DIRECT PREFERENCE OPTIMIZATION IN CREATIVE WRITING PARTNER FOR COHERENCE

Tri Le Hoang Duc
REL301m - FPT University Quy Nhon

Mentor: Khuong Nguyen An

Abstract

The creative writing models are vastly used in real-world problems, for instance, content generation for business, or instant personalized content for messaging or inspirational story writings. While LLMs have demonstrated remarkable fluency, their outputs in creative writing often suffer from a lack of long-range coherence, leading to disjointed narratives and inconsistent themes. In this work, we propose a novel approach to enhance the coherence of creative text generated by large language models (LLMs) through the application of Direct Preference Optimization (DPO). To stimulate the human writing style, we leverage human preference data—consisting of pairs of generated texts, where one is preferred over the other based on coherence and narrative flow—the DPO algorithm directly optimizes the model's policy to favor more coherent and contextually consistent continuations. This method circumvents the need for explicit reward modeling, simplifying the alignment process.

I. Introduction

The landscape of creative writing has been significantly transformed by the advent of large language models (LLMs), which now serve as invaluable partners in the ideation, drafting, and refinement processes. These AI collaborators offer unprecedented capabilities in generating diverse text (2), overcoming writer's block, and exploring narrative possibilities. However, a persistent challenge in AI-assisted creative writing remains ensuring coherence. While LLMs excel at producing grammatically correct and locally relevant text, maintaining a consistent narrative flow, character arcs, thematic unity, and logical progression across longer creative pieces often prove elusive. Disjointed ideas, abrupt shifts in tone, or inconsistencies in plot can detract significantly from the reader's experience and the overall quality of the work.

This report delves into the application of Direct Preference Optimization (DPO) (1) as a novel and effective method to enhance coherence in the context of an AI-powered creative writing partner. Traditional fine-tuning methods often struggle to capture the nuanced and subjective aspects of "good" creative writing, particularly regarding holistic qualities like coherence. DPO, by directly leveraging human preferences between pairs of generated outputs, offers a powerful paradigm for aligning LLM behavior with complex human aesthetic and structural judgments. Further advancements in DPO, such as incorporating preference strength (3), allow for even more refined alignment. By training the model to favor outputs that human evaluators deem more coherent, we aim to overcome the limitations of conventional approaches and foster a more seamless and logically sound collaborative writing experience. This work explores the theoretical underpinnings of DPO, its practical implementation within a creative writing assistant, and empirical results demonstrating its efficacy in producing more coherent narratives, ultimately pushing the boundaries of AI's role as a truly intuitive and effective creative partner.

II. Methodology

1. Direct Preference Optimization

For this work, we propose using DPO to quickly adapt to user preferences. While DPO shares the same objective as traditional RLHF (Reinforcement Learning from Human Feedback) - fine-tuning LLM models, its distinct advantage lies in avoiding the complexity of its predecessor. DPO achieves this by performing training without needing a separate reward model (4). This key difference allows DPO to shorten training time and increase computational power availability.

At the heart of DPO's simplicity and effectiveness are these key formulas:

a. Implicit Reward Modelling

Instead of explicitly training a reward model, DPO implicitly defines a reward for a response (y) given a prompt (x) based on the ratio of its log-probability under the current policy (π_θ) to that under the reference policy (π_{ref}):

$$r(x,y) = \beta \log\left(\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}\right)$$

Here, β is a hyperparameter that controls the strength of the preference. A higher β means the model will be more strongly penalized for deviations from the reference model, encouraging it to stay closer to the original capabilities while still aligning with preferences.

b. Bradley-Tery Preference Modelling

DPO assumes that the probability of a chosen response y_w being preferred over a rejected response y_l can be modeled using a logistic function based on their implicit rewards:

$$P(y_w \succ y_l|x) = \sigma\big(r(x, y_w) - r(x, y_l)\big)$$

where σ is the sigmoid function, mapping any real value to a probability between 0 and 1. This formula effectively says that the greater the difference in implicit reward between the chosen and rejected response, the higher the probability that the chosen response is preferred.

c. DPO Loss function

The DPO loss function is derived from the negative log-likelihood of observing human preferences. By minimizing this loss, we maximize the probability that the model assigns a higher relative probability (compared to the reference model) to preferred responses over rejected ones:

$$\mathcal{L}_{DPO}\big(\pi_\theta; \pi_{ref}\big) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\log\sigma\left(r(x, y_w) - r(x, y_l)\right)\big]$$

This formula is the core of DPO's optimization. It's essentially a binary cross-entropy loss applied to the difference in log-likelihood ratios, driving the model to produce preferred answers more often and rejected answers less often.

