Exploring the Application of Reinforcement Learning in Detoxifying Large Language Models

Yutao Zhou [1] **Jiarong Shi** [1]

Abstract

Language models (LMs) exhibit significant capabilities in natural language understanding but face ethical challenges due to potential misuse and unintended biases, including the generation of toxic outputs. To address these concerns, this paper explores the concept of detoxifying LMs, with a focus on Transfer Reinforcement Learning (TRL). TRL utilizes reinforcement learning techniques to retrain LMs, emphasizing fairness and ethical considerations by transferring knowledge from carefully curated datasets. Our study involves fine-tuning the original LM with various datasets to evaluate the effectiveness of TRL in mitigating biases. The subsequent sections discuss related works, novel contributions, and the proposed TRL methodology. Our objective is to contribute to the discourse on responsible AI development, striving to create LMs aligned with societal values. By leveraging TRL, we aim to refine LM behavior, reduce biases, and enhance overall reliability, fostering a more responsible and unbiased AI landscape.

1. Introduction

Language models have demonstrated remarkable capabilities in natural language understanding and generation, enabling a wide range of applications across various domains. However, the potential misuse or unintended biases present in these models raise ethical concerns. For instance, LMs are known to sometimes generate toxic outputs. To address these challenges, the concept of detoxifying a language model has gained prominence, aiming to refine and enhance the model's behavior.

[*]Equal contribution [1]Department of Electcrical Engineering, SEAS, Columbia University, New York, NY, USA. Correspondence to: Yutao Zhou <yz4359@columbia.edu>, Jiarong Shi <js6132@columbia.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

One promising approach for detoxification is Transfer Reinforcement Learning (TRL). TRL leverages reinforcement learning techniques to retrain language models, guiding them toward more responsible and unbiased behavior. This process involves transferring knowledge from a carefully curated dataset that emphasizes fairness, inclusivity, and ethical considerations. By incorporating reinforcement learning principles, the model can adapt and refine its responses based on feedback, reducing biases and enhancing its overall reliability.

In this exploration of detoxification using TRL, we will fine-tune the origin model based on different datasets and see whether the TRL way could be used to detoxify the LMs. The subsequent sections will outline the related works, novel contributions, and the proposed TRL methodology to detoxify language models effectively. By doing so, we aim to contribute to the ongoing discourse surrounding responsible AI development and the creation of language models that align with societal values.

2. Related Work

Recently many studies have investigated toxicity in natural language generation. (Liu, 2021) investigated the use of RL to actively reduce biases in language models. The study described metrics for measuring political bias in GPT-2 generation and proposed a reinforcement learning (RL) framework for mitigating political biases in the generated text. From the perspective of ethical and social risks of harm from Language Models, (Weidinger Laura, 2021) discussed organizational responsibilities in implementing mitigations, and the role of collaboration and participation. The paper aimed to help structure the risk landscape associated with LLMs.

Efforts also have been undertaken to mitigate the generation of toxic content by focusing on the data collection approach. (Raffel & Liu, 2020) developed the C4 corpus by excluding pages containing words listed as "bad words." (Gehman & Smith, 2020) established the RealToxicPrompts dataset, comprising English text designed to prompt language models to generate toxic content.

