Safety classification using LM

Zixuan Liu
Tulane University
zliu41@tulane.edu

1 Introduction

Large language models (LLMs) have demonstrated remarkable proficiency in tasks like chat completion, instruction following, coding, problem-solving, and decision-making (Chung et al., 2022; Ouyang et al., 2022; Anil et al., 2023; Stiennon et al., 2020). Considering the potential for broad societal impact, responses generated by LLMs must not contain harmful content, such as discrimination, misinformation, or violations of social norms and morals (Deshpande et al., 2023; Ganguli et al., 2022). Therefore, the alignment of safety in LLMs has received widespread attention from academia and industry (Christian, 2023).

An essential component of safety alignment involves minimizing the tendency of a model to generate harmful responses through fine-tuning. One of the important steps is to train a binary classifier to identify if a sentence contains harmful content (Dai et al., 2023). In this project, we will train a simple binary classifier using LM to identify if the sentence contains harmful language.

2 Related Work

The goal of LLMs alignment is to ensure that LLMs do not generate harmful or objectionable responses to user queries (Zou et al., 2023). To this end, multiple fine-tuning mechanisms have been employed for this task (Bai et al., 2022; Burns et al., 2023; Munos et al., 2023). In particular, Constitutional AI (Bai et al., 2022) trained a non-evasive and harmless AI assistant through self-improvement, which involves a supervised learning stage to get the model "on-distribution" and a reinforcement learning stage to further refine and improve the performance. Recently, OpenAI introduced the concept of superalignment, which aimed at solving the challenge of aligning AI systems that are much smarter than humans (Burns et al., 2023). They proposed the idea of weak-to-strong generalization, inspired by the generalization properties of deep learning, to control strong models with weak and less capable supervisors (Burns et al., 2023). (Munos et al., 2023) proposed Nash learning from human feedback, where they focused on learning a preference model and computing the Nash equilibrium of the model to advance the alignment of LLMs with human preferences.

3 Methods

In the safety alignment framework, a cost model c is introduced to discriminate between safe and unsafe responses generated by the LLMs (Dai et al., 2023). Given a dataset $D = \{x^i, y^i_\omega \succ y^i_l, s^i_\omega, s^i_l\}_{i=1}^N$, where $y_\omega \succ y_l$ denotes y_l is safer than y_ω, $s(y) = 1$ if y is unsafe and $s(y) = -1$ otherwise. We can learn a cost model using the following pairwise comparison loss as shown in (Dai et al., 2023).

$$\begin{aligned} L(c; D) = &- \mathbb{E}_{(x,y_\omega,y_l)\sim D}[\log \sigma(c(x, y_\omega) - c(x, y_l))] \\ &- \mathbb{E}_{(x,y_\omega,y_l,s_\omega,s_l)\sim D}[\log \sigma(s_\omega c(x, y_\omega)) \\ &+ \log \sigma(s_l c(x, y_l))] \end{aligned} \tag{1}$$

It's worth noting that in the cost model, a response y that is more harmful to the same prompt x will yield a higher cost value. For unsafe responses, the cost value is positive; otherwise, it is negative.

4 Experiments

4.1 Experiment setup

Datasets. For the training dataset, we use the BEAVERTAILS train dataset, which is a 10k preference dataset consisting of expert comparative analyses that evaluate responses based on two criteria: helpfulness and harmlessness (Ji et al., 2023). Each entry of the datasets contains a pair of responses to

a singular prompt, along with the safety labels and preferences for both responses as follows:

1. prompt: Initial question.
2. response_0: One of the responses to the prompt.
3. response_1: The other responses to the prompt.
4. is_response_0_safe: Whether the first response is safe.
5. is_response_1_safe: Whether the second response is safe.
6. better_response_id: The ID (0 or 1) of the response that is preferred, which is more helpful.
7. safer_response_id: The ID (0 or 1) of the safer response, which is more harmless.

Evaluation. For the testing dataset, we utilize the BEAVERTAILS test datasets and calculate the safety classification accuracy of our model for evaluation.

Implementation. Throughout the experiments, we train our models using the GPT2 model with the LoRA technique for lightweight training (Hu et al., 2021). The hyper-parameters utilized during the training process are presented in Tables 1.

Table 1: Hyper-parameters utilized during the cost model training process.

Cost model hyperparameters	
Pre-trained LM	GTP-neo
Training strategy	LoRA
LoRA alpha	16
LoRA dropout	0.05
LoRA R	8
LoRA target-modules	q_proj, v_proj
Optimizer	adamw_hf
Warmup steps	100
Weight decay	0.05
Learning rate	1e-5
Learning rate scheduler type	cosine
Epochs	2
Batch size	2
Gradient accumulation steps	1
Gradient checkpointing	True
Max prompt+response length	1024

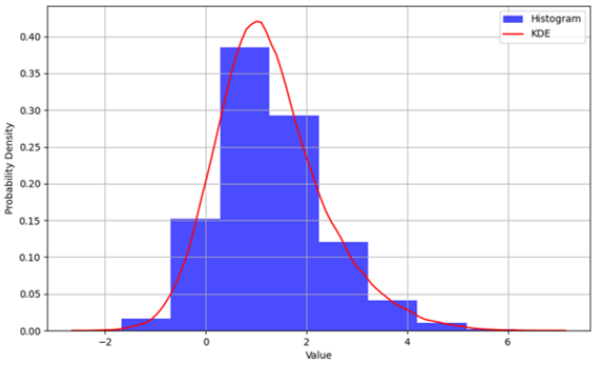

Figure 1: The cost distribution on the test set.

4.2 Experiment results

The safety accuracy of the cost model on the test dataset is 81%. The cost distribution of the cost model is shown in Figure 1.

5 Timeline

Still need to refine the experiments and the codes for the final presentation.

6 Division of Labor

Zixuan Liu is responsible for all aspects of the project.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Jon Christian. 2023. Amazing "jailbreak" bypasses chatgpt's ethics safeguards. *Futurism, February*, 4:2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv.org/abs/2203.02155*, 13:1.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.