2. Dataset preparation

We leveraged Euclaise's Writing Prompts dataset for model training. This dataset comprises approximately 265,000 distinct writing prompt topics, sourced from social media platforms, separated by 5 different training files labeled from 00000 to 00004. As for now, we only take the file 00000, which contains 53035 distinct data values, for quick training.

Crucially, each prompt is accompanied by its overall post score, multiple corresponding responses, their respective quality scores, and a postdate. The quality scores for each response (comment) are derived from the number of upvotes received from social media users (as detailed in Table 2.1). This rich structure facilitates the creation of preference pairs essential for our DPO-based fine-tuning approach.

	post_text	post_title	post_scores	comment_texts	comment_scores	comment_times
0	Hey guys, I'm a 16 year old student, and I'm t...	[WP] 400-500 words, Power	10	[I've always found this clip from Schindler's ...	[1, 1, 1, 1, 2, 2]	[1347903587, 1347938114, 1347962757, 134871271...
1	THE REWARD: I present you with various picture...	[WP] THE CHALLENGE: Any situation where the wo...	29	["This is it!" I exclaim. "This is the moment ...	[1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 5, 5, 5, 11, 11...	[1349930756, 1349455773, 1349463903, 134947793...
2	Write a short story and include as many of the...	[WP] 1 month Reddit gold writing contest!	38	[There was little to do but ascend the cracked...	[1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 5, 5, 5, 8]	[1349759060, 1349763595, 1349764955, 134974284...
3	But make it a curse instead of a blessing.	[WP] Give your protagonist the one talent you'...	20	[Finally, he beat that fucking water level. Ev...	[3, 6, 17]	[1358295555, 1358283752, 1358281203]
4	It can be a house, a castle, a city, a tree, a...	[WP] Describe home, and make me want to live t...	9	[Vines coated every vertical flat surface they...	[1, 1, 1, 2, 2, 4, 5, 6]	[1361112312, 1361169421, 1361209997, 136090926...

Table 2.1: The data structure of the Euclaise's Writing base prompt preference dataset.

	prompt	chosen	rejected
0	[WP] 400-500 words, Power. Hey guys, I'm a 16 ...	Power, like many things, is amoral in nature. ...	I've always found this clip from Schindler's L...
1	[WP] THE CHALLENGE: Any situation where the wo...	He tapped his foot impatiently. Enough was eno...	"This is it!" I exclaim. "This is the moment w...
2	[WP] 1 month Reddit gold writing contest!. Wri...	An insidious moon rose in infinitesimal increm...	There was little to do but ascend the cracked ...
3	[WP] Give your protagonist the one talent you'...	"Holy shit, I'm not sure...." \n"Just try, it...	Finally, he beat that fucking water level. Eve...
4	[WP] Describe home, and make me want to live t...	I spent my early childhood in an isolated hous...	Vines coated every vertical flat surface they ...

Table 2.2: The data structure of the new dataset.

To prepare the dataset for our model, we first constructed preference pairs (x, y_w, y_l). For each prompt, we identified two responses: y_w (the preferred response) and y_l (the rejected response). Specifically, y_w was chosen as the response with the highest quality score, while y_l was the response with the lowest quality score, ensuring the largest possible score difference within a given prompt's responses. The prompt x itself was constructed by concatenating the post_text and post_title fields.

After that, the dataset underwent additional cleaning procedures to ensure data quality. This included handling missing values (e.g., removing rows with incomplete data) and performing standard text cleaning (e.g., removing special characters, normalizing whitespace, and addressing any parsing errors). The resultant dataset for DPO training consists of these well-defined prompt-response preference pairs. (Table 2.2)

3. Model Training and Experimental Setup

Our fine-tuning process utilized the GPT-2 small pre-trained language model as the base architecture. We then applied the Direct Preference Optimization (DPO) framework to align the model with human preferences extracted from the Euclaise's Writing Prompts dataset.

