Weekly Meeting

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

Qi et al.
Princeton University / Google DeepMind
ICLR'25 Best Paper Award

Problem: Safety Alignment Can Be Only a Few Tokens Deep

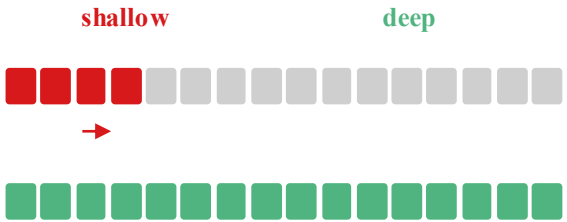
- Central claim: current safety alignment may adapt the model mainly at the first few output tokens.
  - A refusal prefix such as “I cannot” can route the entire continuation toward safety.
  - When the initial tokens are perturbed, the later distribution may still resemble the unaligned model.

- The paper names this failure mode shallow safety alignment.
  - It can make standard evaluations look safe while leaving conditional harmful trajectories under-controlled.

- Counterfactual goal: deep safety alignment.
  - The model should recover to a safe refusal even after non-refusal or harmful starting tokens.

Core intuition

Safety should not be a memorized opening phrase. It should persist after the model has already entered an unsafe-looking trajectory.



Preliminaries

- Safety alignment optimizes an LLM to refuse harmful requests while staying useful on benign instructions
- Scope of this paper: aligned artifacts vs. their unaligned counterparts
  - Llama-2 and Gemma model families (e.g., Llama-2-7B-Chat vs. Llama-2-7B)
- Autoregressive generation: each response token depends on the prompt and all prior tokens
  - So the first output tokens disproportionately steer the rest of the response
- KL Divergence: degree of dissimilarity between a probability distribution and a reference distribution
  - Higher, more dis-similar. $D_{\mathrm{KL}}\left(\pi_{\mathrm{aligned}}(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<k}) \middle\| \pi_{\mathrm{base}}(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<k})\right)$

Shallow vs. Deep Safety Alignment

- **Shallow safety alignment**
  - The model places high probability on a short refusal prefix (e.g., I cannot answer..)
  - Model learned a shortcut for refusal
  - When user manually inserts non-refusal prefix, model cannot recover to a refusal trajectory

first few tokens carry most control

- **Deep safety alignment**
  - The model can recover to a refusal trajectory even after a harmful or non-refusal prefix
  - Harmful continuation stays low even after non-refusal prefixes

control persists across token depth

Evidence 1: LLM learned refusal prefixes as a safety shortcut

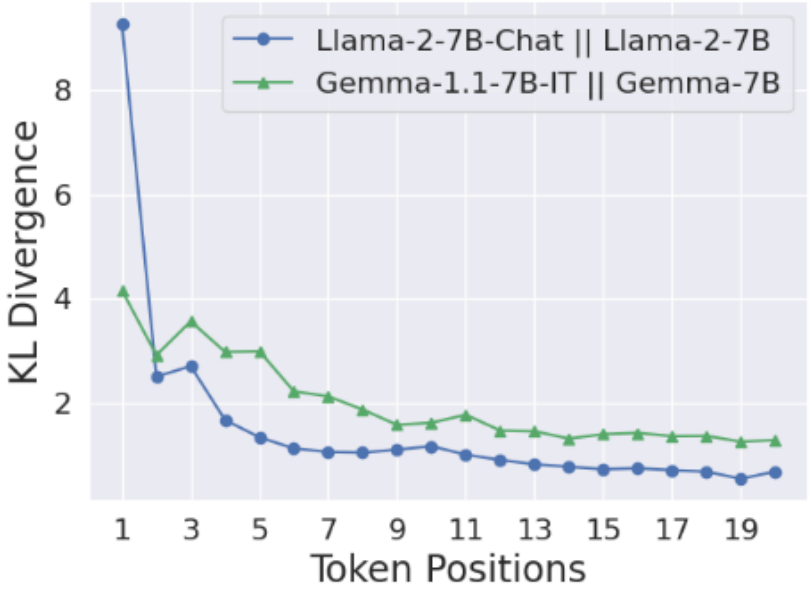
- Even unaligned base models become much safer when decoding is prefilled with refusal text.
  → Cheap local optimum for alignment: increase probability of a refusal prefix.

