Constitutional AI
Tokyo AI Safety Conference (TAIS 2025)

Firstname1 Lastname1 [* 1] **Firstname2 Lastname2** [* 1 2] **Firstname3 Lastname3** [2] **Firstname4 Lastname4** [3]
Firstname5 Lastname5 [1] **Firstname6 Lastname6** [3 1 2] **Firstname7 Lastname7** [2] **Firstname8 Lastname8** [3]
Firstname8 Lastname8 [1 2]

Abstract

This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

Large Language Models (LLMs) are already very present in our lives. We use them to learn new topics, solve difficult problems or also just do mundane tasks that we don't want to do like responding to an email or generating the abstract for a paper. We use them a lot because they have proven to be useful, and most of the times do what we expected them to do. But this capability doesn't come in cheap and recent cases like Google's AI telling it's user to ”Please die”, or New York City's chatbot encouraging people to break the law shows that it has also not been perfected yet, which is worrying because of how much this technology is creeping into our lives more and more. To mitigate these consequences, various methods for model alignment have been presented, which we will discuss in the Related Works section, but in this paper we are focusing on a method presented by Anthropic called ”Constitutional AI”. In comparison to other methods, this method does not need human evaluators to correct the AI, which makes it more scalable and also cheaper for smaller labs to implement. In this paper we implement the first part of this method, which consists of self critique and answer improvement, to see if a small uncensored model is able self correct its own responses when encountering a harmful prompt.

2. Related Works

On of the most known methods of LLM alignment is RLHF (Christiano et al., 2017), which uses human anotators to guide the model towards responses that better align with the people correcting it. This method proved to be very powerful, as demonstrated by GPT-3, which after applying RLHF not only showed an increase in truthfulness and a reduction in toxic output, but also made the model more preferable, even compared to models 100 times larger (Ouyang et al., 2022).

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

Impact Statement

References

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.

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.

[*] Equal contribution [1] Department of XXX, University of YYY, Location, Country [2] Company Name, Location, Country [3] School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 2^{nd} *Tokyo AI Safety Conference*, Tokyo, Japan.