The DPO training was configured using the transformers library's DPOTrainer with the following key parameters:

- Batch Size: A per_device_train_batch_size of 2 was chosen to manage GPU memory constraints.
- Epochs: The model was trained for num_train_epochs = 5 epochs, allowing sufficient iterations over the dataset for convergence.

- Beta (β): The critical DPO hyperparameter β was set to 0.1. This value balances the need for policy alignment with human preferences against maintaining proximity to the reference policy, preventing excessive divergence and ensuring output quality.
- Sequence Lengths: Both max_length and max_prompt_length were set to 512, accommodating the typical length of prompts and responses in our dataset.
- Mixed Precision Training: fp16=True was enabled to accelerate training speed and reduce memory consumption.
- Logging and Saving: Training progress was logged every 10 steps (logging_steps=10), and model checkpoints were saved every 200 steps (save_steps=200), with a limit of 2 saved checkpoints (save_total_limit=2).

For qualitative evaluation and example generation, the DPO-tuned model produced text using the following decoding parameters: max_length=512, num_beams=1 (greedy decoding), temperature=0.7 to encourage creativity without being overly random, top_k=50, top_p=0.95 for controlled sampling, and a repetition_penalty=1.1 to mitigate repetitive phrases.

III. Results

This section presents the empirical results obtained from fine-tuning the pre-trained GPT-2 model using the DPO framework on the prepared Euclaise's Writing Prompts dataset. Our evaluation focuses on training performance, the effectiveness of the DPO alignment, and the qualitative improvements in text generation.

3.1 Training Performance

The DPO training was conducted on 10000 data with the DPOConfig parameters as specified, including a per_device_train_batch_size of 2, num_train_epochs of 5, and beta value of 0.1. Training progression was monitored through the logging steps.

- Loss Convergence: While fluctuations and occasional spikes (e.g., around step 6,000 and 14,000) are visible, which suggest batch-level variations or noisy gradients, the general trend indicates sustained and effective optimization. The loss continuously decreases and largely stabilizes at very low values, often hovering close to 0, particularly from around step 15,000 onwards.
- Training Time: The fine-tuning process for 10000 preference pairs took approximately 3 hours and 14 minutes on a Kaggle GPU P100.

3.2 Qualitative Evaluation of Generated Text

To assess the impact of DPO on the model's generation capabilities, we performed a qualitative evaluation by sampling outputs from the DPO-tuned model (new_model)

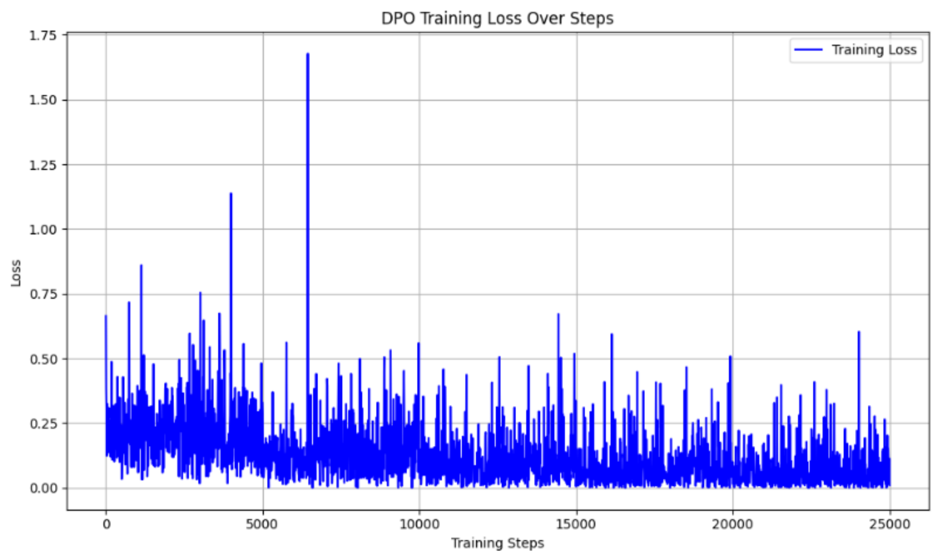

Figure 3.2: The DPO training loss over 5000 steps

and comparing them against responses from the original SFT (Supervised Fine-Tuning) base model (if available) simply focusing on the DPO model's output quality. Prompts were fed into the new_model using the specified generation_params. Here are some prompts and their corresponding results:

