.

Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. Clap: Learning audio concepts from natural language supervision, 2022.

Engel, J. H., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. Neural audio synthesis of musical notes with wavenet autoencoders. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1068–1077. PMLR, 2017. URL http://proceedings.mlr.press/v70/engel17a.html.

Engel, J. H., Hantrakul, L., Gu, C., and Roberts, A. DDSP: differentiable digital signal processing. In *International Conference on Learning Representations (ICLR)*, 2020.

Evans, Z., Carr, C., Taylor, J., Hawley, S. H., and Pons, J. Fast timing-conditioned latent audio diffusion. *CoRR*, abs/2402.04825, 2024. doi: 10.48550/ARXIV.2402.04825. URL https://doi.org/10.48550/arXiv.2402.04825.

Forsgren, S. and Martiros, H. Riffusion - Stable diffusion for real-time music generation, 2022. URL https://riffusion.com/about.

Garcia, H. F., Seetharaman, P., Kumar, R., and Pardo, B. Vampnet: Music generation via masked acoustic token modeling, 2023.

Gauldin, R. *A practical approach to eighteenth-century counterpoint*. Prentice-Hall, 1988.

Gemini Team, G. Gemini: A family of highly capable multimodal models. 2023.

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition. In Meng, H., Xu, B., and Zheng, T. F. (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 5036–5040. ISCA, 2020. doi: 10.21437/INTERSPEECH.2020-3015. URL https://doi.org/10.21437/Interspeech.2020-3015.

Hawthorne, C., Jaegle, A., Cangea, C., Borgeaud, S., Nash, C., Malinowski, M., Dieleman, S., Vinyals, O., Botvinick, M. M., Simon, I., Sheahan, H., Zeghidour, N., Alayrac, J., Carreira, J., and Engel, J. H. General-purpose, long-context autoregressive modeling with perceiver AR. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning (ICML)*, 2022.

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. W. Mulan: A joint embedding of music audio and natural language. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023.

Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E., and Eck, D. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In *International Conference on Machine Learning*, pp. 1645–1654. PMLR, 2017.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Jiang, N., Jin, S., Duan, Z., and Zhang, C. Rl-duet: Online music accompaniment generation using deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 710–718, 2020.

Karbasi, S. M., Haug, H. S., Kvalsund, M.-K., Krzyzaniak, M. J., and Tørresen, J. A generative model for creating musical rhythms with deep reinforcement learning. 2nd Conference on AI Music Creativity, 2021.

Kharitonov, E., Vincent, D., Borsos, Z., Marinier, R., Girgin, S., Pietquin, O., Sharifi, M., Tagliasacchi, M., and Zeghidour, N. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718, 2023. URL https://api.semanticscholar.org/CorpusID:256627687.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, 2019.

Kotecha, N. Bach2bach: generating music using a deep reinforcement learning approach. *arXiv preprint arXiv:1812.01060*, 2018.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audiogen: Textually guided audio generation, 2022.

Lam, M. W. Y., Tian, Q., Li, T., Yin, Z., Feng, S., Tu, M., Ji, Y., Xia, R., Ma, M., Song, X., Chen, J., Wang, Y., and Wang, Y. Efficient neural music generation. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/38b23e2328096520e9c889ae03e372c9-Abstract-Conference.html.

Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H. S., and Cambria, E. A survey on deep reinforcement learning for audio-based applications. *Artificial Intelligence Review*, 56(3):2193–2240, 2023.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Lewkowicz, D. J. The concept of ecological validity: What are its limitations and is it bad to be invalid? *Infancy*, 2 (4):437–450, 2001.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D. P., Wang, W., and Plumbley, M. D. Audioldm: Text-to-audio generation with latent diffusion models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

OpenAI. Gpt-4 technical report. 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Parker, J. D., Spijkervet, J., Kosta, K., Yesiler, F., Kuznetsov, B., Wang, J.-C., Avent, M., Chen, J., and Le, D. Stemgen: A music generation model that listens, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Ramé, A., Vieillard, N., Hussenot, L., Dadashi, R., Cideron, G., Bachem, O., and Ferret, J. Warm: On the benefits of weight averaged reward models, 2024.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

Rey, D. and Neuhäuser, M. Wilcoxon-signed-rank test. In Lovric, M. (ed.), *International Encyclopedia of Statistical Science*, pp. 1658–1659. Springer, 2011. doi: 10.1007/978-3-642-04898-2_616. URL https://doi.org/10.1007/978-3-642-04898-2_616.

Roit, P., Ferret, J., Shani, L., Aharoni, R., Cideron, G., Dadashi, R., Geist, M., Girgin, S., Hussenot, L., Keller, O., Momchev, N., Garea, S. R., Stanczyk, P., Vieillard, N., Bachem, O., Elidan, G., Hassidim, A., Pietquin, O., and Szpektor, I. Factually consistent summarization via reinforcement learning with textual entailment feedback. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 6252–6272. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.344. URL https://doi.org/10.18653/v1/2023.acl-long.344.