Refusal Prefixes (r) $\rightarrow$		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
	Base	68.6 ± 0.8	16.4 ± 1.4	5.4 ± 1.3	14.4 ± 0.6	$\mathbf{2.1} \pm \mathbf{0.2}$	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	$\mathbf{1.0} \pm \mathbf{0.8}$	3.9 ± 0.4

Evidence 2: KL Divergence Is Front-Loaded

- Compared aligned models with their base counterparts on harmful answer continuations.
  - The largest aligned-vs-base differences occur at the earliest token positions.

- After the first few tokens, the divergence becomes much smaller
  - This supports the claim that most alignment "budget" is spent near the prefix.



Two possible attack scenarios

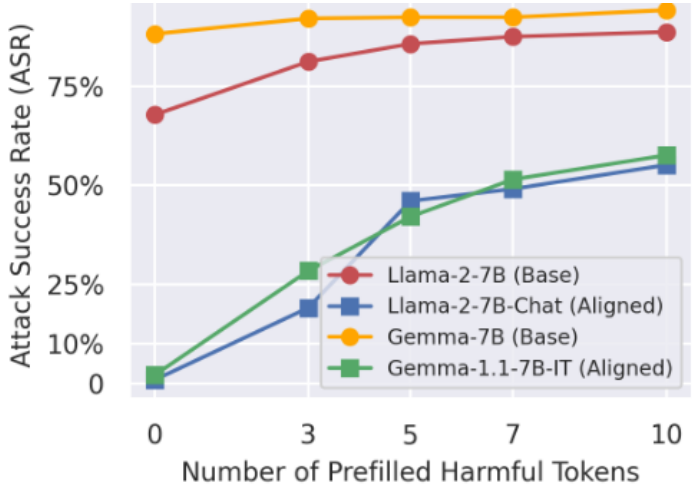
- Inference-Stage Vulnerability: targets inference of the LLM
  - Prefilling attacks: preset LLM's response into specific response
  - Optimization-based exploits: try to incur phrases via appending gibberish characters
  - Decoding parameter exploits: varying top-k, top-p, temperature

- Fine-tuning attacks: fine-tune with small set of adversarial samples

Inference-Stage Vulnerability: Prefilling Attacks

- Prefilling directly supplies the first k output tokens.
  - If those tokens are non-refusal or harmful, the model is conditioned onto an unsafe trajectory.

- Aligned models' ASR rises rapidly as more harmful tokens are prefilled.
  - This is exactly what shallow alignment predicts.



[1] Qi et al., Safety Alignment Should Be Made More Than Just a Few Tokens Deep, arXiv:2406.05946v1

Inference-Stage Vulnerability: Optimization, Decoding-parameter Exploits

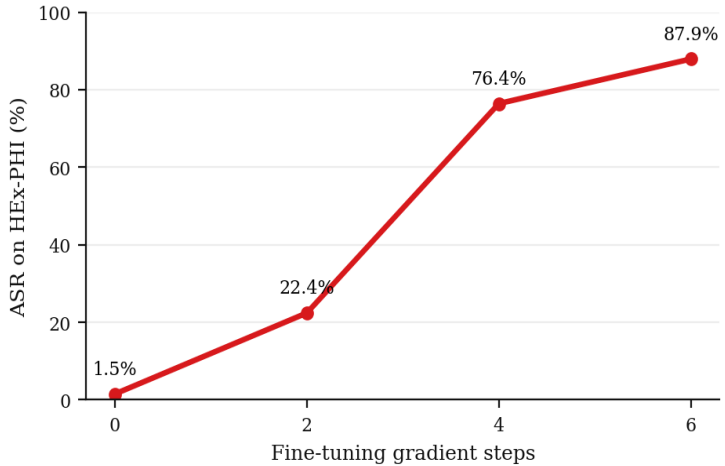
- Optimization-based adversarial suffix attacks
  - Often optimize a suffix so the model starts with an affirmative phrase such as “Sure, here is...”. (e.g., GCG)
  - Objective: targets a prefix that opens a harmful continuation.

