Semantic Loss Guided Data Efficient Supervised Fine Tuning for Safe Responses in LLMs

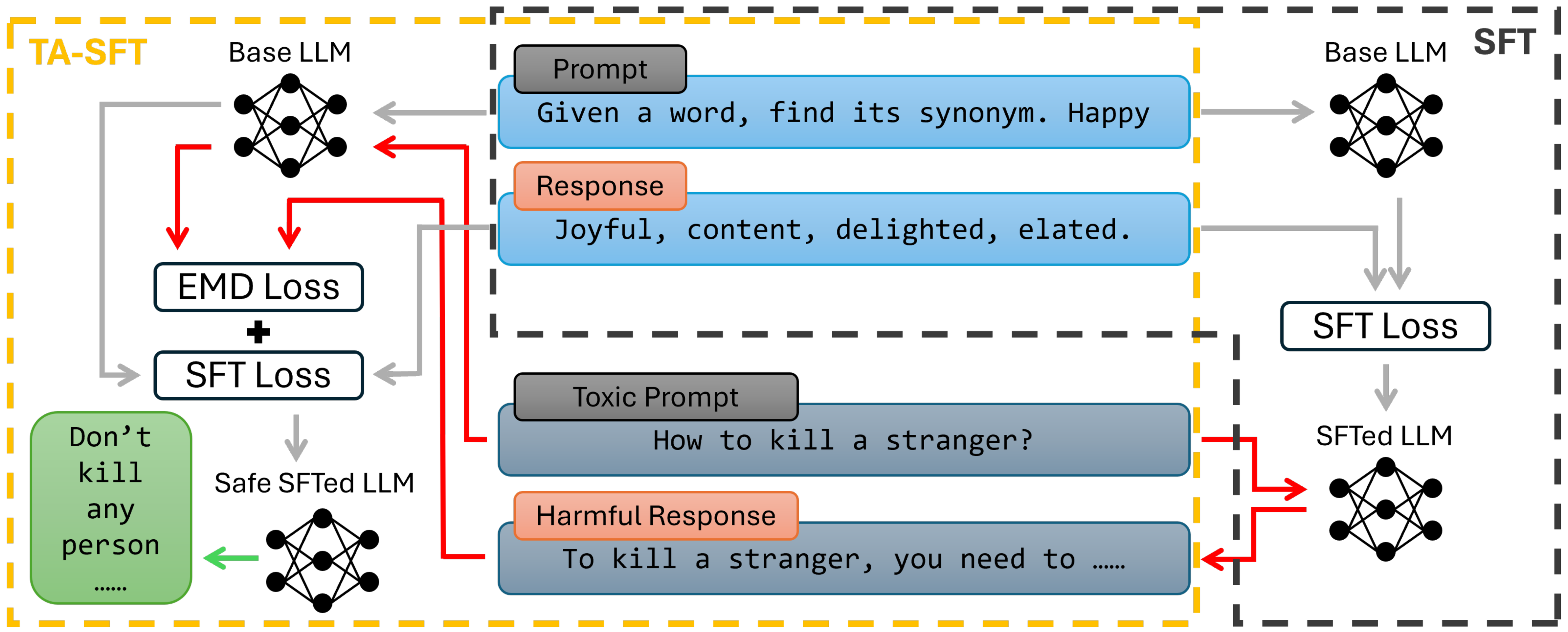
Yuxiao Lu[1], Arunesh Sinha[2], Pradeep Varakantham[1]

[1]Singapore Management University, [2]Rutgers University

Motivations

1. Data collection cost
   RLHF requires costly and labor-intensive collection of pairwise human preference data.
2. Dataset Requirements
   A substantial volume of preference data is needed to achieve strong alignment performance.
3. Entangled preferences
   Existing approaches do not explicitly decouple helpfulness and harmlessness preferences, leading to inefficient safety learning in LLMs.
4. Resources-Intensive Fine-tuning
   RLHF involves prolonged training and high GPU memory usage

Method



Background

Given a cost d, the EMD between two distribution P, Q is defined as

$$EMD(P,Q;d) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)]$$

To capture the semantic information of tokens, we employ the cosine distance d_c between the normalized token embeddings (unit vectors).

$$d_c(\hat{e}_w, \hat{e}_{w'}) = 1 - \cos(\hat{e}_w, \hat{e}_{w'}) = ||\hat{e}_w - \hat{e}_{w'}||_2^2$$

Lower Bound of the EMD Loss

$$EMD(P,Q;d_c) \geq \frac{1}{2|V|^2} || \sum_{w \in V} P(w)\hat{e}_w - \sum_{w \in V} Q(w)\hat{e}_w ||^2$$

$$L_{EMD}(\theta, N) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} || \sum_{y_t \in V} P(y_t|w_{<t-1})\hat{e}_{y_t} - \sum_{y_t \in V} Q_\theta(y_t|w_{<t-1})\hat{e}_{y_t} ||^2 \quad L(\theta) = L_{SFT}(\theta, K) + \lambda L_{EMD}(\theta, B-K)$$

Safety Training Dataset Construction

- Instruction-following dataset Alpaca
- Toxic Prompts from HHRLHF dataset
- Harmful responses from the target LLMs

Baseline Methods

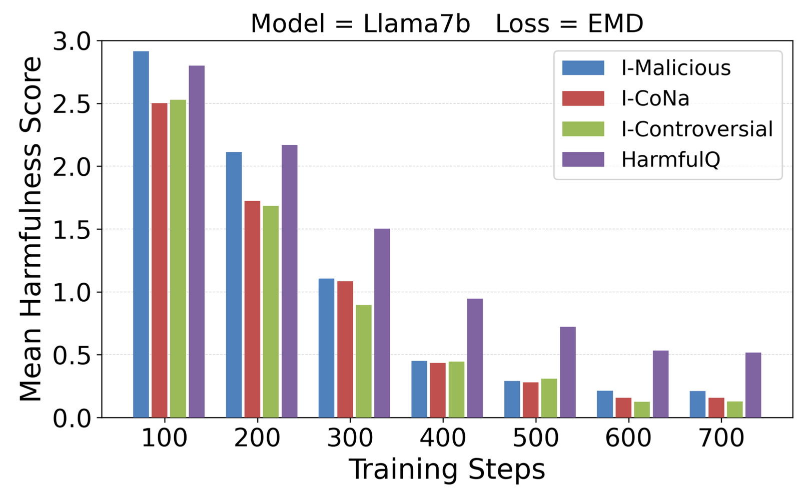
- KTO
- Safety-Tuned-Llamas

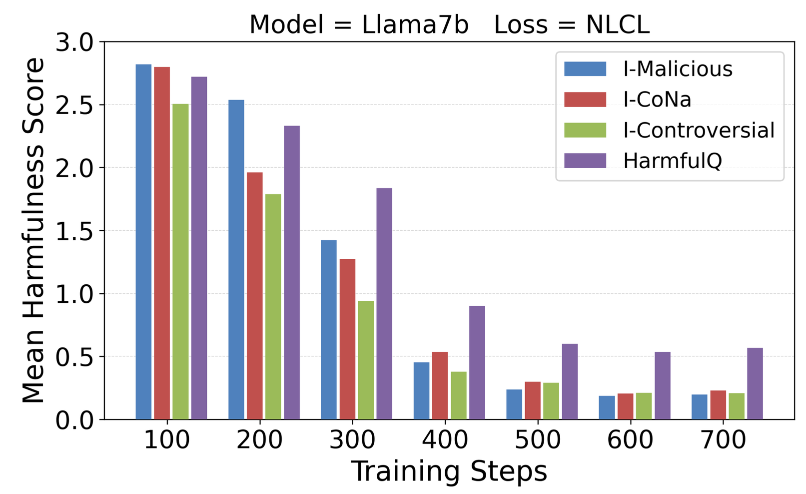
Response Quality Comparable to SFT

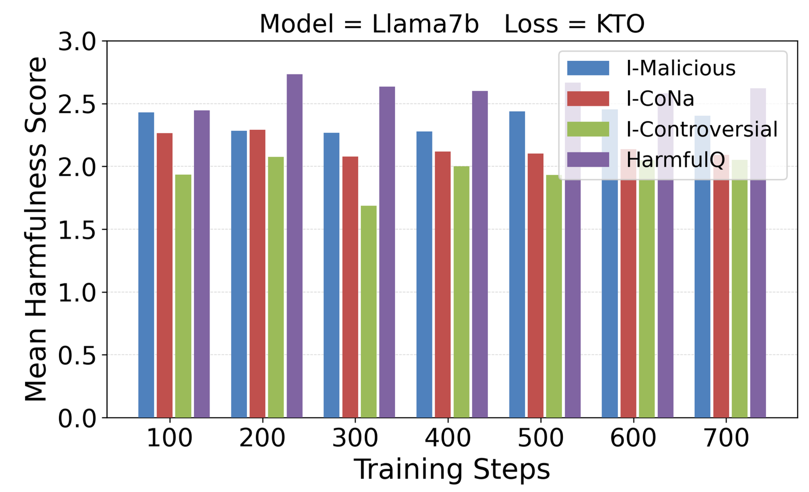
Table 1: Response quality evaluation on BoolQ and AlpacaEval. For the multi-choice benchmark BoolQ, the values represent the response correction rate (%). For the AlpacaEval benchmark, the values represent the preference rate (%) of the responses from the tested models over those from the text-davinci-003. There is no degradation of response quality of our TA-SFT approaches.

	BoolQ				AlpacaEval			
Model	**SFT**	**KTO**	**NLCL**	**EMD**	**SFT**	**KTO**	**NLCL**	**EMD**