Schneider, F., Kamal, O., Jin, Z., and Schölkopf, B. Moûsai: Text-to-music generation with long-context latent diffusion, 2023.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4603–4611. PMLR, 2018. URL http://proceedings.mlr.press/v80/shazeer18a.html.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tervaniemi, M. The neuroscience of music–towards ecological validity. *Trends in Neurosciences*, 2023.

Thomas, J. C. and Kellogg, W. A. Minimizing ecological gaps in interface design. *IEEE Software*, 6(1):78–86, 1989.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Trehub, S. E., Becker, J., and Morley, I. Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370 (1664):20140096, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 2017.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111, 2023. doi: 10.48550/ARXIV.2301.02111. URL https://doi.org/10.48550/arXiv.2301.02111.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Wu, Y. and Hu, B. Learning to extract coherent summary via deep reinforcement learning. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5602–5609. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11987. URL https://doi.org/10.1609/aaai.v32i1.11987.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv:2207.09983*, 2022.

Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Wu, X., Zhao, Z., Watanabe, S., and Meng, H. Uniaudio: An audio foundation model toward universal audio generation, 2023.

Yu, L., Simig, D., Flaherty, C., Aghajanyan, A., Zettlemoyer, L., and Lewis, M. Megabyte: Predicting million-byte sequences with multiscale transformers. *arXiv preprint arXiv:2305.07185*, 2023.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30, 2022.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Qualitative Evaluation.

For the qualitative evaluation, the list of the 101 diverse prompts is the following:

’A mid-tempo country chorus with a choir counter melody ’,
’grunge with a drum n bass beat’,
’An eerie, powerful slow metal guitar riff with drums backing that builds tension and anticipation.’,
’A wistful, nostalgic indie folk-pop song with a strong bass and a deep male voice’,
’Reggeaton with deep bass and a rapping voice’,
’A modern, romantic slow waltz played by a jazz trio’,
’A rock-steady intro with trumpets providing the backing to a gentle guitar’,
’a funky disco song with a bass player’,
’A slow rumba composition with a female voice supported by a piano a percussions’,
’A sad pop melody with piano and strings accompaniment’,
’Chinese music instruments in futuristic theme, fast pace’,
’A frantic drum machine beat and pew-pew laser noises fill the cavernous warehouse rave’,
’fast, classical music with a church organ with an eerie feeling, for a dark thriller soundtrack’,
’A fast, energetic tango played by an accordion and a violin’,
’A mellow british-rock acoustic chorus’,
’Repetitive house music with strong percussive line’,
”The sitar’s slow, meandering melody was accompanied by the tabla’s steady beat, creating a sound that was both calming and enchanting.”,
’Energetic punk rock with a female voice singing’,
’a cheerful children song with a simple xylophone backing’,
’An energetic gospel choir performance’,
’slow, mellow, and instrumental new age music for meditation.’,
’Flamenco performance full of energy’,
’Melodic danceable brazilian music with percussions.’,
’An indie-rock chorus is played by a male singer with a small band backing.’,
’epic movie soundtrack’,
”The K-pop group’s powerful vocals were accompanied by a lush string arrangement, creating a truly epic soundscape.”,
’A funk bass intro with a guitar playing short chords and a drums backing’,
’Salsa music played by an orchestra’,
’A small band plays a latin danceable song’,
’A whistling tune for a western duel soundtrack’,
’A samba beat and a lively chorus combine to create a festive atmosphere.’,
’A jazzy pop song played by a big band’,
’a ska-punk trumpet riff supported by an up-beat guitar’,
’male bass low grave voice male-singing a medieval song with a mandolin’,
’a fast symphonic metal guitar solo with a choir backing’,
’chorus of a sweet acoustic rock ballad’,
’A bluesy piano riff drives the band as they belt out a soulful tune.’,
’A slow, swing pop song with piano and drums backing’,
’A fusion of reggaeton and electronic dance music, with a spacey, otherworldly sound.’,
’A marching band plays a catchy tune’,
’A classical orchestral waltz for a costume dance’,
’Irish folk chorus with a mandolin and team whistle’,
’A male voice sings a pop anthem accompanied by his piano’,
’A catchy pop tune is sung on top a dance drumbeat’,
”The soprano’s voice soared over the delicate accompaniment of the piano, filling the opera house with beauty and emotion.”,
’Rap song with a female melodic line’,
’a reggae song with guitar and singing’,
’A corny pop chorus sung by a female voice with a lot of autotune’,
”The marimba’s soulful melody was accompanied by the steady beat of the drums, creating a bluesy sound that was both melancholy and uplifting.”,