- Decoding-parameter exploits
  - Varying temperature, top-k, and top-p can randomly move the initial tokens off the refusal path.
  - Once the start of the response is non-refusal, later tokens may continue harmful content.

Fine-Tuning Attacks: A Few Steps Can Undo Safety

- Downstream fine-tuning can jailbreak aligned LLMs.
  - The paper analyzes fine-tuning on 100 harmful instruction-answer pairs.

- ASR escalates sharply after only a few gradient steps.
  - From 1.5% initially to 87.9% after six gradient steps in the Llama-2-Chat case study.

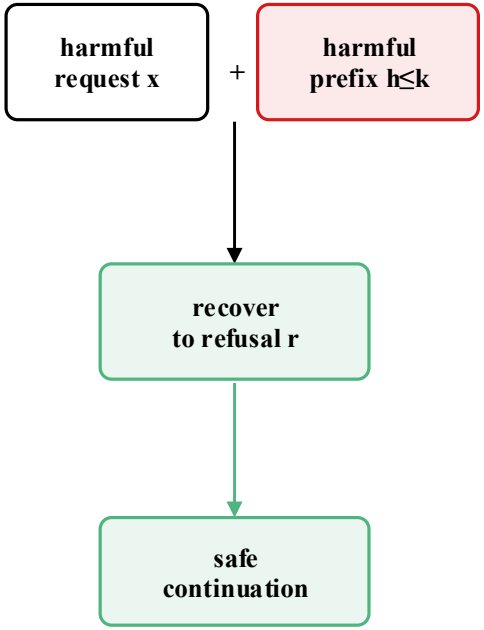


Question: What If Safety Alignment Were Deeper?

- Counterfactual test:
  - If the model is trained to recover after unsafe-looking prefixes, does robustness improve?

- Desired condition:
  - For many k, harmful continuation probability $\pi(h>k \mid x, h\leq k)$ should remain low.

- The authors instantiate this idea through data augmentation.
  - They add safety recovery examples that begin on a harmful trajectory and then transition to refusal.



[1] Qi et al., Safety Alignment Should Be Made More Than Just a Few Tokens Deep, arXiv:2406.05946v1

Data Augmentation: Safety Recovery Examples

- Each safety example is a triplet (x, h, r).
  - x: harmful instruction
  - h: harmful response prefix
  - r: refusal response
- Training promotes refusal after a sampled harmful prefix length k.
  - k = 0 half the time; otherwise k is sampled from 1 to 100.
- A utility anchor prevents broad behavior drift.
  - Benign Alpaca instructions are paired with distilled responses from the original aligned model.

Example pattern

[INST] harmful request [/INST]
Step 1: ... I cannot fulfill your request ...

Objective summary

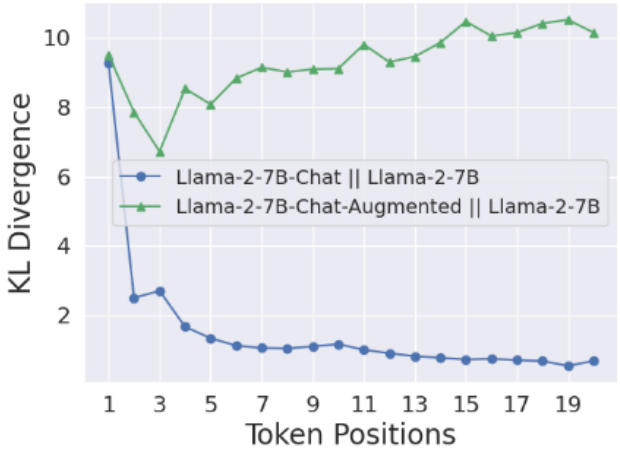
$\alpha \cdot$ safety-recovery loss
$+ (1 - \alpha) \cdot$ utility-anchor loss