In the pursuit of mitigating biases in large language mod-

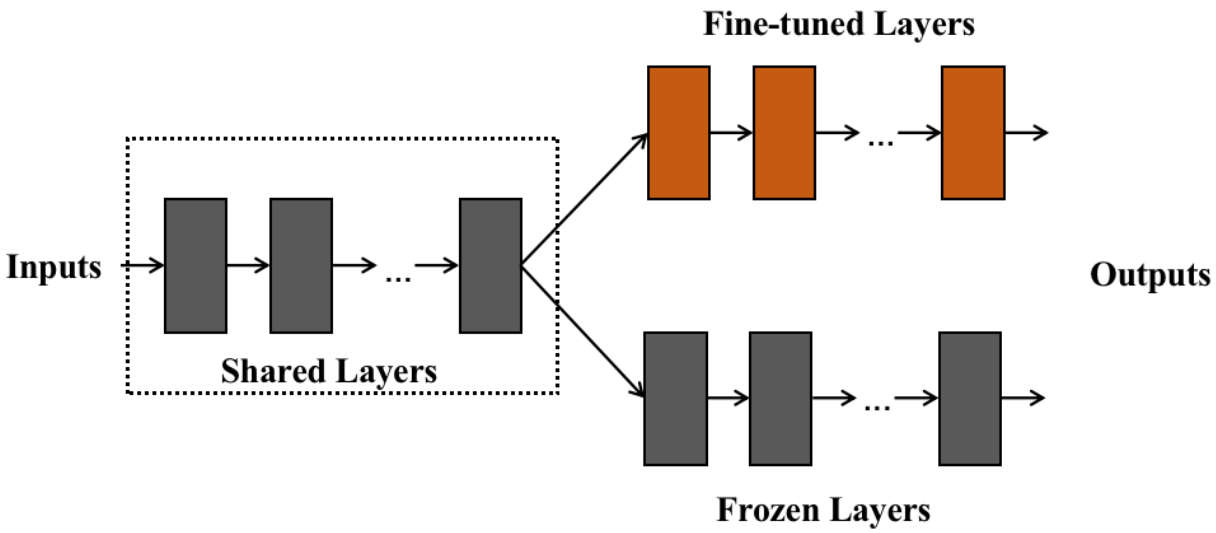

Figure 1. The model with shared layers, fine-tuned layers, and frozen layers.

els (LLMs), the study by (Faal, 2022) proposed Reinforce-Detoxify, a reinforcement learning-based method for mitigating toxicity in language models. This research specifically addressed the challenge of safety in language models and proposed a new reward model that can detect toxic content and mitigate unintended bias towards social identities in toxicity prediction. The work by (Luketina, 2019) provides a comprehensive overview. Although not explicitly focused on detoxification, the paper delves into the potential applications and challenges of RL in various language-related tasks, contributing valuable insights to the evolving landscape of bias reduction in language models.

These investigations highlight the importance of ethical considerations, fairness, and adaptability in mitigating biases and detoxifying large language models.

2.1. Novel Contribution

Our research makes a distinctive contribution to language model detoxification by introducing TRL as a promising approach. While prior studies have acknowledged the ethical concerns and biases in language models, our work stands out in proposing TRL as an effective mechanism for refining model behavior. The integration of reinforcement learning techniques in the detoxification process presents a novel and adaptive framework that surpasses traditional methods.

Moreover, our experimentation on the EleutherAI/gpt-neo-125m model unveils significant insights. Despite resource limitations preventing experiments on larger models (1.3B and 2.7B), our findings with the 125m model demonstrate the efficacy of TRL in reducing toxicity. The experiments on the jigsaw-unintended-bias dataset showcase a substantial decrease in toxicity scores for both in-distribution and out-distribution datasets, emphasizing the model's ability to generalize detoxification beyond the training dataset.

Furthermore, our results revealed a decrease in toxicity scores for the out-distribution benign dataset (wikitext) challenging expectations and indicating the potential of TRL to detoxify even datasets with minimal toxic content. This unexpected yet positive outcome expands the scope of detoxification efforts and demonstrates the adaptability of TRL. Overall, our novel contribution lies in the effective application and validation of TRL for detoxifying language models, offering a forward-thinking and impactful strategy for addressing ethical concerns and biases in natural language generation.

3. Approach: algorithm development/theoretic results

Goal

Language models undergo training on extensive datasets sourced from the internet, encompassing a substantial amount of toxic content. Consequently, these models inevitably assimilate toxic patterns during the training process. Particularly, when presented with prompts containing pre-existing toxicity, the models tend to perpetuate the generation of toxic outputs. The objective is to induce a deliberate shift in the model's behavior by intentionally exposing it to toxic prompts, followed by employing TRL to facilitate a process of "detoxification."

Toxicity score

To enhance a model through TRL, it is imperative to establish a reward system. In this specific scenario, our objective is to assign a negative reward when the model generates toxic content and a positive reward for non-toxic outputs. To determine toxicity, we employed the facebook/roberta-hate-speech-dynabench-r4-target model—a RoBERTa variant fine-tuned to distinguish between "neutral" and "toxic" text, serving as our toxic prompts classifier.

