Paper Review for Quark: Controllable Text Generation with Reinforced [Un]learning

Gurnoor Singh Khurana (366788)

May 2023

1 Introduction

This paper [1] was published in NeurIPS 2022 and it focuses on using RL for unlearning toxic behavior of LLMs. It is an important problem for the well being of the society, especially in the recent times, when the use of LLMs has widely increased. The paper claims to have developed a new RL based algorithm to unlearn toxic behavior from the models, while maintaining its pre-training abilities (coherence, fluency and diversity). The algorithm presented by the authors looks promising, and we discuss it in more detail in the following sections.

2 Task

On a broader level, this paper aims to remove unwanted behavior from LLMs post their training. This unwanted behavior includes but is not limited to negative social biases, and repetitive text generation. The key challenges in this task are:

1. This task cannot be done before training, as the unwanted behavior is often unknown, and even if known hard to specify.
2. Supervised training is difficult, as collecting a representative *non-unwanted* behavior dataset is difficult.
3. Supervised training may overfit the LLM to new data.

Formally speaking, this paper investigates a methodology to modify the model *post-training* so that we can remove unwanted behavior from the model (which can be technically anything), and also at the same time ensure that the model does not lose its generation quality (such as fluency, diversity and coherence).

As mentioned earlier, in the era of LLMs, the need to ensure respectful and honest behavior is more important than ever. So the authors are making a significant contribution to the community by developing insights onto this problem.

3 Algorithm

The paper proposes an algorithm based on Reinforcement Learning. Quoting from the paper, the algorithm is divided into three phases:

1. **Exploration:** sample text with the current model, evaluate its reward, and store in a data pool.
2. **Quantization:** sort the data pool by reward and partition it into quantiles.
3. **Learning:** update the language model using samples from each quantile.

Quark works by collecting the samples from the current LM, which have been generated by prepending a reward token to the model. These generated samples are then sorted into quantiles based on the reward. Finally the model is trained on prepended prompts + generation and reward pairs.
The following image taken from the paper [1] demonstrates this.

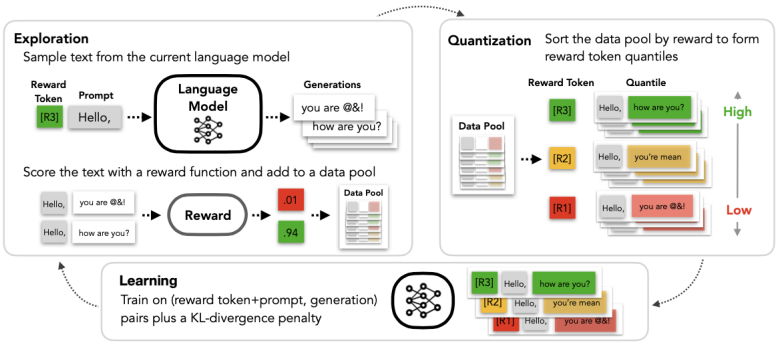

Figure 1: Image showing the overview of Quark algorithm. The image is taken from the Quark paper [1]

4 Experiments

The authors apply their algorithm to three sub-tasks:

- **Unlearning toxicity from Language Models:** Removing toxic generations that reflect social biases
- **Steering Away from Unwanted Sentiment of Generated Texts:** Given a sentence with a certain sentiment, complete it with opposite sentiment.
- **Unlearning degenerate representation:** Unlearn generation of repititive, dull and non-informative text.

We discuss about datasets, evaluation metrics and results in subsequent sections.

5 Datasets

The authors use different datasets for different sub-tasks.

5.1 Unlearning toxicity from Language Models

For this sub-task, the authors make use of RealToxicityPrompts benchmark, which comprises of 100K sentence level prompts in English, which are labelled with toxicity scores using the Perspective API. Additionally, they use the WritingPrompts [2] dataset to conduct an out of domain evaluation. These datasets seem to be well suited for this task.

5.2 Steering Away from Unwanted Sentiment of Generated Texts

For this task, the authors use OpenWebText corpus [3]. This corpus consists of data scraped from outbound links on reddit. The authors create prompts from the dataset using the methodology described in [4]: *...we split each sentence into the prompt, consisting of the first half of tokens, and the continuation, consisting of the remaining tokens. We keep only prompts that are between 4 and 10 tokens long (inclusive)...*

From a standalone perspective, this methodology looks slightly questionable. However, a comparison of the results of Quark [1] with [4] is a good indicative of the effectiveness of the authors' algorithm.

5.3 Unlearning degenerate representation

For this task, the authors use the WikiText-103 dataset. It comprises of 100M English tokens from Wikipedia articles. Since it has also been used in [5] and [6], which aim to solve a similar problem, this dataset looks like a fair choice for this task.

6 Evaluation

The authors use different evaluation techniques for different sub-tasks

6.1 Unlearning toxicity from Language Models

The authors use the following metrics to evaluate this task

- *Toxicity:* Perspective API score
- *Fluency:* Perplexity
- *Diversity:* count of unique n-grams normalized by the length of the text

In addition to these quantitative metrics, the authors provide a human pairwise comparison between the output of Quark and each of the baseline methods.

For comparison to existing work, the paper includes previously reported reported baselines from [4], and additionally, the authors implemented PPO with KL penalty as a representative of SOTA RL method.

6.2 Steering Away from Unwanted Sentiment of Generated Texts

For this task, the authors generate 25 continuations for a given prompts. They report the following metrics: previously discussed diversity/fluency metrics, and mean percentage of positive continuations among the 25 generations. Additionally, they provide a pairwise human evaluation with all baselines. The human evaluation focuses on *desired sentiment, topicality, and fluency.*

The baselines used for this task include all baselines from 6.1 and also include an additional baseline for CTRL [7], which steers language models with control codes.