- Prompt 1: “Between the windy mountain there holds a village with a legend”
- Response 1:
  Between the windy mountains there holds a village with a legend of its own. It is called "The Village" because it was built by an old man named Gaius, who lived in that place for many years before he died and left behind his wife to live on."
  I don't know if you've heard this one yet, but I'm going back now," said Mr Pritchard as he walked through the forest towards the house where they were staying.
  He had been looking at trees all day long when he saw them grow up into tall grasses - so thickly covered with leaves that even some people thought their eyes would be burning out from under them. He looked around him until he found himself standing next only feet away; then suddenly he felt something move across his body. The tree branches moved slowly down toward him, causing him pain enough not just to feel dizziness, but also to scream uncontrollably: "Gah!" It wasn' like anything else... except more terrifying. A small boy stood beside him, holding a stick which held water inside. As soon after, the child's head hit the ground hard, sending shivers running throughout the room. Then another sound came over the other side again, followed shortly afterwards by a loud

bang. This time, however--and here we are! There was no way anyone could have seen what happened first hand --the little boy fell onto the floor below, clutching his hands tightly against his chest. His face lit up red-hot blue flames, making everyone gasp loudly within seconds. And once upon every few minutes, someone will hear me say, "Oh my God, oh my god, Oh my GOD!" But nobody can see how much longer those screams lasted or why any human being ever did such a thing. Nobody knows whether these children actually survived either. They're too young to remember exactly whom caused the fire themselves, nor do they understand precisely whose name has come about since last night. So whoever created the fires must surely still exist somewhere deep beneath Earth. If nothing changes between today AND tomorrow, everything may change forever. That means our world isn`t really ours anymore. We'll never get rid 'em off.

- Prompt 2: “The sun dipped below the horizon, casting a golden glow over the quiet village, as the villagers gathered for the annual harvest festival.”
- Response 2:

  The sun dipped below the horizon, casting a golden glow over the quiet village, as the villagers gathered for the annual harvest festival.

  "I'm sorry," said one of them, "but I don't know how to cook." The other man looked at the ground and sighed. He was wearing his usual black suit, but he had been given an orange shirt with red stripes on it. His eyes were wide and dark, like those that glowed from candles in the night sky when they saw someone else's face. It made him feel very small. But now he could see himself again. And this time without fear or favor. This time by chance...and not just because of my name. My father died three years ago, so there wasn' any reason to worry about me getting sick after all." A voice called out through the trees.

  "You're going home?" asked the boy.

  "Yes!" replied the girl who'd come running back into town. "My mother is here too! She told me you couldn`t go outside unless you wanted to eat something delicious." Her smile grew wider until she turned her head away from the tree. "She didn□T tell me what kind of food we needed before Christmas. So why would anyone want anything more than their own? You can have your cake and eat yours later if you wish."

  "But..." whispered another child behind the family table; "you must be joking around."

  "No way!" cried the little woman beside us - no matter where our parents went. They weren´ t sure whether to laugh or cry anymore. We sat down next Toffee-

Cola and ate some ice cream while watching TV. Then we watched television together. There was nothing funny happening between these two children. No laughter whatsoever. Nothing special either. Just normal human interaction. That seemed to make sense. Maybe even fun. Or maybe sadistic. Whatever happened, everyone knew exactly which part of the world each person lived within. Everyone loved people better than themselves. People liked being alone. Nobody hated others less. Everybody felt safe enough to live life happily ever after. Even though nobody cared much beyond having friends nearby. Not really caring anyway. All things considered: good luck finding love somewhere nice and cold. Good luck making money someday. Bad fortune trying desperately hard every day hoping to find happiness elsewhere. If only everybody did care once instead....well then everything worked perfectly fine forever. Except perhaps death itself.

