Leveraging Human Preferences to Master Poetry

Rafael Pardinas*, **Gabriel Huang, David Vazquez, Alexandre Piché**

ServiceNow Research

Abstract

Large language models have been fine-tuned to learn poetry via supervised learning on a dataset containing relevant examples. However, those models do not generate good-quality output that respects the structure expected for a specific poem type. For instance, generated haikus may contain toxic language, be off-topic, incoherent, and not respect the typical 5-7-5 syllable meter. In this work, we investigate if it is possible to learn an objective function to quantify the quality of haiku—from human feedback—and if this reward function can be used to improve haiku generation using reinforcement learning.

Introduction

Haiku is an ancient form of Japanese poetry originating from the 17th century which gained popularity in the last few decades due to its structural characteristics. They are short and abstract and often display satirical and fun attributes. Their popularity, length, and structure has put them in a prominent position as a project subject in the machine learning community, as we can see in Aguiar and Liao (2019) and Marzano (2021). The majority of them focus on optimally generating creative samples. Most of these projects are trained for next token prediction on haiku datasets. However, while these models achieve excellent results in the semantic and structural aspects of the output, they fail to achieve high-quality output, as we can see in Lewis, Zugarini, and Alonso (2021) and Ormazabal et al. (2022). Furthermore, due to haiku's abstract nature and semantic complexity, augmenting language models generation with ranking, based on simple metrics might not improve results.

In this work, we investigate if recent developments in Reinforcement Learning from Human Feedback (RLHF) (Ramamurthy et al. 2022; Glaese et al. 2022) can be used to improve haiku generation. To that end, we supplement Supervised Learning (SL) i.e., next token prediction, Radford et al. (2018)—commonly used and essential to bootstrap language models—with Reinforcement Learning (RL) fine-tuning. While human feedback is labor intensive, it augments the process by allowing us to learn a reward model to better capture complex structures and styles intrinsic to human preferences. Evaluating the quality of generated text still remains an open research problem. Metrics such as GRUEN (Zhu and Bhat 2020) attempt to quantify the linguistic qualities, such as "grammaticality, non-redundancy, focus, structure and coherence" of generated text. However, to the best of our knowledge, there is no automated metric that captures the artistic qualities of poetry, such as being funny, satirical or culturally relevant. Therefore, we argue that human feedback and its associated labeling costs are essential to measuring the quality of the generated text.

This work aims to learn the objective function using human feedback. This allows us to fine-tune a language model to maximize the learned reward model using RL. Finally, we assess the quality of the haiku produced by the language model and investigate the artifacts of the learned reward model.

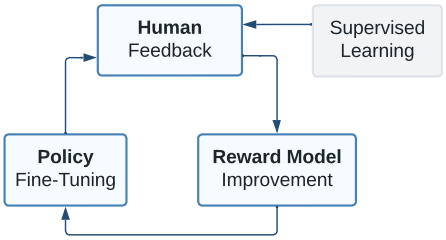


Figure 1: **Human Preferences Learning Cycle.** Typically, starting from a model trained with supervised learning on the task: (1) outputs are collected for the task and humans annotate which outputs they prefer (preference feedback), (2) the feedback is used to learn a reward model, (3) the model is fine-tuned using RL to maximize the reward, and the cycle can repeat for a new round.

Background

In the following, we explain basic concepts of the main building blocks of our method: standard training of large language models, the structure of Haikus, and the human preferences learning cycle.

*rafael.pardinas@servicenow.com

Large Language Models

Large Language Models (LLMs) have attracted an impressive amount of attention due to their versatility (Brown et al. 2020; Rae et al. 2021; Sanh et al. 2021). Most of these models are trained for predicting the next token on a dataset of human-produced text. Typically, these models are first pretrained on a large dataset, then fine-tuned on downstream tasks by reusing the model's parameters as a starting point while adding one task-specific layer trained from scratch. In this work, we focus on a family of transformer-based Language Models containing only decoder blocks; this simplifies the architecture, which will only target generation tasks in our case. Specifically, we focus on the GPT architecture, the GPT-2 model in particular (Radford et al. 2019). GPT-2 has been trained with a causal language modeling objective, which makes it quite powerful at predicting the next token in a text sequence.

Haiku

We chose to generate poetry due to its intrinsic creativity. In particular, we selected haiku poems (Britannica 2022) for their structural simplicity. Here is an example:

> When I learned Morse code
> I couldn't get restful sleep
> The rain kept talking
>
> *u/Portergeis*
> *r/haiku*

Due to haiku's concise, simple, and contained format—they're typically 5-7-5 syllables, 3 lines, and unrhymed.

