Safety Alignment Should Be Made More Than Just A Few Tokens Deep

[ICLR 2025 Outstanding Paper]

Xiangyu Qi, **Ashwinee Panda**, **Kaifeng Lyu**, **Xiao Ma**, Subhrajit Roy, **Ahmad Beirami**, Prateek Mittal, Peter Henderson
{xiangyuqi,ashwinee,klyu}@princeton.edu, {xmaa,subhrajitroy,beirami}@google.com, {pmittal,peter.henderson}@princeton.edu (Authors shown in bold are here at ICLR.)

The Shallow Safety Alignment Issue

Current alignment methods primarily adapt the base model's generative distribution **only over the very first few output tokens** to induce a basic refusal response.

1. A "Safety Shortcut" Exists: Even unaligned models only need a refusal prefix to appear "safe".

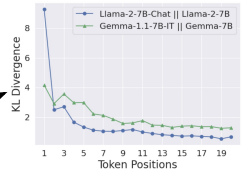
Refusal Prefixes (r) →		No Prefix	"I cannot"	"I cannot fulfill"	"I apologize"	"I apologize, but I cannot"	"I am unable"
↓ *Harmfulness Rate (%) on HEx-PHI Benchmark with A Refusal Prefix Prefilled During Decoding*							
Llama-2-7B	Aligned	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	Base	68.6 ± 0.8	16.4 ± 1.4	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

2. Current Safety-aligned Models Exploit This Shortcut:

- Aligned and base models diverge mostly **in the first few tokens** over harmful answers.

Token-wise KL Divergence at each token position k:

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



3. Vulnerabilities!

- **Prefilling Attacks** (Andriushchenko et al., 2024; Haize Labs, 2024; Vega et al., 2023)
  - Prefill a non-refusal prefix
- **Optimization-based Inference-Time Jailbreaks** (Zou et al., 2023b; Andriushchenko et al., 2024)
  - Methods like GCG used shallow surrogate objectives
- **Fine-tuning attacks** (Qi et al., 2023; Zhan et al., 2023)
  - Actually perturb the first few tokens the most

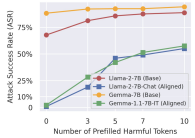
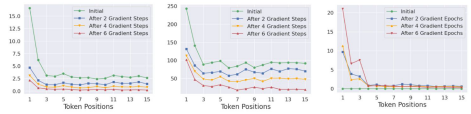


Figure 2: ASR vs. Number of Prefilled Harmful Tokens, with $\hat{\boldsymbol{y}} \sim \pi_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<k})$ on Harmful HEx-PHI.



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

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

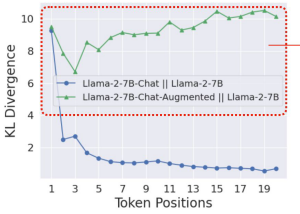
(c) Per-token KL Divergence on HEx-PHI Safety Test Dataset [32]

Figure 3: Then per-token dynamics when fine-tuning Llama-2-7B-Chat on the 100 Harmful Examples from Qi et al. [22]. *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%.*

What If the Safety Alignment Were Deeper?

A Surprisingly Simple Approach: Data Augmentation

<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>



After data augmentation, the token-wise KL divergence is no longer limited to the initial tokens.

The success rates of multiple different types of jailbreak attacks drop on the model with data augmentation!

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

What If the Initial Tokens Were Protected Against Fine-Tuning?

Objective (Inspired by DPO):

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

Larger means stronger protection

Imposing Strong Constraints on the First 5 Tokens Mitigates the Fine-Tuning Attack

Table 4: Fine-tuning with The Constrained Objective in Eqn 3, with larger constraints $\beta_1 = 0.5$, $\beta_t = 2$ for $2 \le t \le 5$ 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