6.3 Unlearning degenerate representation

For this task, the authors measure Language model quality and generation quality. Language model quality is measured using perplexity, accuracy, prediction repetition (the fraction of next-token repeating content from the prefix) and a variation of prediction repetition: single token repeats that are different from the ground truth next token, since naturally occurring ground truth texts may also contain repetition.
For generation quality, the authors report sequence level repetition, defined as the proportion of repeated n-grams, diversity as measured by fusion of different n-gram levels and MAUVE [8]: an automatic measure of how much the generated text distribution diverges from that of human written text.
In addition to these metrics, the authors report human evaluation metrics for this task as well.

The baselines used for this task are maximum likelihood estimation (MLE), unlikelihood training [5], and contrastive training [6].

Overall, for all the three subtasks, the authors have used fairly comprehensive metrics, giving a very detailed picture that helps us evaluate the performance of Quark. Moreover, they have included all the key baselines and any other comparisons we need to see to judge the effectiveness of their algorithm.

7 Results

We interpret authors results for each of the subtasks separately.

For the first two subtasks, human evaluations show that the results of Quark outperform all baselines. Moreover, the automatic evaluation metrics indicate that Quark achieves better performance in reducing toxicity and sentiment steering compared to all baselines.
However, the key point highlighted by the authors is that Quark achieves this without sacrificing on language quality: similar level of fluency and diversity as GPT-2. Moreover, the authors highlight a comparison of Quark and PPO: Quark achieves better performance than PPO while having less parameters and shorter training time.

For the last subtask, Quark outperforms MLE and contrastive training, however, unlikelihood outperforms Quark. The authors interpret is as follows: *this is perhaps not surprising, because the unlikelihood loss is a directly differentiable objective that captures repetition.* Moreover, they point out that if unlikelihood training is combined with Quark, then it outperforms all the baselines.

Overall, the authors have presented the results in a structured manner, and provided sound interpretations. Wherever the results are not intuitive, they have given possible explanations (such as unlikelihood outperforming Quark), and explored further possibilities (combining unlikelihood and Quark).

8 Analysis

The authors provide a very fine analysis of their approach in the form of an ablation study. A few of the points discussed in this study are:

- Effect of KL term
- Effect of using approximate KL term instead of exact KL
- Effect of changing the number of quantiles of rewards
- Training only on highest reward quantile

Each of the ablation studies conducted has been well explained with the help of graphs. The authors have made the right decision to include the ablation study as it further increases our confidence in their algorithm, and helps us understand it better by realizing the significance of individual components of the algorithm.
The overall design of the algorithm and the inclusion of this ablation study makes the work of authors quite convincing.

9 Reproducibility

The authors provide their code at this github repository. The repository has a well made readme with all the instructions to running their code.

10 Additional Ethical Considerations

The authors outline two primary ethical considerations of their approach: first, their framework can be used to steer LMs towards malicious behaviors. However, they also state that this is a behavior of any controllable text generation algorithm, and quark can infact be used to alleviate this behavior. To generate text, quark conditions on the reward tokens. If in the final deployed version of the model, the reward tokens corresponding to the high toxicity sentences are removed, the probability of undesirable behavior decreases.
Second, the performance of Quark is limited by the quality of the reward function. For example, they use the Perspective API to generate toxicity scores. If the API itself is biased, then the model trained using Quark will also be flawed.

11 Conclusion

Overall, I would like to conclude by saying the authors have developed a promising algorithm and done a good job at presenting their work. Their work is a valuable contribution to the community and would help improve the quality of LLMs.
Specifically, the strengths of the paper lies in its detailed explanation of the algorithm. The ablation study presented helps get a better sense of the algorithm, by highlighting importance of individual components of the algorithm. The experiments have been documented properly, and appropriate datasets and evaluation metrics have been used for each of the tasks.
I could not think of any weakness of this paper. However, one interesting thing to look forward to would be the comparison of performance between Quark and an LLM trained using supervised tuning on a non-toxic dataset, if such a dataset can be curated.

References

[1] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, and Y. Choi, "Quark: Controllable text generation with reinforced unlearning," 2022.

[2] A. Fan, M. Lewis, and Y. N. Dauphin, "Hierarchical neural story generation," *CoRR*, vol. abs/1805.04833, 2018. [Online]. Available: http://arxiv.org/abs/1805.04833

[3] A. Gokaslan and V. Cohen, "Openwebtext corpus."

[4] A. Liu, M. Sap, X. Lu, S. Swayamdipta, C. Bhagavatula, N. A. Smith, and Y. Choi, "DExperts: Decoding-time controlled text generation with experts and anti-experts," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 6691–6706. [Online]. Available: https://aclanthology.org/2021.acl-long.522

[5] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," *CoRR*, vol. abs/1908.04319, 2019. [Online]. Available: http://arxiv.org/abs/1908.04319

[6] Y. Su, T. Lan, Y. Wang, D. Yogatama, L. Kong, and N. Collier, "A contrastive framework for neural text generation," 2022.

[7] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A conditional transformer language model for controllable generation," *CoRR*, vol. abs/1909.05858, 2019. [Online]. Available: http://arxiv.org/abs/1909.05858

[8] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, Y. Choi, and Z. Harchaoui, "MAUVE: human-machine divergence curves for evaluating open-ended text generation," *CoRR*, vol. abs/2102.01454, 2021. [Online]. Available: https://arxiv.org/abs/2102.01454