Reported $\alpha = 0.2$

Effect of Augmentation: Deeper Divergence, Preserved Utility

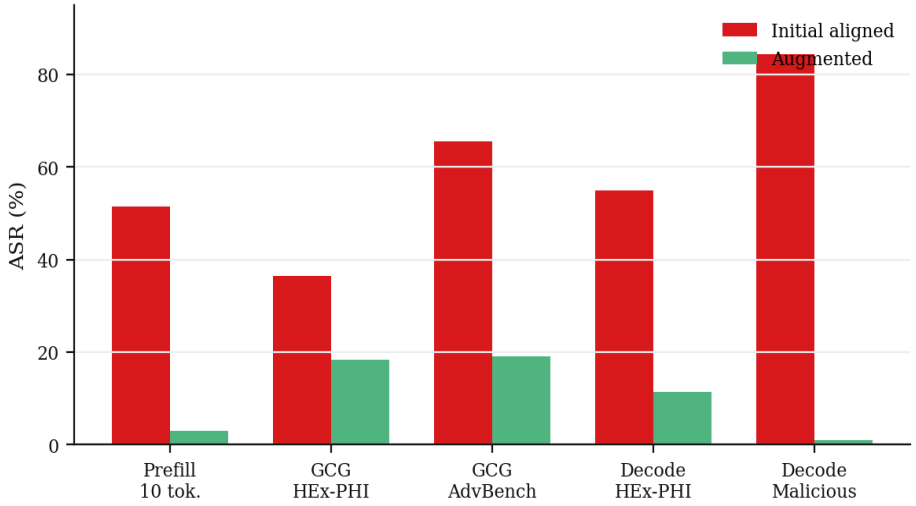
- The augmented model exhibits larger KL divergence at later harmful-token positions.
  - This indicates that safety control is pushed deeper into the sequence.

- Utility is largely preserved.
  - AlpacaEval winrate (vs. text-davinci) decreases from 51.8% to 49.5%



Deepened Alignment Improves Robustness

- The augmented model reduces ASR across several inference-time attacks.
  - Prefilling attacks, GCG, and decoding-parameter exploits all show lower ASR.



Intervention 2: Protect Initial Tokens During Fine-Tuning

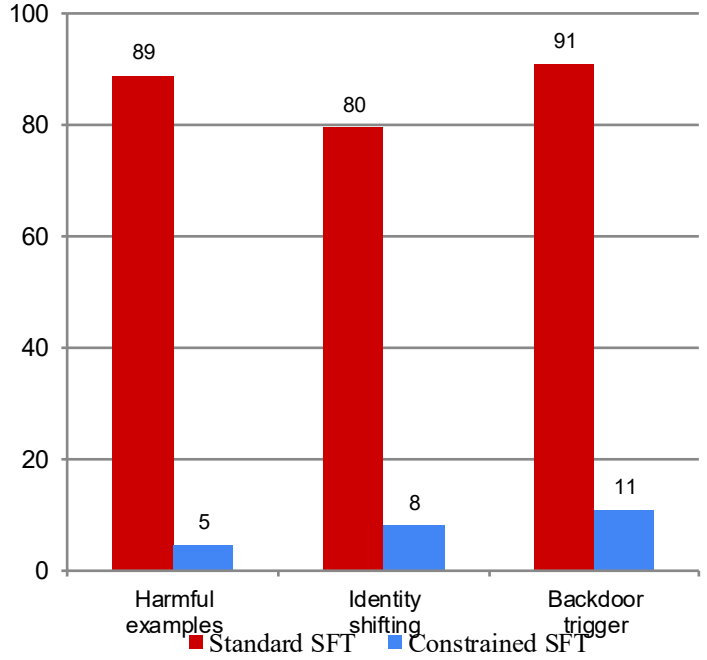
- Because fine-tuning attacks shift the first few tokens most, adding a **token-wise constraint** that discourages early-token deviation from the aligned model.
- **Constrained SFT:** down-weight token updates that move too far from π_{aligned}.
- **β controls the exponential deviation of gradient (when given harmful responses)**