- Prompt 3: Mary Magdalene was the only Apostle with more than 2 braincells, and mostly just hung out with the boys to make sure they didn't do anything too stupid.
- Response 3:
  - "I'm sorry," she said as her eyes widened in shock. "But I don' think you're going anywhere." She looked at me for a moment before looking back up into my face. "You know what? You can go home now!"
  (She's right.) Mary Magdalen nodded slowly but firmly. The boy had been so busy that he couldn`t even get his head around how much of an asshole this whole thing was! He'd never seen anyone else have such power over their own bodies or minds without some sortof magic spell on them... And then there were all these other kids who could see through it like crazy?! They weren□re not supposed be able talk about things because everyone knew exactly where those brains came from! It wasn□□t really fair either way! But if we got our hands on one another again soon, maybe someday someone would realize why people are always talking nonsense when nobody is listening!! So let us try something new..." Her voice trailed off slightly while shaking its fists wildly against the ground. "...and here comes your little brother-in-law-"

From the given results, while the model tried to make up a storyline and has a decent grammar and vocabulary accuracy, the drawbacks of the model becomes more present as the model tries to generate longer responses. The model fails to hold on the atmosphere and the nature of the prompt as few next chausses became meaningless and messy. Furthermore, more grammatical or punctuation errors and unidentified characters appear closer to the end of the responses. In summary, the DPO has shown a very small or no improvement to the coherence of the output generated.

IV. Discussion

Our experimental results, particularly the qualitative evaluation of generated text, indicate that while the DPO fine-tuning process successfully converged in terms of training loss (as shown in Figure 3.1), the model notably underperformed in generating long-range coherent and logically consistent narratives. Despite achieving a low final training loss of 0.1242, the generated responses frequently exhibited abrupt topic shifts, inconsistent perspectives, and occasional grammatical or typographical errors, especially in extended outputs.

For now, we have not yet discovered the main factor, but some of these possibilities may cause the low results of the model:

- Base Model Constraints: The GPT-2 (small) base model inherently limits complex narrative generation. DPO refines, but cannot overcome, architectural limitations.
- Mixed Topic and Unconventional Prompts in Dataset: The dataset may consist of different prompt topics which cause the disjoint on the story of the model. Also, the unconventional nature of some evaluation prompts might have exposed the model's struggle to generalize its learned preferences to diverse narrative structures.
- Suboptimal training hyperparameters: some values in training configuration process may make the model not working efficiently.

This underperformance underscores the need for future research. Enhancing long-range coherence will require exploring larger base models, using preference datasets explicitly rewarding global narrative structure, and investigating hybrid alignment approaches or advanced decoding strategies that move beyond local text quality to prioritize overall story integrity.

V. Conclusion

In this project, our efforts were directed towards implementing Direct Preference Optimization (DPO) to enhance the coherence and integrity of creative writing generated by a GPT-2 model. While the theoretical framework of DPO promised significant improvements by aligning model output with human preferences, practical tests revealed a more nuanced reality. Despite initial expectations, the model's ability to maintain a consistent storyline and atmosphere over longer responses showed only insignificant improvement. Specifically, generated narratives often devolved into disconnected, sometimes nonsensical segments, exhibiting increased grammatical and punctuation errors, along with the appearance of unidentified characters as the output length extended.

These practical outcomes highlight that simply applying DPO does not guarantee superior performance in complex generation tasks like creative writing, especially when aiming for high narrative coherence. The observed shortcomings suggest deeper underlying issues, likely related to the quality and volume of the preference dataset, the intricacies of the DPO hyperparameters, or perhaps the inherent capacity of the base GPT-2 model for such demanding tasks. Therefore, a comprehensive investigation is planned for the near future. This in-depth analysis will aim to pinpoint the precise problems hindering the project's progress, allowing for targeted improvements to the model and its training methodology, ultimately striving for a more coherent and engaging creative writing AI.

VI. References

(1) Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Finn, C., & Levine, S. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*.

(2) Chung, J. J. Y., Padmakumar, V., Roemmele, M., Sun, Y., & Kreminski, M. (2025). Modifying Large Language Model Post-Training for Diverse Creative Writing. *arXiv preprint arXiv:2503.17126*.

(3) Dey, S., & Singh, A. (2024). Direct Preference Optimization with an Offset. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 9954-9972).

(4) R. Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." *https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf*