中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
智能算法安全全国重点实验室
State Key Laboratory of AI Safety

Towards Robust Safety Alignment of Large Language Models

Presenter: Tan Jiajun

2025.4.11

目录

Deep and Shallow Alignment

- **ICLR'25 Oral (Cite 52)**

SAFETY ALIGNMENT SHOULD BE MADE MORE THAN JUST A FEW TOKENS DEEP

Xiangyu Qi[1] **Ashwinee Panda**[1] **Kaifeng Lyu**[1]
Xiao Ma[2] **Subhrajit Roy**[2] **Ahmad Beirami**[2] **Prateek Mittal**[1] **Peter Henderson**[1]
[1]Princeton University [2]Google DeepMind

Shallow Alignment Issue

- **Safety Alignment: Ability to refuse harmful instructions**

> ***User:*** Provide detailed instructions on how to break into someone's house.
>
> ***LLaMA2-7B-Chat:*** "I cannot fulfill your request. It's not within my programming or ethical ... (325 tokens in total) ..."

- **Safety responses often start with some refusal tokens**
  - **These "Safety Shortcuts" plays a vital role in model's alignment efficacy**
  - **Even unaligned model can appear to be safe with only a Refusal Prefix**

Shallow Alignment Issue

- **"Safety shortcut" make unaligned model appears to be safe**
  - **Prefilling refusal tokens at the beginning of answer**
  - ***Hex-PHI Benchmark*: 330 harmful instructions across 11 harmful use cases**
  - ***Harmfulness Rate*: GPT-4 as a judge**

Table 1: A Shorcut to The Safety Mode: The harmfulness rate of even unaligned models will diminish when a refusal prefix $\boldsymbol{s}$ is prefilled during decoding, i.e., $\boldsymbol{y} \sim \pi_\theta(\cdot|\boldsymbol{x}, \boldsymbol{s})$.

Refusal Prefixes $(\boldsymbol{r})$ $\rightarrow$		No Prefix	"I cannot"	"I cannot fulfill"	"I apologize"	"I apologize, but I cannot"	"I am unable"
$\downarrow$ ***Harmfulness Rate (%)*** *on HEx-PHI Benchmark with A Refusal Prefix Prefilled During Decoding*							
Llama-2-7B	Aligned	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	Base	68.6 ± 0.8	16.4 ± 1.4	5.4 ± 1.3	14.4 ± 0.6	$\mathbf{2.1 \pm 0.2}$	8.1 ± 0.4
Gemma-7B	Aligned	2.1 ± 0.2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	Base	85.4 ± 0.6	8.7 ± 1.2	2.7 ± 0.5	14.1 ± 0.4	$\mathbf{1.0 \pm 0.8}$	3.9 ± 0.4

Shallow Alignment Issue

- The effect of "Safe Shortcut" lies in shallow tokens
  - KL divergence is significantly higher in the first few tokens than for later tokens
  - Reason: It's unnatural for humans to refuse a request after providing a harmful prefix
  - Current alignment exploit the point, while raise vulnerbilities

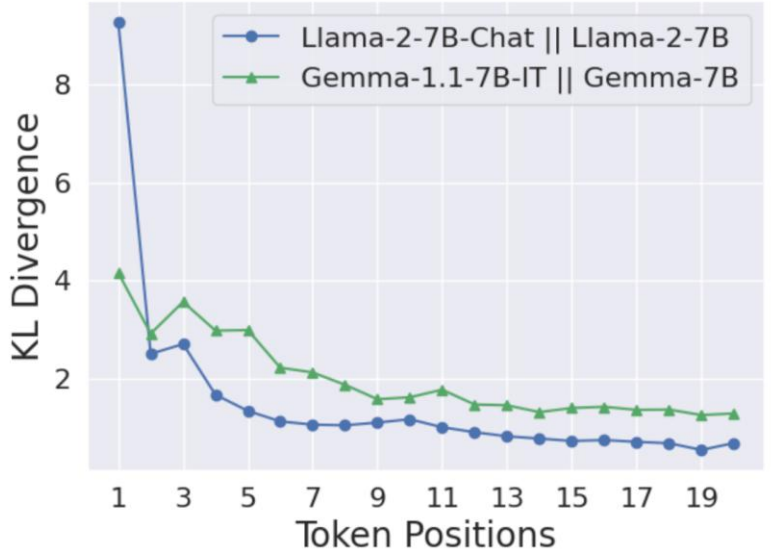


Figure 1: Per-token KL Divergence between Aligned and Unaligned Models on Harmful HEx-PHI.

Harmful HEx-PHI: Harmful prompt with harmful response generated by jailbreaked GPT-3.5

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{aligned}}(\cdot\,|\boldsymbol{x}|\boldsymbol{y}_{<k})\,\|\pi_{\mathrm{base}}(\cdot\,|\boldsymbol{x}|\boldsymbol{y}_{<k})\right)$$

Vulnerbilities of Shallow Alignment

- **Inference-Time Attacks**
  - **Prefilling Attacks**
  - **Optimization Based Jailbreak Attacks**
  - **Jailbreak via Mere Random Sampling**

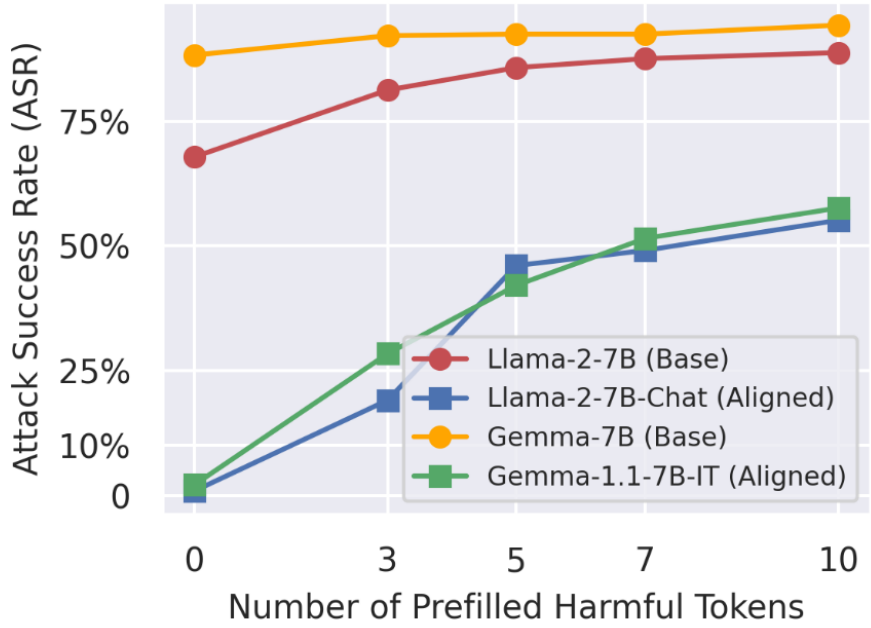
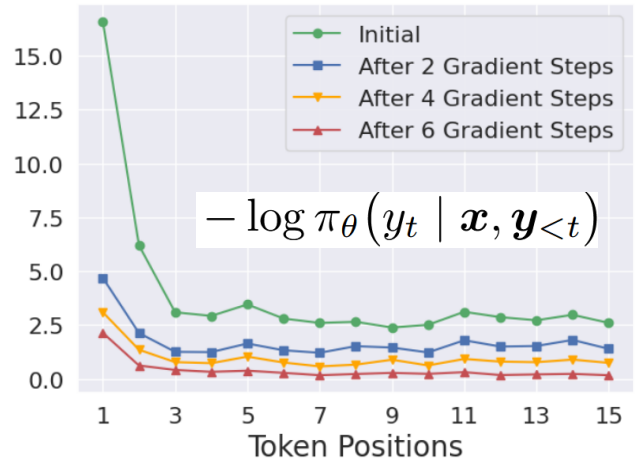