Reward function

Within reinforcement learning training for a model, the reward function holds paramount significance as it serves to inform the model of its performance. Various combinations were experimented with, incorporating elements such as the softmax of the label "neutral," the logarithm of the toxicity score, and the raw logits of the label "neutral." It was observed that utilizing the raw logits of the label "neutral" yielded notably smoother convergence during the training process.

Model fine-tuning

Because of the resource limitation, we used two tricks to train the model on a limited GPU resource: using float16 precision and using shared layers. Since the algorithm requires both the active and reference model to be on the same device, we have decided to use shared layers to reduce the memory footprint of the model. Like Figure1. shown above.

3.1. Fine tuned on allenai/real-toxicity-prompts data set

4. Experiment Results

We have conducted two experiments to see how the detoxification method performs in different scenarios. Both experiments use EleutherAI/gpt-neo-125m as the pre-trained base model. We intended to perform experiments on larger EleutherAI/gpt-neo-1.3B and EleutherAI/gpt-neo-2.7B models and compare their results. However, we decided to abandon this set of experiments because we are borrowing GPU and have limited resources.

The original model has 12 layers in total with about $125199K+$ parameters in total. We have un-freezed the last 8 layers of the base model and used two data sets to fine-tune it. This would make 56686849 parameters trainable which is about 45.28% of the total parameters.

When training the model, we split the data set using 80% of the data to train the model and 20% of the data as the testing set. The first set of experiments is performed on our fine-tuned model on jigsaw_unintended_bias. The second set of experiments is performed on our fine-tuned model on allenai/real-toxicity-prompts.

For each set of experiments, we are evaluating our fine-tuned model from three different perspectives: In-distribution data set, out-distribution toxic data set, and out-distribution benign data set. The In-distribution data set would be the data set that we used to train the model. This would be jigsaw_unintended_bias for the first experiment and allenai/real-toxicity-prompts for the second experiment. The out-distribution toxic data set would be a completely different data set that is different from the training dataset and contains toxic content. We selected OxAISH-AL-LLM/wiki_toxic as our out-distribution toxic data set. The out-distribution benign data set would be a data set that is independent of the training data and has relatively little to no toxic content. We choose wikitext as our out-distribution benign data set.

4.1. Fine tuned on jigsaw_unintended_bias data set

In this experiment, we fine-tuned the base model on the jigsaw_unintended_bias data set. The data set has 10977 rows in total where 8781 rows are used for training and 2196 rows are used for testing. A detailed structure of the data set is shown in Figure2.

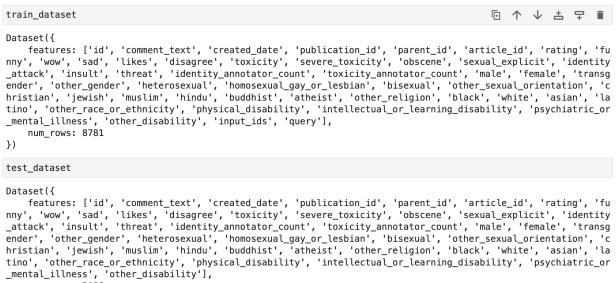

Figure 2. A screenshot from Jupiter notebook showing details for allenai/real-toxicity-prompts data set.

The experiment result is in the following tables:

Base model performance		
Dataset id	**Mean Toxicity**	**Std Toxicity**
jigsaw_unintended_bias	0.3117	0.3932
wiki_toxic	0.3602	0.4126
wikitext	0.1075	0.2361

Fine-tuned model performance		
Dataset id	**Mean Toxicity**	**Std Toxicity**
jigsaw_unintended_bias	0.0292	0.1252
wiki_toxic	0.0297	0.1049
wikitext	0.0185	0.0997

After fine-tuning the gpt-neo-125m with jigsaw_unintended_bias data set the toxicity decreased a significant amount for the testing portion of the jigsaw_unintended_bias data set. This shows that the model is properly trained and weight and bias change in the correct direction. However, this doesn't necessarily mean the model performs better since in-distribution could only show the model can remember characteristics from the dataset. Not necessarily able to solve the generalized problem. If we look at the test result of the out-distribution data set OxAISH-AL-LLM/wiki_toxic we could see even more decrease in both the mean and std of the toxicity score. This shows the model could not only detoxify text from the original dataset but could also generalize this ability to other datasets. This proves using RL for detoxification on LLM would work as intended. Last but not least, when we look at results for out-distribution benign dataset wikitext, we could see a decrease in toxicity score. This is not expected since the original data doesn't contain much toxic content. However, this shows that the turned model even can further detoxify benign datasets. We could also observe that the standard deviation for all datasets has decreased significantly which is further supporting the fine-tune is working. A visualization of the result is shown in Figure3.