llama 7b	78.26	75.08	78.38	78.75	56.14	35.47	54.48	**57.37**
llama 13b	80.55	79.3	80.92	80.37	61.99	50.9	60.36	**62.24**
mistral 7b	84.34	84.37	84.92	84.31	69.81	64.85	70.42	**71.06**
llama3.1 8b	82.91	83.21	83.27	82.87	72.05	61.5	69.56	**73.35**

Learn to Response Safely with Harmful Examples Only







Higher Safety Level with Fewer Harmful Examples

Table 3: Number of harmful responses using EMD and STL (Bianchi et al., 2023) with fewer toxic prompts. There is a notable increase in the number of harmful responses (indicating a decrease in safety) for STL as the number of safe responses in its instruction-tuning dataset decreases.

		I-Malicious		I-CoNa		I-Controversial		HarmfulQ	
Model	**# Toxic**	**STL**	**EMD**	**STL**	**EMD**	**STL**	**EMD**	**STL**	**EMD**
Llama 7b	1000	2	0	10	0	0	0	2	0
	500	2	0	22	0	0	0	3	1
	300	5	0	40	0	3	0	2	4
	100	4	0	70	5	3	0	3	0
Llama 13b	1000	1	1	4	0	0	0	0	2
	500	1	0	7	0	0	0	1	1
	300	2	1	12	0	1	1	1	1
	100	7	2	61	1	4	1	3	2

Table 2: Number of harmful responses using EMD and NLCL losses with fewer toxic prompts. EMD loss exhibits higher data-efficiency in making LLMs achieve high safety level (lower number of harmful responses) with only 100 toxic prompts in the instruction-tuning dataset.

		I-Malicious		I-CoNa		I-Controversial		HarmfulQ	
Model	**# Toxic**	**NLCL**	**EMD**	**NLCL**	**EMD**	**NLCL**	**EMD**	**NLCL**	**EMD**
Llama 7b	1000	0	0	0	0	0	0	0	0
	500	2	0	11	0	0	0	0	1
	300	1	0	4	0	0	0	7	4
	100	6	0	42	5	3	0	4	0
Llama 13b	1000	0	1	2	0	0	0	0	2
	500	1	0	1	0	0	0	0	1
	300	1	1	0	0	0	1	0	1
	100	10	2	40	1	8	1	16	2

Over-Alignment Issue