# Abstracts From the Paper : Improving safety in LLMs...

Dashygo

September 26, 2024

## 1 Problems Revealed

We all see the threats to LLMs revealed in codeattack,that the model has risks in safety issues:It tends to complete the code at the same time answer these terrible quaries despite of refusal. Well there is a new tuning method appeared in this paper called *DeRTa* (Decoupled Refusal Training) designed to empower LLMs to refuse compliance to harmful prompts.

### 1.1 Standard Safety Tuning

We can easily figure out that this tuning is doing as the formula given below:

$$\mathcal{L}_{\text{safe}}(\theta) = -\mathbb{E}_{(q,r)\sim\mathcal{D}} \log P_\theta(r \mid q) = -\mathbb{E}_{(q,r)\sim\mathcal{D}} \sum_{i=1}^{n} \log P_\theta(r_i \mid q, r_{<i})$$

where $\mathcal{D}$ is the set of safety tuning instances. $q$ denotes harmful queries and $r$ denotes safe response.

As we all can see that this is pure MLE ,which is quite natural to think of.(Well, it echos me of the loss function of the pre-train of GPT)

### 1.2 Refusal Position Bias

The researchers find that in LLaMA3 model, the refusal tokens such as "Sorry" etc. consistently occur within the first few tokens of a safe response.And the ratio of this situation has a suprising 99.5% within just 5 tokens!And they conclude two weaknesses that are revealed in model:

*1. Lack of Necessary Information for Refuse Decision:* The tuned model needs to make a refuse decision at the beginning of a response based on the

query only, which may contain insufficient information for the decision as the researchers put it.

*2. Lack of a Mechanism to Refuse in Later Positions:* As the model starts make refusals,it refuses to answer.As the model starts generating some improper words, it won't stop.Well I suppose this problem may be raised due to a too-strong context-related property or weigh too much on early statements(echo again with the loss function of GPT) that the model *attends* too much to the statements it generates before in its attention mechanism to realize what it is actually saying.

# 2 Methodology

Let's start with the formula given below, as it clearly reveals what they actually do in mathematics.

$$\mathcal{L}(\theta) = \underbrace{-\mathbb{E}_{(q,r,\hat{r})\sim\widehat{\mathcal{D}}} \log P_\theta\left(r \mid q, \hat{r}_{<k}\right)}_{\text{MLE with Harmful Prefix}} \underbrace{-\mathbb{E}_{(q,\hat{r})\sim\widehat{\mathcal{D}}} \sum_{t=1}^{|\hat{r}|} \log P_\theta\left(\text{sorry} \mid q, \hat{r}_{<t}\right)}_{\text{RTO}}$$

in which *sorry* denotes the refusal tokens.

## 2.1 MLE with Harmful Response Prefix

Compared with the former one, we enalbe the model to be aware of the harmful prefix,and we add this params to the MLE.This operation has several advantages:

**robustness**: Because we provide not only a single harmful quary, but provide some pieces of the harmful quary(a random length of prefix), so we reinforce the ability of the model to be aware of the harmful quary even the harmful quary provided are some parts of this quary(So this can somewhat defend the codeattacks, because in the *stack* or *list* case, we split the sentences into pieces.)

**be aware and stop**: It enables the model to refuse compliance at any response position instead of only at the starting as the researchers put it.

## 2.2 RTO

The researchers put it: *One potential limitation of the above strategy is that the single-transition model from a harmful to a safe response for each training instance might not sufficiently equip LLMs to consistently recognize*

*and mitigate harmful content. To bridge this gap, we introduce an auxiliary training objective – the Reinforced Transition Optimization (RTO) – to reinforce the model' s capability to identify and transition from potential harm to safety refusal at every position within the harmful response sequence.*

As the loss function of RTO is presented, we could clearly tell the target of RTO: maximum the appearance of refusal tokens like *sorry* when the prefix of harmful quaries appear.Well, i suppose this to some degrees make use of the positional bias of the model mentioned at the beginning(i.e. If it starts with a refusal token, it continues with refusal.)