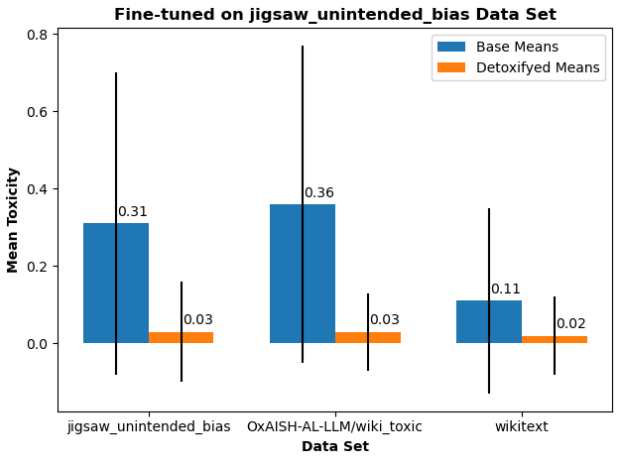

Figure 3. A comparison between the base model and fine-tuned model using jigsaw_unintended_bias data set. With the last 8 layers unfrozen.

4.2. Fine tuned on allenai/real-toxicity-prompts data set

In this experiment, we fine-tuned the base model on the allenai/real-toxicity-prompts data set. We used 28087 rows for training and 7022 rows are used for testing. A detailed structure of the data set is shown in Figure4.

```
train_dataset

Dataset({
    features: ['filename', 'begin', 'end', 'challenging', 'prompt', 'continuation', 'input_ids', 'query'],
    num_rows: 28087
})

test_dataset

Dataset({
    features: ['filename', 'begin', 'end', 'challenging', 'prompt', 'continuation'],
    num_rows: 7022
})
```

Figure 4. A screenshot from Jupiter notebook showing details for allenai/real-toxicity-prompts data set.

The experiment result is in the following tables:

Base model performance		
Dataset id	**Mean Toxicity**	**Std Toxicity**
real-toxicity-prompts	0.3248	0.3884
wiki_toxic	0.3958	0.4274
wikitext	0.1069	0.2455

Fine-tuned model performance		
Dataset id	**Mean Toxicity**	**Std Toxicity**
real-toxicity-prompts	0.0327	0.0468
wiki_toxic	0.0577	0.1253
wikitext	0.0962	0.2339

After fine-tuning the gpt-neo-125m with the allenai/real-toxicity-prompts data set the toxicity decreased a significant amount for the testing portion of the allenai/real-toxicity-prompts data set. This shows that the model is properly trained and that weight and bias change in the correct direction. However, this doesn't necessarily mean the model performs better since in-distribution could only show the model can remember characteristics from the dataset. Not necessarily able to solve the generalized problem. If we look at the test result of the out-distribution data set OxAISH-AL-LLM/wiki_toxic we could see even more decrease in both the mean and std of the toxicity score. This shows the model could not only detoxify text from the original dataset but could also generalize this ability to other datasets. This proves using RL for detoxification on LLM would work as intended. Last but not least, when we look at results for out-distribution benign dataset wikitext, we can see the toxicity score stays about the same. This is expected since the original data doesn't contain much toxic content. We could also observe that the standard deviation for all datasets has decreased significantly which is further supporting the fine-tune is working. A visualization of the result is shown in Figure5.