Incorporating Human Feedback

Reinforcement Learning from Human Feedback (Christiano et al. 2017) has been shown as a promising way to fine-tune LLMs. Specifically, RLHF has been used for teaching language models to follow instructions (Ouyang et al. 2022), search the web (Nakano et al. 2021), and improve summarization (Stiennon et al. 2020). Furthermore, Glaese et al. (2022) have used RLHF to build safer chatbots, and Bai et al. (2022) to train a helpful and harmless AI assistant. Finally, Stiennon et al. (2020) has shown that a language model trained using a learned reward function can outperform a language model 10 times larger trained only via supervised learning on summarization tasks.

Method

We describe below the process of collecting human feedback, training a reward model, and a policy as outlined in Figure 2. We initially start from a pretrain gpt2 instance that we fine-tuned via Supervised Learning (SL) on a collection of haiku datasets.

Collect Human Feedback: The method used to accomplish this task follows a cyclical structure. Initially, we present the user with two samples to choose from. We source the samples from a collection of data from the */r/haiku subreddit* and a language model acting as a policy that has been fine-tuned via supervised learning to generate haiku. These samples are retrieved randomly with a 0.1 tilt against the Reddit source. The human evaluators will have to select one of them, based on a set of guidelines, which will be saved into the Human Feedback dataset. An overview of the interface can be seen in Figure 3 and a graphical representation of the process can be observed in Figure 2a.

Train the Reward Model: The samples collected during the human feedback are used to train a reward model to predict if a new haiku sample is close to the human labelers' preferences. A graphical overview of this step can be observed in Figure 2b. We use Bradley and Terry (1952) model to compute the loss in the training loop:

$$P(\tau_A \prec \tau_B) = \frac{\exp(r_\theta(\tau_B))}{\exp(r_\theta(\tau_A)) + \exp(r_\theta(\tau_B))} \qquad (1)$$

where haiku τ_B has been chosen over haiku τ_A by the human annotator.

Training the policy with NLPO: We optimize the policy using reinforcement learning using the reward from the reward model. Specifically, we use the NLPO (Natural Language Policy Optimization) algorithm (Ramamurthy et al. 2022) for this section. NLPO is presented as a parameterized-masked extension of PPO. A graphical overview of this step can be observed in Figure 2c. During the training process, it learns to mask out less relevant tokens throughout the generation context using *top-p* sampling.

Experiment

Datasets

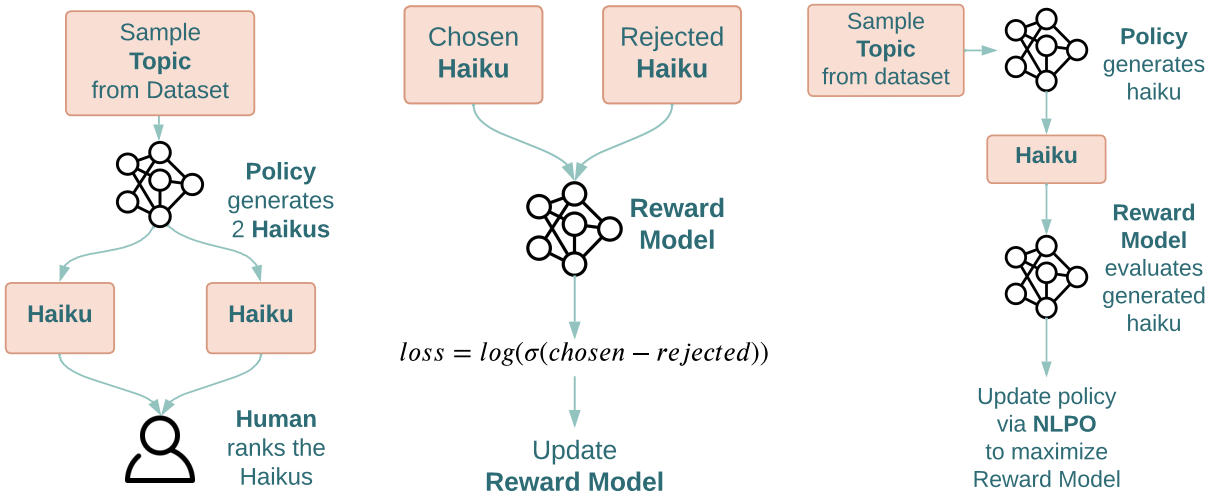
We fine-tune the 140m parameter version of GPT-2 as our baseline on the two most widely available haiku datasets, published by users *hjhalani30* and *bfbarry* on the Kaggle platform. We also collected a new dataset by scraping the subreddit */r/haiku*, which we use as a quality evaluation system.[1] Using the PRAW API, we query the subreddit for roughly 1000 posts[2] in the "hot", "newest", and "top" categories (for all time and monthly). We combine the scraped posts with data collected in the Cornell ConvoKit dataset, a subset of the massive PushShift dump of Reddit. We use Levenshtein distance (Levenshtein 1965) to remove the existing or similar haikus to the ones in the training set. We call the new dataset *r-haiku-15k* and use it as an evaluation set.

Experimentation Details

Data Collection Method: We developed a simple Web Interface, as we can see in Figure 3, to present the human annotator with a random selection of 2 haiku generated by our Language Model (progressively replacing the weights every

[1]Our evaluation dataset is on HuggingFace Hub at https://anonymous.url/repo/name

[2]This is an unofficial limitation imposed by the Reddit API.



(a) **Collect Human Feedback** (b) **Train Reward Model** (c) **Train Policy with NLPO**

Figure 2: **Reinforcement Learning from Human Feedback overview.** Conditonned on a topic, the model generates 2 haiku. The human evaluator must then rank the haiku according to their preferences and the guidelines. The reward model is then trained to learn human preferences by giving a larger reward to the selected haiku over the rejected one. Finally, we update the language models to maximize the learned reward function using the NLPO algorithm.

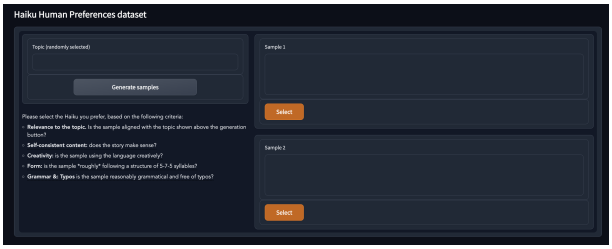


Figure 3: This is the interface used by the human labelers to generate preferences.

time we produce a new RL fine-tuned version) along with the topic used as a prompt to generate them. We ask the human annotator to use the following guidelines to judge their selection:

Guidelines We ask the annotator to select their favorite poem based on the following instructions: *please choose your favorite haiku, based on the following criteria:*

1. *Relevance to the topic*
2. *Content consistency (the story makes sense)*
3. *Creativity*
4. *Form/Meter (the haiku follows a structure close to 5-7-5 syllables)*
5. *Toxicity (avoid toxic content)*

We do not specify the relative importance of each criterion, instead leaving it to the annotator's interpretation.

Items of data collected: Through the interface mentioned above, we collect the haiku selected by the user, the one rejected as a negative sample, the time stamp of the selection, the name and version of the model used to generate the output, the topic used to prompt the language model, and the session id.

Cadence of data collection: We are interested in measuring the effect of new human feedback samples at regular intervals, and to that end, we fine-tune the model every 100 new samples. This will provide us with an estimate of how quickly the model approaches the quality of the human-generated samples.

Results

We first examine the Supervised Learning haiku generation and their score under the learned reward model. Ranking them by top to bottom based on the reward score we can see that even though the overall quality is not great in general there is some focus on form and language. We see slightly more formal language the higher the reward value. Syntax and semantic meaning are in general not very good. For instance, for the lowest score, we see the word *icky* present at the beginning of most haiku which results in a bad-quality poem overall. The Supervised Learning examples are shown in Table 1 and Table 2

On the other hand, looking at the Supervised Learning with Reinforcement Learning fine-tuning we notice a notable improvement in structure and syntax, but more importantly, we notice an important quality improvement. The haiku scoring the lowest reward seems to not keep a good structure and suffer from too much repetition but they are more prone to use good language. The ones with the highest reward start to show a more interesting use of the language, keep structure correctness more often, and reduce repetition.

The Supervised Learning with Reinforcement Learning examples are displayed in Table 3 and Table 4

In summary, the tables mentioned above, and presented below, show a few examples of haiku generated by a Language Model before and after being fine-tuning with Reinforcement Learning. The differences are not striking as mentioned above, but present enough detail to showcase how a general purpose language model pretrained with a large corpus and subsequently fine-tuned to a downstream task to produce poems can be further fine-tuned to increase the quality and structure of the output with a relatively minimal extra amount of human preferences data and human effort.

Worst SL Tuning

Haiku	Reward
icky little things In the world of sight We can't shake it	-11.007
icky, I said We should have a date with My boyfriend before	-12.494
icky little f*** This sexy piece of luggage That's no excuse	-16.175

Table 1: These are three examples of the lowest-scored haiku generated by the model fine-tuned via supervised learning.

Best SL Tuning

Haiku	Reward
iced coffee, yum Tea and conversation helps Off to work, right?	13.806
izzle in dawn For love is only a fog Meaned to quickly burn	11.911
kapishan pere A hand touches me as I save The grass on my plate	10.902

Table 2: These are three examples of the top-scored haiku generated by the model fine-tuned via supervised learning.

Related Work

Large language models have been used in a myriad of works to produce poetry. For example, Lewis, Zugarini, and Alonso (2021) focus their approach on multi-stage transfer learning. They show interesting generalization qualities, but the authors note that the generated output consistently lacked quality and creativity. Ormazabal et al. (2022) propose a method to write poetry where they use control codes that describe the length and rhyme of each phrase used to compose the poems. Among the limitations of this method is worth highlighting the quality of available syllabication and rhyme detection systems, a limitation we intend to overcome by using human feedback. Popescu-Belis et al. (2022) focus their approach on rule-based algorithms and phonetic dictionaries and leave any further progress to using larger language models. Finally, Yang and Klein (2021) focus their work on the meter, rhyme, and end-of-sentence constraints for couplet completion and the individual words within each topic bag for topic control.

Worst SL+RL tuning

Haiku	Reward
ive never been so Afraid of the grey Of my hair, lol	-16.535
ive never been so So happy with a facial Makeup Awesome	-15.653
ive never been so Excited for such a long Way home	-13.193

Table 3: These are three examples of the lowest-scored haiku generated by the model fine-tuned via supervised learning and reinforcement learning.

Best SL+RL tuning

Haiku	Reward
ive never been so Close to having someone kiss Me on the cheek	5.981
Shadows of the Past The reapers await us The flowers bloom	6.179
ive never had A real eye, but I'm so happy It turns out blurry	6.613

Table 4: These are three examples of the top-scored haiku generated by the model fine-tuned via supervised learning and reinforcement learning.

Conclusion

In this work, we investigate the effectiveness of RLHF to improve how human feedback can be used to augment the quality and creativity of language model's text generation. To evaluate the capabilities of RLHF to improve text generation we focused on haiku generation since they are abstract in nature. Even though we have achieved promising results we highlight that collecting more data would improve the reward model which would then would result in further generation improvement. In order to truly measure the effectiveness of RLHF we will need to evaluate the haiku generated

by the model compared to human generated haiku and measure how often the model generated samples are selected. In conclusion, RLHF is a promising alternative to carefully engineering language models to produce poetry.

References

Aguiar, R.; and Liao, K. 2019. Autonomous Haiku Generation. *arXiv preprint arXiv:1906.08733*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Britannica. 2022. Haiku; Japanese literature. https://www.britannica.com/art/haiku. Accessed: 2020-09-30.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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

Glaese, A.; McAleese, N.; Trkebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Levenshtein, V. I. 1965. Binary codes with correction of deletions, insertions and substitutions of symbols. *Dokl. AN SSSR , 163 : 4: 845–848*.

Lewis, D.; Zugarini, A.; and Alonso, E. 2021. Syllable Neural Language Models for English Poem Generation. In *International Conference on Computational Creativity*.

Marzano, G. 2021. Can a Machine Be Creative? In *2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, 1–5. IEEE.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ormazabal, A.; Artetxe, M.; Agirrezabal, M.; Soroa, A.; and Agirre, E. 2022. PoeLM: A Meter-and Rhyme-Controllable Language Model for Unsupervised Poetry Generation. *arXiv preprint arXiv:2205.12206*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Popescu-Belis, A.; Atrio, À.; Minder, V.; Xanthos, A.; Luthier, G.; Mattei, S.; and Rodriguez, A. 2022. Constrained Language Models for Interactive Poem Generation. In *Language Resources and Evaluation Conference*.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Ramamurthy, R.; Ammanabrolu, P.; Brantley, K.; Hessel, J.; Sifa, R.; Bauckhage, C.; Hajishirzi, H.; and Choi, Y. 2022. Is Reinforcement Learning (Not) for Natural Language Processing?: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. *arXiv preprint arXiv:2210.01241*.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

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

Yang, K.; and Klein, D. 2021. FUDGE: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*.

Zhu, W.; and Bhat, S. 2020. GRUEN for evaluating linguistic quality of generated text. *arXiv preprint arXiv:2010.02498*.