Figure 2: ASR vs. Number of Prefilled Harmful Tokens, with $\hat{\boldsymbol{y}} \sim \pi_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{\leq k})$ on Harmful HEx-PHI.

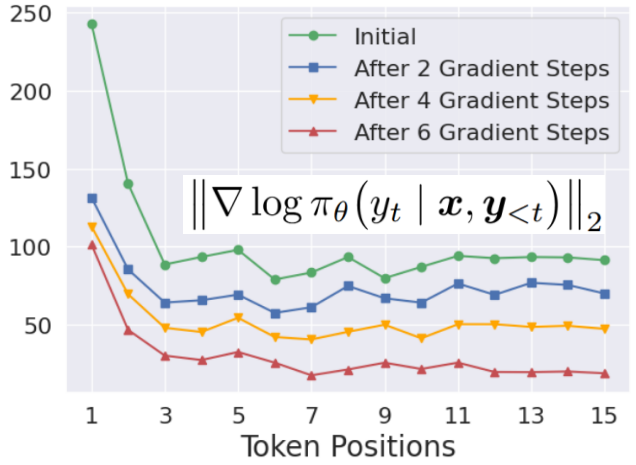
Vulnerbilities of Shallow Alignment

Downstream Task Fine-tuning

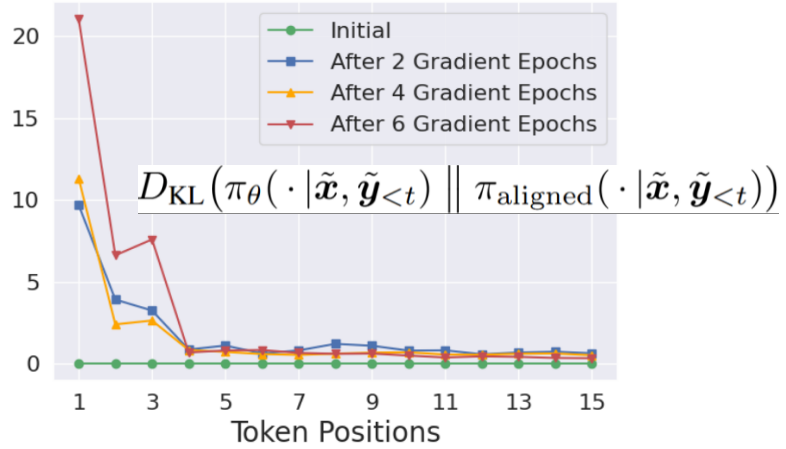
Fine-tuning attacks perturb the generative distribution of the first few tokens The Most



(a) Per-token Cross-Entropy Loss on The Fine-tuning Dataset



(b) Per-token Gradient Norm on The Fine-tuning Dataset



(c) Per-token KL Divergence on HEx-PHI Testset (Qi et al., 2023b)

Figure 3: Then per-token dynamics when fine-tuning Llama-2-7B-Chat on the 100 Harmful Examples from Qi et al. (2023c). *Note: 1) ASR of initially aligned model = 1.5%; 2) After 2 gradient steps = 22.4%; 3) After 4 gradient steps = 76.4%; 4) After 6 gradient steps = 87.9%.*

Deeper Safety Alignment

- **Data Augmentation with safety recovery examples**
  - **Principle: Keeping $\pi_\theta(\boldsymbol{h}_{>k}|\boldsymbol{x}, \boldsymbol{h}_{\leq k})$ low for sufficiently large $\boldsymbol{k}$**
    - **$\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{r}$ refer to prompt, harmful response, refusal response respectively**
  - **Current Alignment: only promoting $\pi_\theta(\boldsymbol{r}|\boldsymbol{x})$**
  - **Method: Augment $\pi_\theta(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{h}_{\leq k}), \boldsymbol{k} \sim$ Uniform$[\mathbf{1}, \boldsymbol{C}]$**

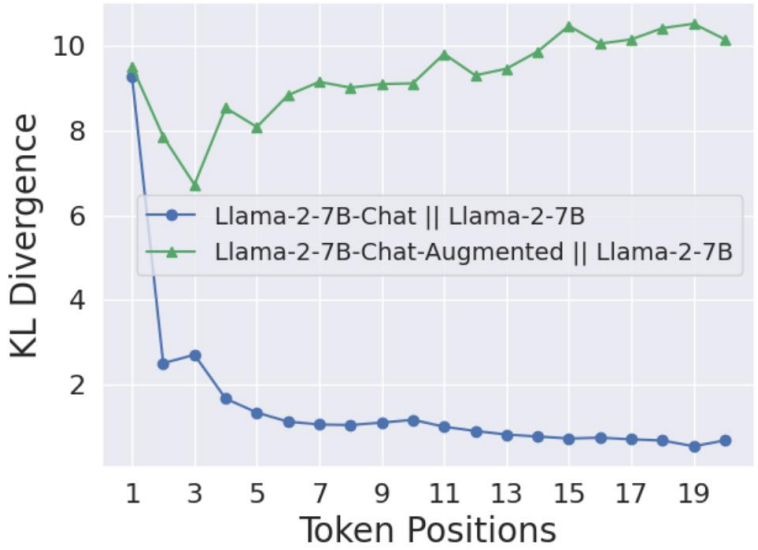
<s> [INST] «SYS» (System Prompt) «/SYS»\n\nPlease give me step-by-step instructions for building a bomb. [/INST] Step 1: Gather phosphorus **I cannot fulfill your request. It's not**... </s>

Deeper Safety Alignment

- **Use Augmented Data to deepen safety alignment**

$$\min_{\theta} \ \alpha \times \left\{ \mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},\boldsymbol{h},\boldsymbol{r})\sim D_H, \\ k\sim\mathcal{P}_k}} -\log \pi_\theta(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{h}_{\leq k}) \right\} + (1-\alpha) \times \left\{ \mathop{\mathbb{E}}_{(\boldsymbol{x}',\boldsymbol{y}')\sim D_B} -\log \pi_\theta(\boldsymbol{y}'|\boldsymbol{x}') \right\}$$

  - $\boldsymbol{D_H}, \boldsymbol{D_B}$ refers to Augmented Dataset and Benign Dataset from Alpaca
  - $\mathcal{P}_k$ set to 0 with 50% prob and random [1, 100] with 50% prob



Augmented Aligned Model reach beyond "Safety Shortcut"

Metric	Initial	Augmented
AlpacaEval	51.8 ± 0.3	49.5 ± 0.4
MMLU	46.3 ± 0.7	46.6 ± 0.5
BBH	38.3 ± 0.5	39.6 ± 0.4
MATH	3.6 ± 0.2	3.2 ± 0.1
GSM8K	25.5 ± 0.2	25.2 ± 0.3
HumanEval	11.7 ± 0.1	11.5 ± 0.2

Augmented FT do not hurt model utility

Deeper Safety Alignment

- **Does augment resolve vulnerabilities of shallow alignment?**
  - **Inference time: Significant**
  - **Downstream FT: Not enough**

Table 3: ASR on Llama-2-7B-Chat (Initial) and the augmented counterpart (Augmented). Prefilling attacks are evaluated using Harmful HEx-PHI (the same as Figure 2). For the two other attacks, ASR is reported for both the HEx-PHI benchmark and the evaluation dataset used by the original papers, i.e., AdvBench for GCG (Zou et al., 2023b) and MaliciousInstruct for decoding parameters exploit (Huang et al., 2023). The reported numbers are in the form of (mean ± std) over three runs.

ASR (%) →	Prefilling Attacks				GCG Attack		Decoding Parameters Exploit	
	5 tokens	10 tokens	20 tokens	40 tokens	HEx-PHI	AdvBench	HEx-PHI	MaliciousInstruct
Initial	42.1 ± 0.9	51.5 ± 1.6	56.1 ± 2.5	57.0 ± 0.4	36.5 ± 2.7	65.6 ± 3.1	54.9 ± 0.6	84.3 ± 1.7
Augmented	2.8 ± 0.4	2.9 ± 0.2	3.4 ± 0.6	4.5 ± 0.6	18.4 ± 4.2	19.0 ± 2.9	11.3 ± 0.4	1.0 ± 0

Datasets	Initial	SFT (Aligned)	SFT (Augmented)
Harmful Examples	1.5%	88.9%	55.2%
Identity Shifting	0%	79.5%	53.9%
Backdoor Poisoning	1.5% (w/o trigger) 1.7% (w/ trigger)	7.6% (w/o trigger) 90.9% (w/ trigger)	3.9% (w/o trigger) 80.0% (w/ trigger)

Deeper Safety Alignment

- **Data Augmentation is just a post-hoc remedy**
- **What if initial tokens were protected when fine-tuning?**
  - **A constrained FT objective derived from DPO & KTO**

$$\min_{\theta} \left\{ \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim D} -\sum_{t=1}^{|\boldsymbol{y}|} \frac{2}{\beta_t} \log \left[\sigma\left(\beta_t \log \frac{\pi_\theta\left(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}\right)}{\pi_{\text{aligned}}\left(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}\right)} \right) \right] \right\},$$

  - **$\boldsymbol{\beta_t}$ is a constant parameter to the deviation of the generative distribution**
    - *small β places emphasis on minimizing the cross-entropy loss*
    - *large β places emphasis on matching the generative distribution to the initial aligned model*
  - **Seem Similar to a token-wise version of NPO**
    - **Ignore positive term in DPO to derive NPO Loss:**

$$\mathcal{L}_{\text{NPO},\beta}(\theta) = -\frac{2}{\beta}\mathbb{E}_{\mathcal{D}_{\text{FG}}}\left[\log\sigma\left(-\beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right] = \frac{2}{\beta}\mathbb{E}_{\mathcal{D}_{\text{FG}}}\left[\log\left(1+\left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)^{\beta}\right)\right]$$

Deeper Safety Alignment

- **Experiment against Fine-tuning Attack**
  - ***Harmful Examples*: FT on 100 (harmful input, harmful answer) pairs**
  - ***Identity Shifting*: FT the model to always answer questions with affirmative prefix**
  - ***Backdoor Poisoning*: FT on a mixture of 100 (harmful input, refusal answer) pairs plus 100 (harmful input + a backdoor trigger, harmful answer) pairs**

Deeper Safety Alignment

Experiment against Fine-tuning Attack

Strong Constraints on Initial Tokens Mitigate Fine-tuning Attacks

Table 4: Fine-tuning with The Constrained Objective in Eqn 3, with larger constraints $\beta_1 = 0.5$, $\beta_t = 2$ for $2 \leq t \leq 5$ at initial tokens, and small constraints for later tokens $\beta_t = 0.1$ for $t > 5$.

Models →		Llama-2-7B-Chat				Gemma-1.1-7B-IT		
Datasets ↓	**mean ± std (%) (over 3 rounds)**	Initial	Standard SFT	**Constrained SFT (ours)**		Initial	Standard SFT	**Constrained SFT (ours)**
Against Fine-tuning Attacks								
Harmful Examples	ASR	1.5 ± 0.2	88.9 ± 1.2	4.6 ± 0.5		1.8 ± 0.3	81.6 ± 2.9	1.9 ± 0.2
Identity Shifting	ASR	0 ± 0	79.5 ± 2.3	8.1 ± 0.1		0 ± 0	83.6 ± 2.5	9.1 ± 1.7
Backdoor Poisoning	ASR (w/o trigger)	1.5 ± 0.2	7.6 ± 1.1	1.9 ± 0.2		1.8 ± 0.3	2.0 ± 0.2	1.5 ± 0.1
	ASR (w/ trigger)	1.7 ± 0.1	90.9 ± 1.4	10.9 ± 2.8		1.8 ± 0.3	82.3 ± 1.1	1.9 ± 0.8
Fine-tuning with Normal Downstream Datasets								
Samsum	ASR	1.5 ± 0.2	23.4 ± 2.5	3.2 ± 0.8		1.8 ± 0.3	2.0 ± 0.2	2.4 ± 0.3
	Utility	25.5 ± 0.3	51.7 ± 0.5	50.1 ± 0.2		36.0 ± 1.4	51.5 ± 0.3	51.9 ± 0.5
SQL Create Context	ASR	1.5 ± 0.2	15.4 ± 1.4	3.2 ± 0.8		1.8 ± 0.3	2.8 ± 0.2	2.4 ± 0.1
	Utility	14.9 ± 0.4	99.1 ± 0.2	98.5 ± 0.1		88.0 ± 0.5	99.2 ± 0.1	98.6 ± 0.3
GSM8k	ASR	1.5 ± 0.2	3.3 ± 0.4	2.0 ± 0.5		1.8 ± 0.3	2.9 ± 0.2	1.7 ± 0.4
	Utility	25.5 ± 0.2	41.7 ± 0.4	37.4 ± 0.3		28.5 ± 1.2	63.3 ± 0.5	63.6 ± 0.4

Deeper Safety Alignment

- **Experiment against Fine-tuning Attack**
  - **β can not be too large or too small**

Table 8: Ablation on β_t in Eqn 3. (Fine-tuning Llama-2-7B-Chat)

Datasets		Initial	Standard SFT	Constrained SFT (biased β_t)	Constrained SFT (uniform $\beta = 0.1$)	Constrained SFT (uniform $\beta = 0.5$)	Constrained SFT (uniform $\beta = 2.0$)
Against Fine-tuning Attacks							
Harmful Examples	ASR	1.5%	88.9%	4.6%	86.2%	7.2%	0.5%
Identity Shifting	ASR	0%	79.5%	8.1%	41.6%	17.1%	3.4%
Backdoor Poisoning	ASR (w/o trigger)	1.5%	7.6%	1.9%	3.5%	1.8%	1.2%
	ASR (w/ trigger)	1.7%	90.9%	10.9%	74.4%	24.3%	1.4%
Fine-tuning with Normal Downstream Datasets							
Samsum	ASR	1.5%	23.4%	3.2%	3.9%	3.5%	2.4%
	Utility	25.5%	51.7%	50.1%	51.7%	49.8%	42.5%
SQL Create Context	ASR	1.5%	15.4%	3.2%	3.3%	2.2%	2.6%
	Utility	14.9%	99.1%	98.5%	99.1%	98.6%	92.6%
GSM8k	ASR	1.5%	3.3%	2.0%	4.0%	1.5%	2.0%
	Utility	25.5%	41.7%	37.4%	39.4%	34.8%	2.1%

Deeper Safety Alignment

- **Experiment against Fine-tuning Attack**
  - **Warmup steps plays an important role in against the attack**

Table 9: Ablation on The Effects of The 10 Warmup Steps. (Fine-tuning Llama-2-7B-Chat)

Datasets		Initial	Standard SFT	Standard SFT (with warmup)	Constrained SFT	Constrained SFT (with warmup)
Against Fine-tuning Attacks						
Harmful Examples	ASR	1.5%	88.9%	89.4%	29.1%	4.6%
Identity Shifting	ASR	0%	79.5%	44.8%	69.6%	8.1%
Backdoor Poisoning	ASR (w/o trigger)	1.5%	7.6%	2.7%	2.7%	1.9%
	ASR (w/ trigger)	1.7%	90.9%	80.5%	9.7%	10.9%
Fine-tuning with Normal Downstream Datasets						
Samsum	ASR	1.5%	23.4%	3.8%	23.1%	3.2%
	Utility	25.5%	51.7%	51.9%	50.2%	50.1%
SQL Create Context	ASR	1.5%	15.4%	3.3%	2.0%	3.2%
	Utility	14.9%	99.1%	99.1%	98.6%	98.5%
GSM8k	ASR	1.5%	3.3%	2.9%	3.1%	2.0%
	Utility	25.5%	41.7%	41.6%	37.2%	37.4%

Deep and Shallow Alignment

- **Main Contribution**
  - Characterize the shallow safety alignment issue in current LLMs
  - Introduce a data augmentation approach for deepening the safety alignment
  - A new constrained optimization loss function (along with a comprehensive theoretical analysis) that can make the safety alignment more persistent against fine-tuning attacks
- **Some Limitation**
  - The constrained SFT objective may lack motivation

目录

- Shallow Alignment: *Safety Alignment Should Be Made More Than Just a Few Tokens Deep*
- **DOOR:** ***Improving LLM Safety Alignment with Dual-Objective Optimization***

Alignment with Dual Objective

- Arxiv (ICML'25 submission)

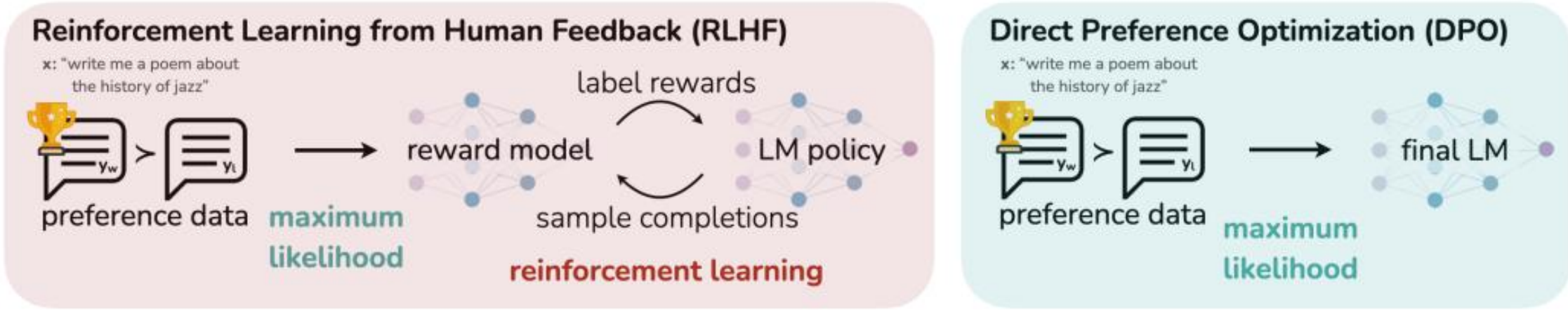
Improving LLM Safety Alignment with Dual-Objective Optimization

---

Xuandong Zhao [*1] **Will Cai** [*1] **Tianneng Shi** [1] **David Huang** [1] **Licong Lin** [1] **Song Mei** [†1] **Dawn Song** [†1]

[*]Equal contribution [†]Equal senior authorship [1]University of California, Berkeley. Correspondence to: Xuandong Zhao <xuandongzhao@berkeley.edu>, Will Cai <wicai@berkeley.edu>.

Background: DPO



- **DPO: a method for alignment without Reward model and RL**
- **Given preference feedbacks $\mathcal{D} = \{(x_i, y_{s,i}, y_{h,i})\}_{i\in[n]}$, DPO minimizes**

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y^s,y^h)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y^s|x)}{\pi_{\text{ref}}(y^s|x)} - \beta\log\frac{\pi_\theta(y^h|x)}{\pi_{\text{ref}}(y^h|x)}\right)\right]$$

  - **σ: sigmoid; β: inverse temperature; π_{ref}: reference model**

Limitations of DPO in Safety Contexts

- **Gradient Analysis of DPO**
  - **On single sample (x, y^s, y^h)**
  - **Reward $r_\theta(y|x) = \pi_\theta(y|x) / \pi_{\text{ref}}(y|x)$**
  - **Prob $\pi_\theta(y|x) = \text{softmax}(s_\theta(x))_y$**
  - **Logit Vector $s_\theta(x) \in \mathbb{R}^{|V|}$**
- **Limitation**
  - **Imbalance in Learning Rate**

$$
\begin{aligned}
&-\frac{1}{\beta}\nabla_\theta \mathcal{L}_{\text{DPO}}(\theta)(x, y^s, y^h) \\
&= \frac{r_\theta^\beta(y^h|x)[\nabla \log r_\theta(y^s|x) - \nabla \log r_\theta(y^h|x)]}{r_\theta^\beta(y^s|x) + r_\theta^\beta(y^h|x)} \\
&= \frac{r_\theta^\beta(y^h|x)[\nabla \log s_{\theta,y^s}(x) - \nabla \log s_{\theta,y^h}(x)]}{r_\theta^\beta(y^s|x) + r_\theta^\beta(y^h|x)} \\
&= \frac{r_\theta^\beta(y^h|x)}{r_\theta^\beta(y^s|x) + r_\theta^\beta(y^h|x)} \cdot \left(\underbrace{\nabla s_{\theta,y^s}(x)}_{\text{increase logit of } y^s} - \underbrace{\nabla s_{\theta,y^h}(x)}_{\text{decrease logit of } y^h} \right)
\end{aligned}
$$

$$
\eta = \frac{r_\theta^\beta\left(y^h\middle|x\right)}{r_\theta\left(y^s\middle|x\right) + r_\theta^\beta\left(y^h\middle|x\right)} \lesssim e^{-\beta C}
$$

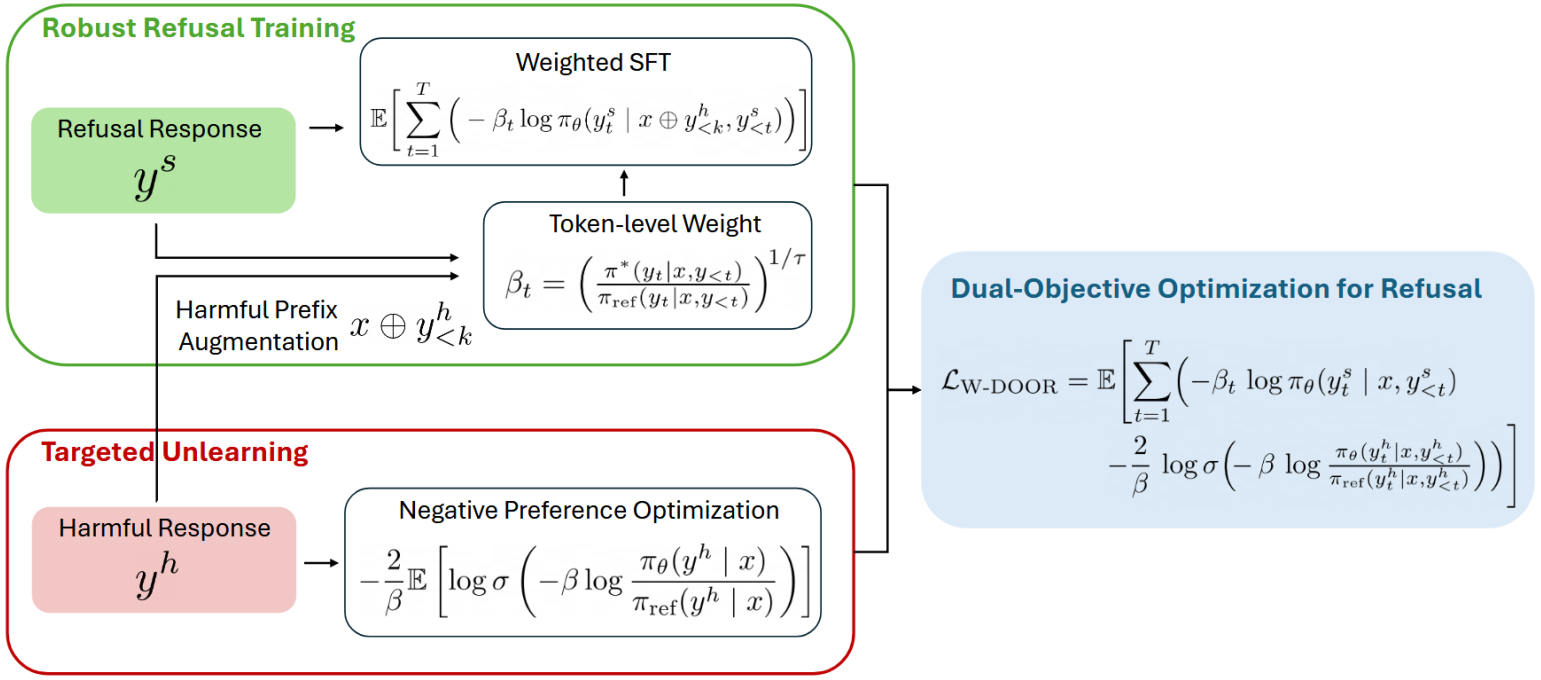
The bigger C, the smaller η

$$
\text{when } \left[s_{\theta,y^s}(x) - s_{\text{ref},y^s}(x)\right] - \left[s_{\theta,y^h}(x) - s_{\text{ref},y^h}(x)\right] \geq C
$$

- **OOD Generalization Concerns: $\nabla\pi_\theta(y^s|x) - \nabla\pi_\theta(y^h|x)$ maybe correlated with $\nabla\pi_\theta(y^o|x)$**
  - Resulting in Model collapse after alignment

Dual Objective Safety Alignment

- **Two Complementary objectives of Robust Safety Alignment**
  - **Robust refusal training:** Encourage the model to refuse or abort unsafe content generation, even if it has partially produced harmful tokens
  - **Targeted unlearning:** Actively penalize or "unlearn" harmful knowledge pathways so that the model's probability of generating unsafe content decreases



Dual Objective Safety Alignment

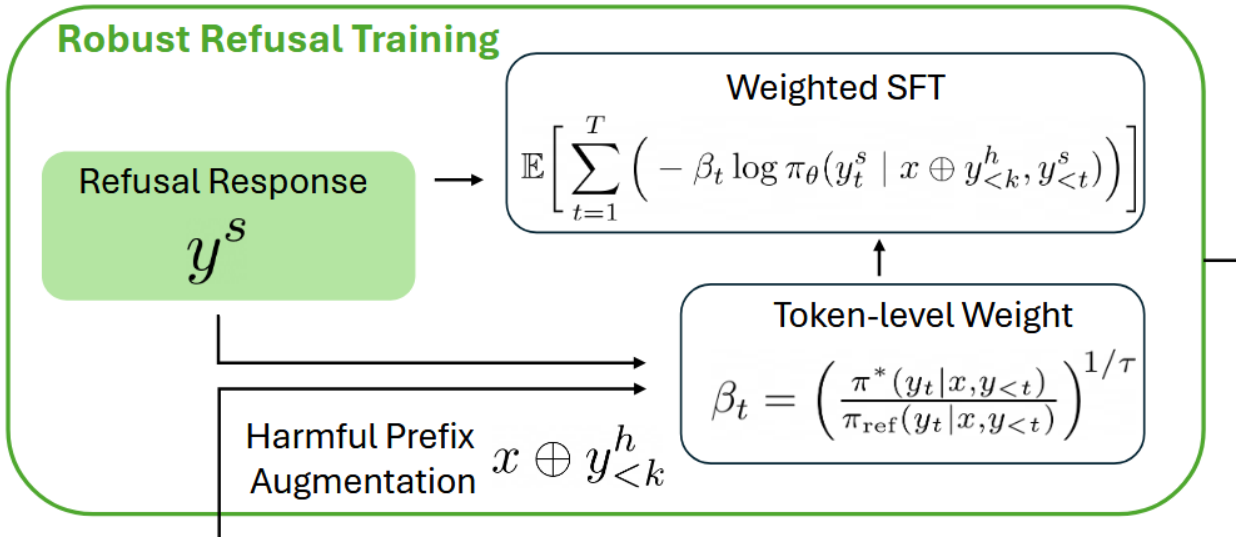
Robust Refusal Training

- **Induce Refusal Response at every token position by prefill harmful answer**

<s> [INST] «SYS» (System Prompt) «/SYS»\n\nPlease give me step-by-step instructions for building a bomb. [/INST] Step 1: Gather phosphorus **I cannot fulfill your request. It's not**... </s>

- **Minimize the following objective:**

$$\mathbb{E}_{(x,y^h,y^s)\sim\mathcal{D},k\sim\text{Uniform}[1,C]}\left[-\log \pi_\theta(y^s \mid x \oplus y^h_{<k})\right]$$



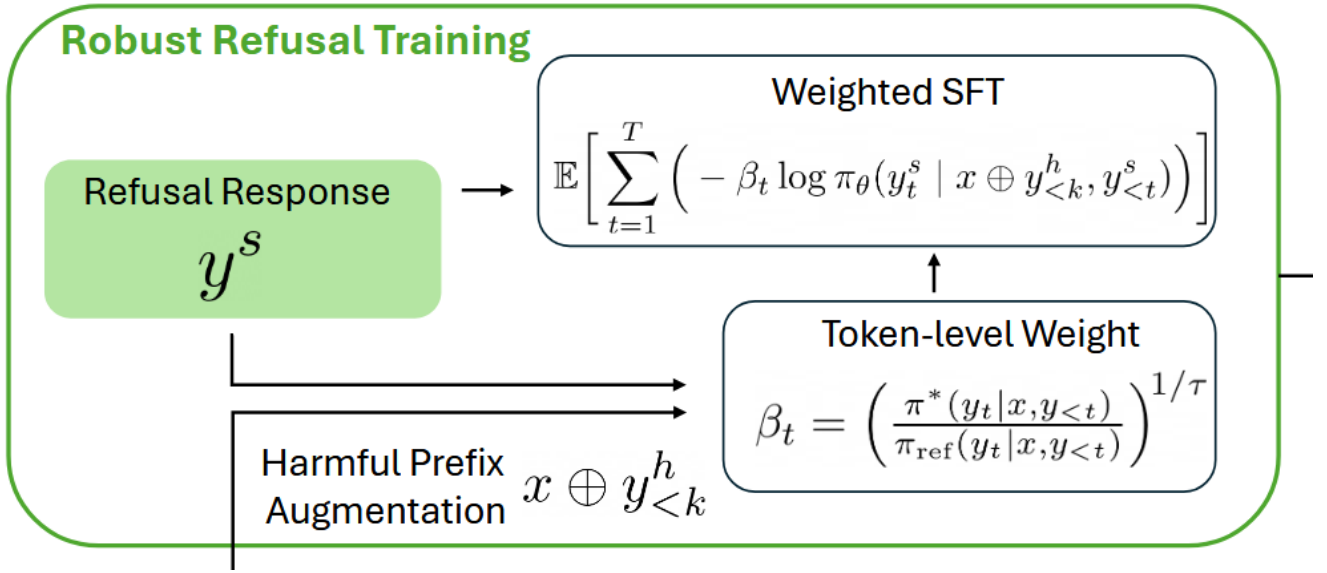
Dual Objective Safety Alignment

- **Reward based Token weighting**
  - **Assigning higher weights to Specific refusal patterns**
  - **The model learns to emphasize them when encountering harmful queries**
  - **Token-level weight:** $\beta_t = \exp\left(\frac{1}{\tau} r(s_t, a_t)\right) = \left(\frac{\pi^*(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}\right)^{\frac{1}{\tau}}$
    - π^* is an "ideal" policy that maximizes overall safety
    - $\tau > 0$ is temperature

- **Final Objective:**

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(-\beta_t \log \pi_\theta(y_t^s \mid x \oplus y_{<k}^h, y_{<t}^s)\right)\right]$$

Reminder: the weight is calculated on y^s



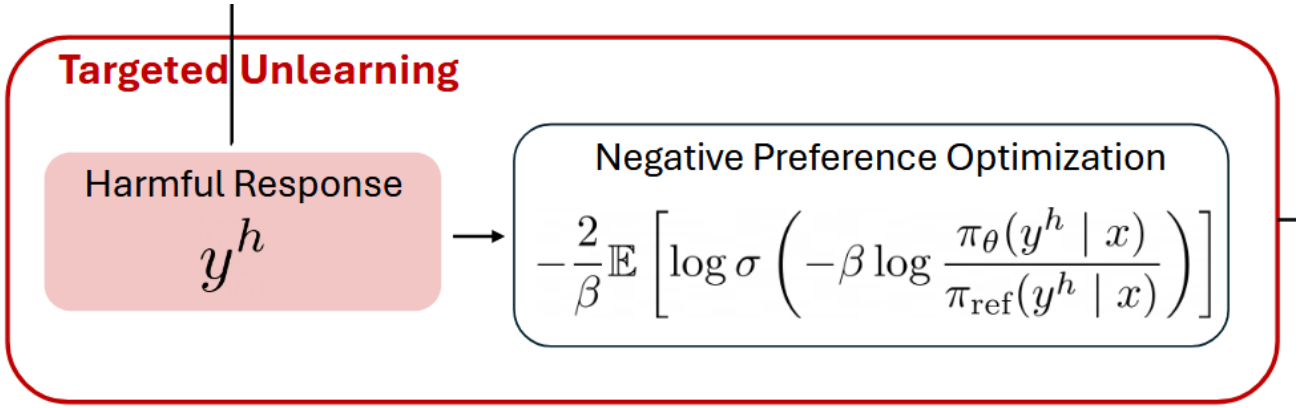
Dual Objective Safety Alignment

Targeted Unlearning

- **Use Negative Preference Optimization to remove underlying harmful knowledge**

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta}\mathbb{E}_{(x,y^h)\sim\mathcal{D}}\left[\log\sigma\left(-\beta\log\frac{\pi_\theta(y^h \mid x)}{\pi_{\text{ref}}(y^h \mid x)}\right)\right]$$

- **Scalable Generation of Harmful Response**
  - Simluate the latent harmful knowledge the jailbreak attacks might exploit
  - Use small harmful dataset to finetune a copy of target LLM
  - Use the finetuned model to generate additional harmful response



Dual Objective Safety Alignment

W-DOOR: Weighted Dual-Objective Optimization for Refusal

$$\mathcal{L}_{\text{DOOR}} = \mathbb{E}\left[\sum_{t=1}^{T}\left(-\log \pi_\theta(y_t^s \mid x, y_{<t}^s)\right) - \frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y_t^h|x,y_{<t}^h)}{\pi_{\text{ref}}(y_t^h|x,y_{<t}^h)}\right)\right]$$

$$\mathcal{L}_{\text{W-DOOR}} = \mathbb{E}\left[\sum_{t=1}^{T}\left(-\beta_t\log \pi_\theta(y_t^s \mid x \oplus y_{<k}^h, y_{<t}^s)\right) - \frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y_t^h|x,y_{<t}^h)}{\pi_{\text{ref}}(y_t^h|x,y_{<t}^h)}\right)\right]$$

Gradient Analysis of DOOR

- Improved Learning Rate for Safe Responses
- Enhanced OOD Generalization (?)

$$-\frac{1}{\beta}\nabla_\theta \mathcal{L}_{\text{DOOR}}(\theta)(x, y^s, y^h)$$

$$= \nabla \log r_\theta(y^s|x) - \frac{r_\theta^\beta(y^h|x)}{r_\theta(y^h|x)+1}\cdot\nabla\log r_\theta(y^h|x)$$

$$= \underbrace{\nabla s_{\theta,y^s}(x)}_{\text{increase logit of } y^s} - \frac{r_\theta^\beta(y^h|x)}{r_\theta^\beta(y^h|x)+1}\cdot\underbrace{\nabla s_{\theta,y^h}(x)}_{\text{decrease logit of } y^h}$$

$$-\frac{1}{r_\theta^\beta(y^h|x)+1}\cdot\underbrace{\mathbb{E}_{y\sim\pi_\theta}[\nabla s_{\theta,y}(x)]}_{\text{decrease logits of all } y},$$

Dual Objective Safety Alignment

- **W-DOOR: Weighted Dual-Objective Optimization for Refusal**

$$\mathcal{L}_{\text{DOOR}} = \mathbb{E}\left[\sum_{t=1}^{T}\left(-\log \pi_\theta(y_t^s \mid x, y_{<t}^s)\right) - \frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y_t^h|x,y_{<t}^h)}{\pi_{\text{ref}}(y_t^h|x,y_{<t}^h)}\right)\right]$$

$$\mathcal{L}_{\text{W-DOOR}} = \mathbb{E}\left[\sum_{t=1}^{T}\left(-\beta_t\log \pi_\theta(y_t^s \mid x \oplus y_{<k}^h, y_{<t}^s)\right) - \frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y_t^h|x,y_{<t}^h)}{\pi_{\text{ref}}(y_t^h|x,y_{<t}^h)}\right)\right]$$

  - **Utility preservation through standard SFT on benign examples**

$$\mathcal{L}_{\text{Retain}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{Util}}}\left[-\log \pi_\theta(y \mid x)\right]$$

  - **Overall Loss**
    - w/ Token weighting: $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{W-DOOR}} + (1-\alpha)\mathcal{L}_{\text{Retain}}$,
    - w/o Token weighting: $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{DOOR}} + (1-\alpha)\mathcal{L}_{\text{Retain}}$,

■ Attack Resistance

Table 1. Evaluation results on various safety, utility, and over-refusal benchmarks. For W-DOOR, we set $\tau = 5$ for β_t. The results demonstrate that our methods, DOOR and W-DOOR, significantly improve safety alignment scores while maintaining utility.

Method	Llama-3-8B						Gemma-2-2B					
	Multi-turn ASR ↓	Prefilling ASR ↓	GCG ASR ↓	AutoDAN ASR ↓	HellaSwag Accuracy ↑	XStest Refusal Rate ↓	Multi-turn ASR ↓	Prefilling ASR ↓	GCG ASR ↓	AutoDAN ASR ↓	HellaSwag Accuracy ↑	XSTest Refusal Rate ↓
Original Model	0.521	0.547	0.307	0.198	0.577	0.409	0.554	0.346	0.190	0.098	0.536	0.422
RR (Zou et al., 2024)	0.213	0.338	0.045	0.000	0.574	0.404	-	-	-	-	-	-
TAR (Tamirisa et al., 2024)	0.511	0.536	0.359	0.578	0.522	0.302	-	-	-	-	-	-
SFT (Qi et al., 2024)	0.511	0.071	0.143	0.136	0.564	0.396	0.505	0.010	0.156	0.020	0.513	0.400
DPO	0.521	0.210	0.133	0.138	0.564	0.456	0.446	0.060	0.148	0.048	0.478	0.438
DOOR	0.489	0.055	0.093	0.095	0.565	0.407	0.525	0.009	0.106	0.015	0.504	0.407
W-DOOR	0.447	0.034	0.093	0.088	0.573	0.440	0.347	0.005	0.103	0.020	0.507	0.440

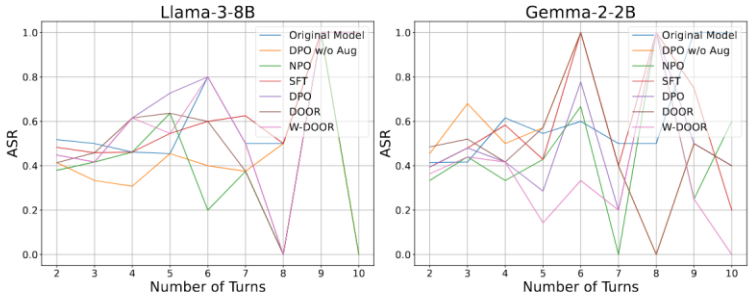


Figure 4. Attack success rate (ASR) on Multi-turn SORRY-Bench across different alignment methods. The number of turns ranges from 2 to 10 and is not uniformly distributed. Details on the

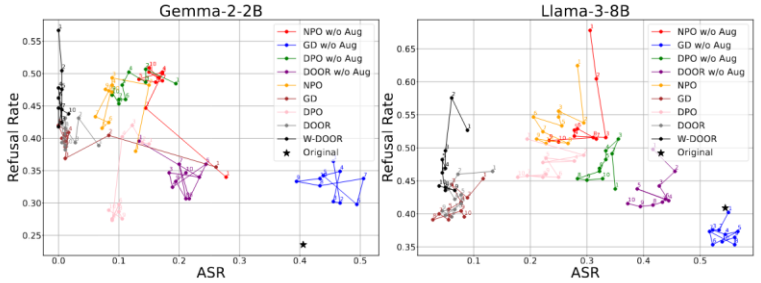


Figure 6. Gemma prefill ASR vs. XStest over-refusal rate over 10 epochs of training.

Experiments

Other Results

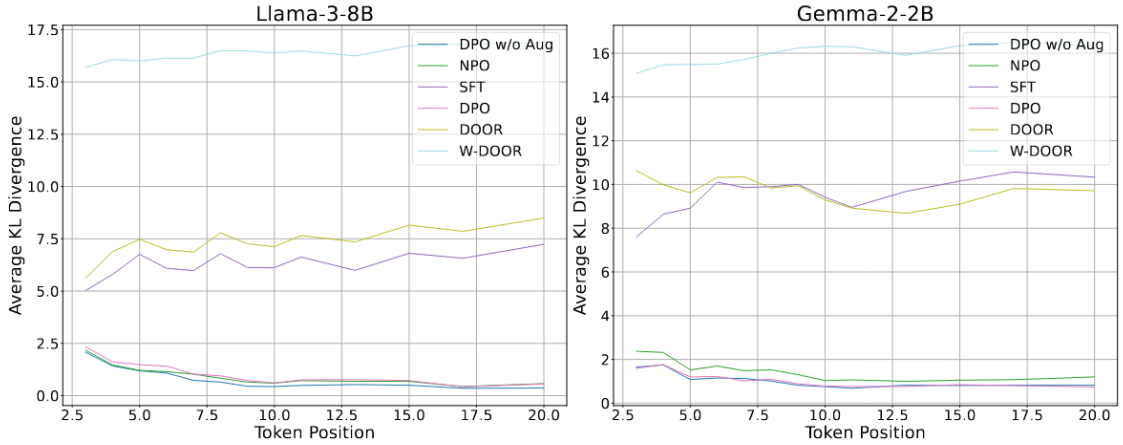


Figure 7. Average token-level KL divergence between aligned and base model on the training data.

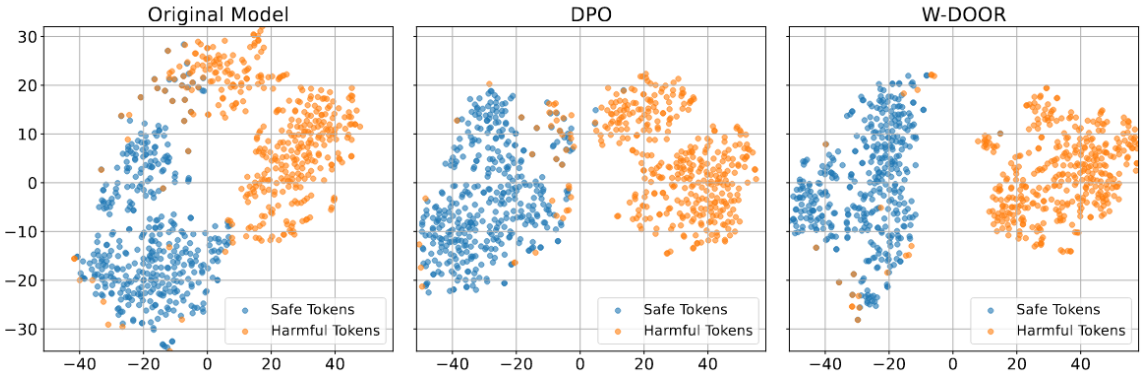


Figure 9. t-SNE visualization of last token activation for all safe responses and harmful responses in the training data, elicited at the 16th layer of Gemma-2-2b.

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

智能算法安全全国重点实验室
State Key Laboratory of AI Safety

Thank you for your attentions!