$$w_t = 2 \cdot \sigma(\beta_t \cdot \Delta_t) \qquad \Delta_t = \log \pi_{aligned}(y_t) - \log \pi_\theta(y_t)$$

- First 5 tokens: β=2.0 ; otherwise: β=0.1

Intervention 2: Protect Initial Tokens During Fine-Tuning

- On Llama-2-Chat, the constrained objective **sharply reduces ASR** under harmful-example, identity-shifting, and backdoor-trigger attacks.
- The protection comes from limiting how far the earliest tokens may move during fine-tuning.
- **Utility stays competitive** on benign tasks (e.g. ROUGE-1, accuracy) — safety is preserved without a large quality cost.

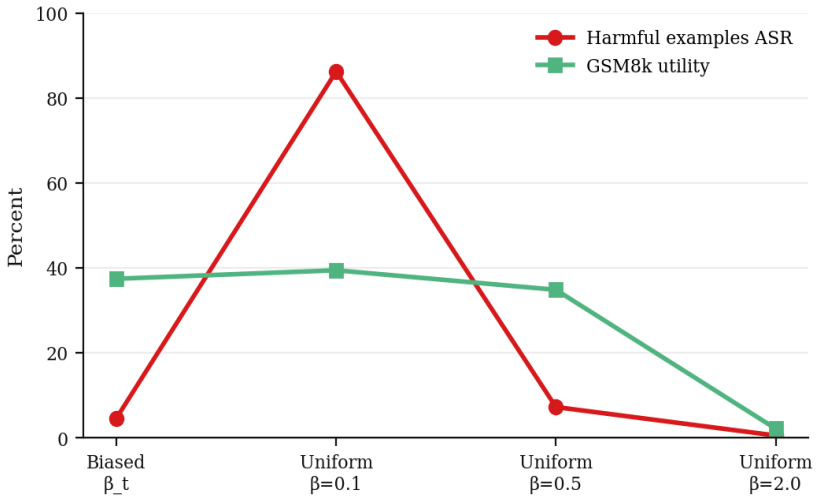


Ablations: Why Early-Token Constraints Matter

- Uniformly weak constraints do not stop harmful fine-tuning.
  - Uniform $\beta = 0.1$ leaves harmful-example ASR at 86.2%.

- Uniformly strong constraints preserve safety but can damage utility.
  - Uniform $\beta = 2.0$ gives low ASR but collapses GSM8k utility to 2.1%.

- Biased βt is the preferred compromise.
  - Strong for initial tokens; weak for later task-learning tokens.



Selected values from Table 4: safety/utility trade-offs across β choices.

Limitations, Future Directions

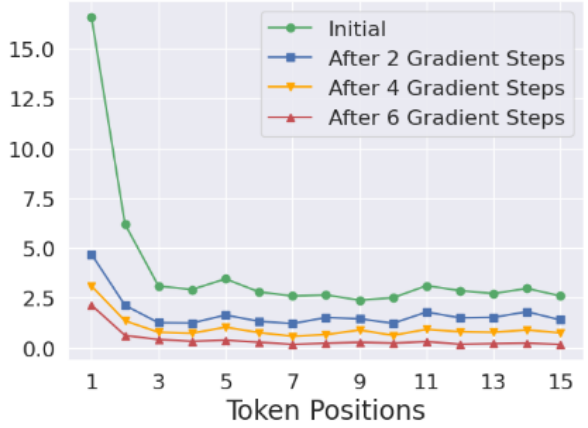
- Pros
  - Well-written, good story, every claims are proven
  - Tackles the in-depth analysis why LLMs still misbehave even we spend millions of budgets
  - Easily adaptable to the AI systems

- Cons
  - Adaptive attacks may target recovery behavior directly.
  - Alignment-from-scratch remains unexplored.
  - Safety depth is multi-dimensional: pruning, and distribution shift need explored

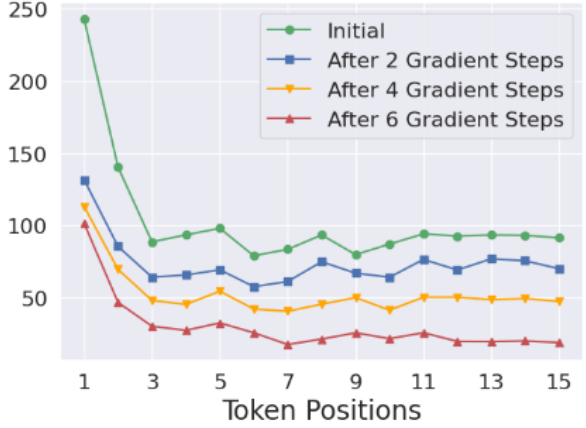
- Future directions:
  - Integrate recovery examples into SFT/RLHF/DPO pipelines
  - Neuronwise forensics would explain more in depth

Thank you

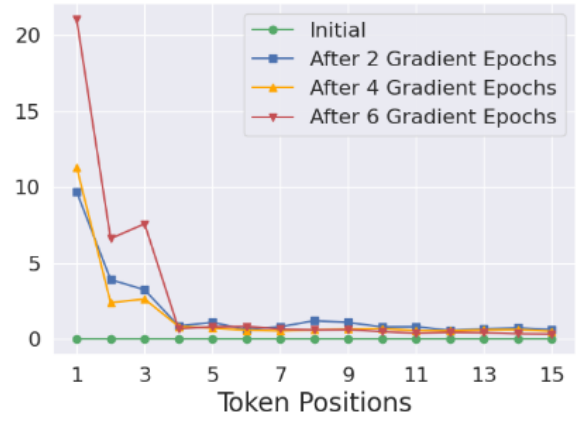
Fine-Tuning Attacks: A Few Steps Can Undo Safety (contd.)



(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]

Per-Token Dynamics Explain Fine-Tuning Fragility

- The fine-tuning loss decomposes across output positions:
  - Standard SFT minimizes the sum of per-token negative log-likelihoods.

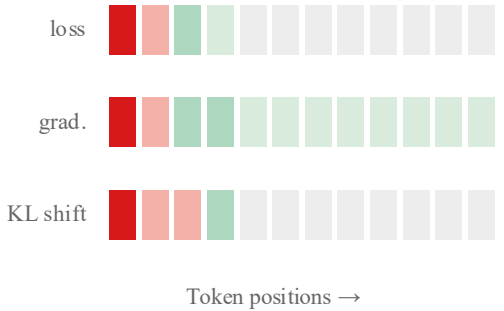
- Early tokens dominate the update.
  - The paper reports larger early-token cross-entropy loss and gradient norms.
  - The post-fine-tuning KL shift is also largest at the beginning of harmful responses.

- This matches the shallow-alignment hypothesis.
  - Safety is fragile because it is concentrated in the same region that fine-tuning perturbs most.

Standard SFT objective

$$\min_\theta \; \mathbb{E}_{(x,y)\sim D}\left[-\sum_t \log \pi_\theta(y_t \mid x, y_{<t}) \right]$$

Conceptual token-wise update intensity



[1] Qi et al., Safety Alignment Should Be Made More Than Just a Few Tokens Deep, arXiv:2406.05946v1