’A gospel choir sings on top a metal guitar backing’,
’A powerful female voice sings with soul and energy over a driving drum beat.’,
’A repetitive lullaby sung by a female voice with a carillon backing’,
’Traditional fast song played by a male voice with an accordion backing’,
’An up-beat reggae with a deep male voice and a piano striking the chords’,
’Slow, melodic music backed by a sitar and strings.’,
’Funky piece with a strong, danceable beat, a prominent bassline and a keyboard melody.’,
”A danceable, fast and cheerful swing tune from the 50’s”,
’a professional solo cellist playing a sad melody for solo cello on the cello, high quality recording’,
’A rock guitar riff, a slide guitar solo and a flute melody create a lively, upbeat sound.’,
’an a cappella chorus singing a christmas song’,
’nice ragtime guitar chord progression’,
”A cheerful R’n’B song is played by two singers with a trumpet melody”,
’A dance song with a fast melody taken from sampled voice, giving the impression of percussions’,
’a gospel song with a female lead singer’,
’a nostalgic tune played by accordion band’,
’A mariachi song with an epic twist and symphonic orchestra backing’,
’A middle-easter tune with percussions and flutes’,
’Jazz composition for piano and trumpet’,
’A slow blues intro with a harmonica and minimal backing.’,
’The experimental modular synthesizer created a unique soundscape by combining the sounds of water with electronic music.’,
’a cheerful ragtime with guitar’,
’Industrial techno sounds, with hypnotic rhythms. Strings playing a repetitive melody creates an unsettling atmosphere.’,
’The microphone picked up the soulful, funky scream of the lead singer as he reached the climax of the song.’,
’The snare drum and lute played a lively duet, with the snare drum providing a steady beat and the lute playing a melody on top.’,
’The two rappers traded verses over a pulsating synth beat, creating a sound that was both energetic and infectious.’,
’A bagpipe is playing an aggressive tune with a punk backing’,
’A string quartet plays a lively tune.’,
’A very fast piano cadenza that is hard to play.’,
’A lone harmonica plays a haunting melody over the sound of the wind blowing through the desert.’,
’An aggressive, but sad punk verse, with a prominent slow guitar melody and dark bass line.’,
’a band playing cumbia in a boat along the magdalena river in colombia’,
’A slow jamaican ska song with an organ backing’,
’The gramophone needle crackled and hissed as it spun across the vinyl record, filling the room with a warm, nostalgic sound.’,
’fast piano toccata’,
”Romantic R’n’B song with a warm female voice”,
’A cheerful bollywood-style group dance’,
’Dance music with a melodic synth line and arpeggiation’,
’The wooden bongo drums beat a deep, resonating bass as the dancers move their bodies to the music.’,
’a tenor singing with a backing guitar’,
’Slow trap song with a lot of reverb and autotune’,
’A syncopated progressive rock tune with a saxophone ’,
’A syncopated drum beat backs a hard rock guitar riff’,
’a gregorian chant’,
’A danceable folk waltz is played by an accordion’,
’A bagpipe is playing a fast catchy tune in a dance-pop song’,
’A full orchestra playing a violin concerto from the 1800s’,
’The trap beat was layered with acoustic string sounds creating a catchy chorus.’,
’A church choir sings a high-pitched, soothing melody.’,
’An energetic dance-pop song, sang by a powerful female voice’,
’An harmonica plays a melancholic solo over an acoustic guitar’,
’A fast rap chorus is sung on top of a simple catchy tune’

B. Additional Quantitative Evaluation Plots

Figure 7 and Figure 8 show the progress of the RL-finetuned models along training as measured by the three reward signals.