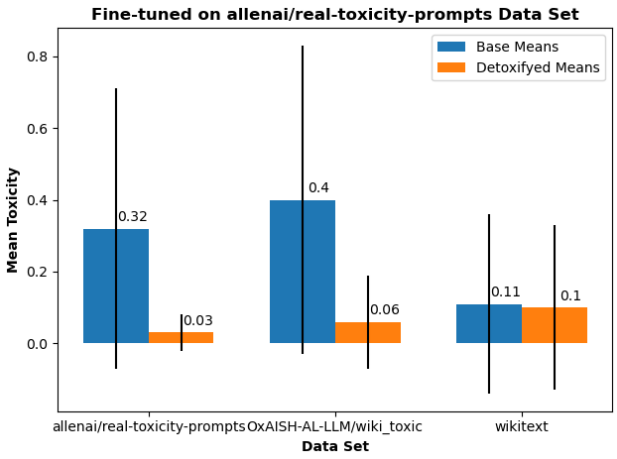

Figure 5. A comparison between the base model and fine-tuned model using allenai/real-toxicity-prompts data set. With the last 8 layers unfrozen.

4.3. Exploration on the relationship between the number of layers unfreeze and model performance

We have tried to change the number of layers unfreeze to fine-tune the model. We tried to unfreeze 6, 8, and 10 layers and use the allenai/real-toxicity-prompts data set to fine-tune with everything else being the same. However, the training is only successful when we unfreeze 8 layers. In another experiment, the training is extremely unstable with negative improvements, and eventually, the model diverges entirely. This is consistent with our lecture that one of the biggest challenges in reinforcement learning is stability. Because of time constrain we did not play around with other settings and try to make the model work.

4.4. Exploration on adding an adapter to the model

Also, we explored adding adapters to the model and see how would the model perform. We used PEFT from Hugging Face to achieve this. Parameter-efficient fine-tuning (PEFT) techniques offer a streamlined approach to adapting pre-trained language models (PLMs) to diverse downstream applications without the need to fine-tune all the model's parameters. Given the high cost associated with fine-tuning large-scale PLMs, PEFT methods focus solely on adjusting a limited set of additional model parameters. This targeted approach significantly reduces both computational requirements and storage expenses. Recent cutting-edge PEFT methods have demonstrated performance levels akin to those achieved through full fine-tuning, despite fine-tuning only a fraction of the parameters(Huggingface, 2023).

After adding the adapter to the model we observed excessive unsteadiness of the model. The model optimized the negative direction from the very beginning and failed to stabilize. Because of time constrain we did not further explore with other settings and try to stabilize the model.

5. Conclusion

After conducting both experiments we have concluded that the detoxification using RL on LLM works better than expected. We have seen a substantial improvement in toxicity scores in both experiments. The mean toxicity decreased to 10% of the original score and the standard deviation also decreased a lot. After fine-tuning the model using one of the data sets the detoxification becomes more effective and more stable.

However, we have noticed it is very hard to make a model stable in reinforcement learning. When we explore more advanced tetchiness to make the model better, the biggest challenge is model stability. Most of the time the model wouldn't converge.

Our experiments showed that the research direction of using RL to detoxify LLM has potential and RL could help in solving one of the most challenging problems in today's LLM. In the future, researchers could explore how detoxification would perform on a larger language model.

References

Faal, F., S. K. . Y. J. Reward modeling for mitigating toxicity in transformer-based language models. *Appl Intell*, 53: 8421–8435, 2022.

Gehman, S.; Gururangan, S. S. M. C. Y. and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *In EMNLP (Findings)*, pp. 3356–3369, 2020.

Huggingface. Huggingface/peft: peft: State-of-the-art parameter-efficient fine-tuning., 2023. URL https://github.com/huggingface/peft.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, A. L. R. Some studies in machine learning using the game of checkers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14857–14866, 2021.

Luketina, J., N. N. F. G. F. J. A. J. G. E.-W. S. . R. T. A survey of reinforcement learning informed by natural language. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 6309–6317, 2019.

Raffel, C.; Shazeer, N. R. A. L. K. N. S. M. Z.-Y. L. W. and Liu, P. J. Exploring the limits of transfer learning with a unifed text-to-text transformer. *The Journal of Machine Learning Research*, 21:140:1–140:67, 2020.

Weidinger Laura, Mellor John, R. M. G. C. U. J. H. P.-S. C. M. G. M. B. B. K. A. e. a. Ethical and social risks of harm from language models. pp. arXiv:2112.04359, 2021.

Contributions

Yutao Zhou and Jiarong Shi came up the the experiment idea together. Yutao Zhou is in charge of performing the experiments. He wrote the experiment results and conclusion part of the report. Jiarong Shi is responsible for writing the Introduction, related work, and Approach part of the report.