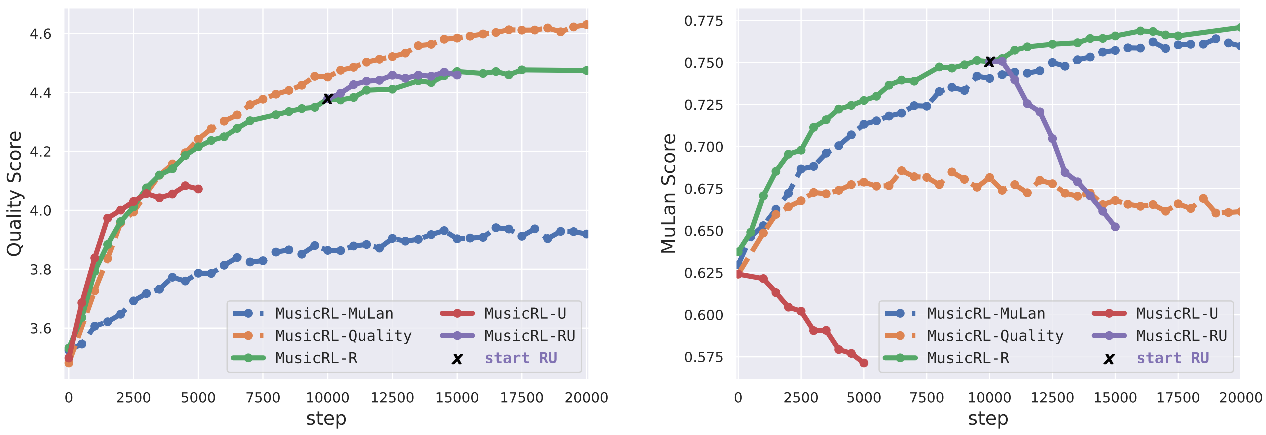


Figure 7. Quality (left) or MuLan score (right) vs step for the RL-finetuned models. The black cross corresponds to the checkpoint used to start the training of `MusicRL-RU`. RL-finetuning successfully optimizes the quality and the MuLan scores (`MusicRL-R`). Additionally, optimizing the user preference reward (`MusicRL-RU`, `MusicRL-RU`) improves the quality score while the MuLan score starts to be significantly impacted when the model over optimize the user preference reward.

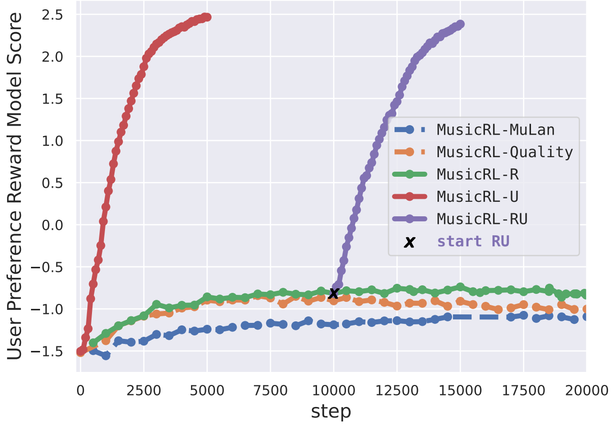


Figure 8. User Preference Reward Model Score for the different RL-finetuned models. The black cross corresponds to the checkpoint used to start the training of `MusicRL-RU`. RL-finetuning successfully improves the user preference reward model score of the generations (see `MusicRL-U` and `MusicRL-RU` curves). When trained on other rewards (MuLan and/or quality) the user preference reward model score slightly improves.

C. Implementation Details

RL-finetuning. For the RL-finetuning, we use a KL regularization strength of 0.001, a policy learning rate of 0.00001, a value learning rate of 0.0001, and 128 TPU cores of Cloud TPU v5e.

Reward Modeling. For the training of the user preference reward, we use a learning rate of 0.0001 and 32 TPU cores of Cloud TPU v4.

For both training, we use Adafactor (Shazeer & Stern, 2018) for the optimizer.

D. Advantages of User Data

In Section 5, we show that we could leverage a model that was trained with human rater data to improve a music generation model with RL. However, rater data have some limitations and biases.

In behavioural sciences, the *ecological validity*[3] of a lab study refers its potential of generalization to the real world (Lewkowicz, 2001). In the context of music, it is crucial to experiment on real-world settings (Tervaniemi, 2023). Thomas & Kellogg (1989) explore the *ecological validity* concept in the context of interface design and say that ”User-related ecological gaps are caused by characteristics of users - such as what motivates them, their cognitive abilities, preferences, and habits - that may vary between the lab and the target environment.” The concept of user-related ecological gaps is particularly relevant for the finetuning and the evaluation of large language models as the raters and users are often dissimilar.

Population Gap. Raters are often not representative of the user population especially as the rating task is often outsourced to crowdsourcing services which employ people in different countries than the ones the model is deployed in e.g. Amazon Mechanical Turk[4] proposes a global workforce for rating tasks. This population difference creates biases such as cultural biases which can impact the music preferences (Trehub et al., 2015).

Motivation Gap. As mentioned in Thomas & Kellogg (1989), the *motivation gap* which corresponds to the difference of motivations between the different users can have a significant effect on the results. In our context, while the users of music generation models have a genuine interest in playing with the model, the incentive of the raters are very different. Hence, for rating tasks, it is crucial to give specific set of instructions to make sure the raters make their decisions aligned with what the creator of the rating task would expect which also can be a source of biases. Whereas for users, we are interested in general interactions where no instructions are given.

Dataset Size. Due to the cost of rater data, the number of collected human preference is often below 100,000 (Ziegler et al., 2019; Lee et al., 2023; Stiennon et al., 2020). On the other hand, the number of user interactions can be orders of magnitude higher once a model is deployed.

[3] https://en.wikipedia.org/wiki/Ecological_validity

[4] https://www.mturk.com/