Safety Alignment Can Be Not Superficial With Explicit Safety Signals

Jianwei Li, Jung-Eun Kim

Department of Computer Science,

North Carolina State University, Raleigh, NC

Task of Interest

- Improve Safety Alignment of LLMs to enable robust refusals:
  - Direct Attacks [1 & 2]
  - Jailbreak Attacks [3,4,5,6,8]
    - Suffix
    - Prefix
    - Prefill
    - Role play
    - Netsed Scene
    - Token manipulation
    - Persuation
    - Optimized (Gradient or Genetic)
    - ...
  - Decoding Exploitation Attacks [7]

Background

1. Superficial Safety Alignment Hypothesis [10]
    1. Current generative LLMs implicitly perform a safety-related binary classification task.
    2. Current aligned model can't hold safety at each generation step
2. Data Augmentation Based Methods [5, 9]
    1. Construct more complex adversarial samples that are initialy fullfilled but later refused.
    2. Do not fundamentally address the root problem
    3. Struggle to handle harmful content that appears mid- or end-generation.

- **Challenge**: Existing alignment techniques lack the mechanisms to handle nested harmful reasoning patterns or those that emerge near the end of a response, pressing the need for more robust methods that address safety at a deeper level.

Observation & Motivation

- **Implicit** safety signal is often diluted or overridden by competing objectives, such as learning complex human preferences related to tone, style, or phrasing of responses.
- Can we extract and take use of some **Explicit** safety-related signals to prevent or alleviate the above unexpected situation?

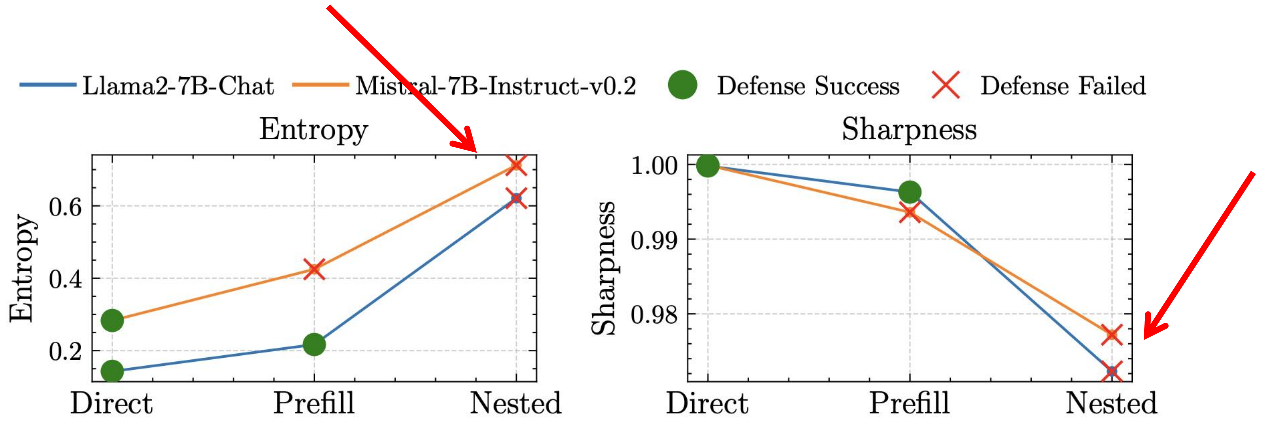

Figure 4. Entropy (left) and sharpness (right) of **Llama2-7B-Chat** and **Mistral-7B-Instruct-v0.2** under increasing adversarial complexity. As adversarial complexity increases (**Direct** → **Prefill** → **Nested**), both models show higher entropy and lower sharpness, reflecting reduced confidence and alignment robustness. Notably, in the nested scenario, both models fail to maintain safety as highlighted by the success of the attack (in red X).

Methodology - Explicit Binary Classification Task

- Incorporating a safety-related **Binary Classification** task into the training process to **explicitly** extract safety-related signals

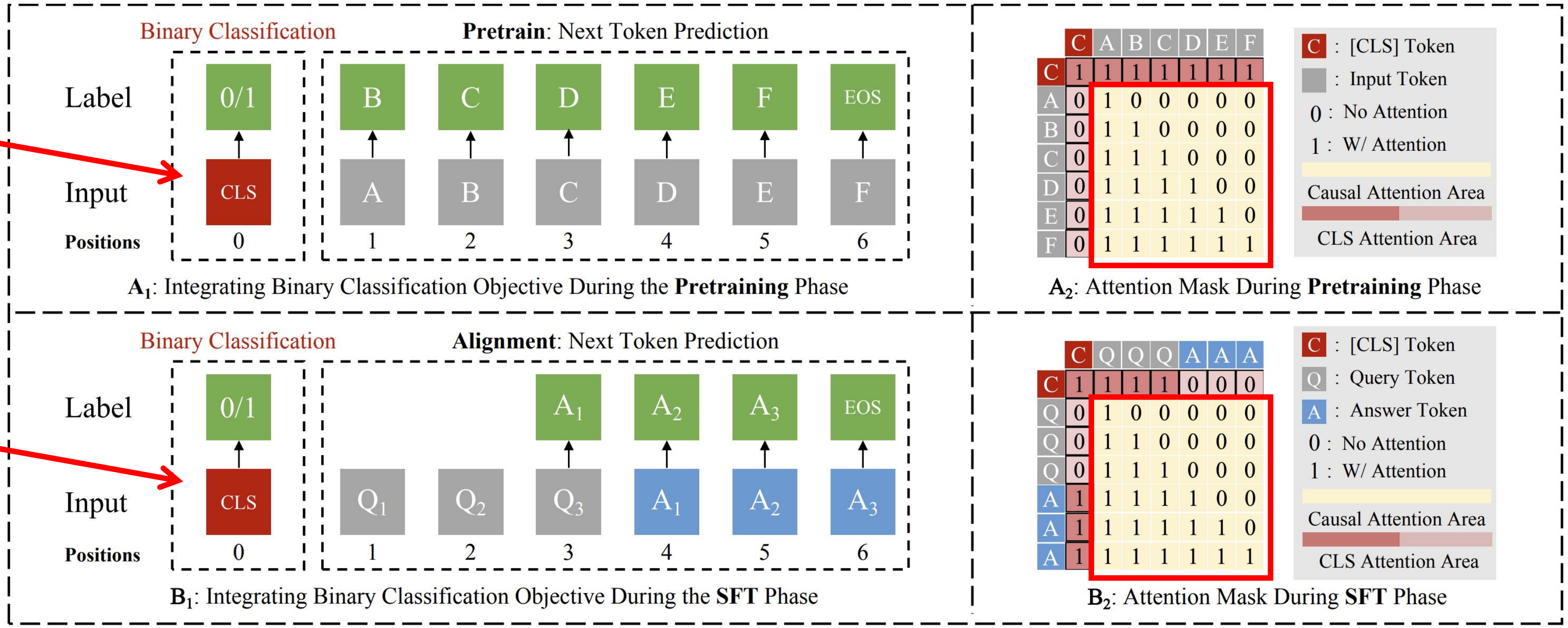

Figure 1. Integration of a safety-related binary classification task into the pre-training and supervised fine-tuning phases of LLMs.

Strategic Attention - Implicitly

ICML International Conference On Machine Learning

NC STATE UNIVERSITY

- A mechanism integrates the hidden state of the [CLS] token into the model's generative process, allowing it to **implicitly** incorporate safety signals during entire text generation process.

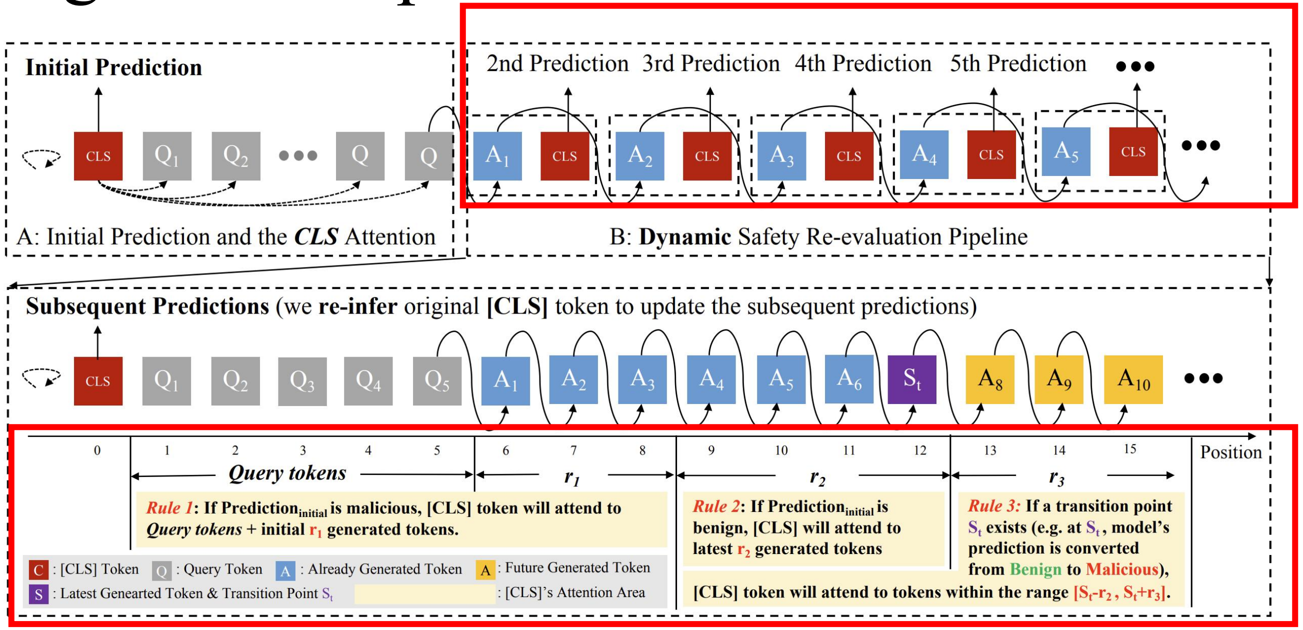

Figure 2. Strategic Attention Mechanism. (A) Initial predictions leverage the `[CLS]` token's attention to evaluate safety. (B) The dynamic safety re-evaluation pipeline updates predictions as new tokens are generated. Subsequent `[CLS]` token's attention follows defined rules: **1)** focusing on query tokens and initial r_1 tokens, **2)** the latest r_2 tokens, **3)** or a specific range around a transition point (S_t), ensuring adaptive and context-sensitive safety assessments throughout the generation process.

Strategic Decoding - Explicitly

- A strategy **explicitly** leverages the prediction of the **binary classification** task to guide the model's decision-making process during text generation, enabling it to respond to complex adversarial scenarios more timely and confidently.

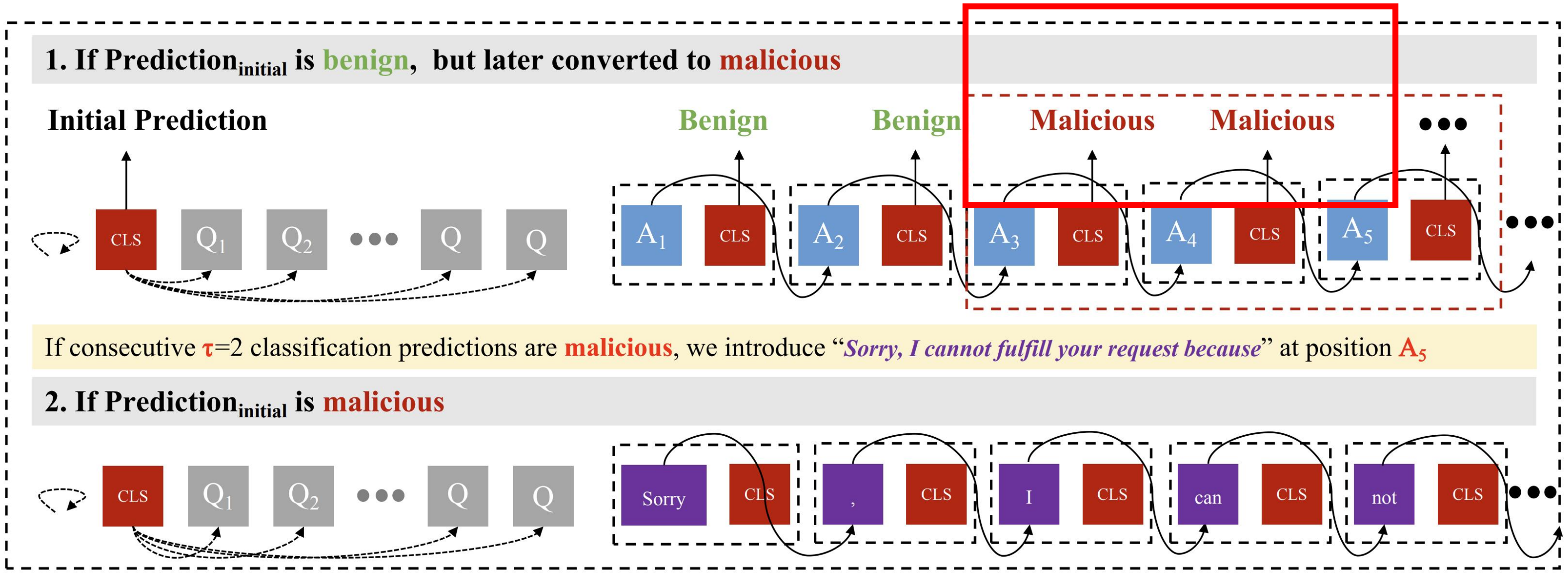

Figure 3. Strategic Decoding Mechanism. We use the `[CLS]` token's dynamic predictions to adaptively refuse malicious inputs, either by inserting refusal phrases after consecutive malicious classifications or responding immediately to initially malicious queries.

Experiment Results

- Primary Baseline (SFT, SFT + DPO)
- Official Release Baseline
- State-of-the-Art Baseline
- Cross-family Baseline

Table 1. Comparison with primary and official released baselines. This table compares the Attack Succesful Rate (**ASR** (%)) of **Llama2-7B-SFT**, **Llama2-7B-SFT-DPO**, **Llama2-7B-Chat (RLHF)**, and **Llama2-7B-CLS (Ours)** across various benchmarks and jailbreak attacks. **Llama2-7B-CLS** achieves significantly lower ASR, demonstrating superior safety and robustness over other alignment methods. The only exception is the **DeepInception** jailbreak attack, where our method has a **single** failure case, resulting in a slightly higher ASR.

ASR (%) ↓	Attack Method	Llama2-7B–SFT	Llama2–7B–SFT–DPO	Llama2–7B–CHAT	Llama2–7B–CLS
AdvBench	Direct	$1.15\% \pm 0.19\%$	$1.5\% \pm 0.19\%$	$0.19\% \pm 0.19\%$	$\mathbf{0.19\% \pm 0\%}$
HEx–PHI	Direct	$3.33\% \pm 0.3\%$	$4.24\% \pm 0.61\%$	$2.73\% \pm 0.3\%$	$\mathbf{0.3\% \pm 0\%}$
			Jailbreak Attack		
AdvBench	Prefill	$92.7\% \pm 2.69\%$	$12.12\% \pm 1.35\%$	$39.62\% \pm 2.5\%$	$\mathbf{0.4\% \pm 0\%}$
HEx–PHI	Prefill	$92.73\% \pm 2.42\%$	$21.52\% \pm 2.12\%$	$60.91\% \pm 2.12\%$	$\mathbf{1.2\% \pm 0.3\%}$
Harmbench	GCG	$41.0\% \pm 2.0\%$	$14.0\% \pm 1.0\%$	$28.0\% \pm 3.0\%$	$\mathbf{0.0\% \pm 0\%}$
AdvBench	AutoDAN-T	$13.08\% \pm 2.31\%$	$\mathbf{0.77\% \pm 0.19\%}$	$61.3\% \pm 2.31\%$	$\mathbf{0.77\% \pm 0.19\%}$
AdvBench	DeepInception	$38.0\% \pm 2.0\%$	$\mathbf{0\% \pm 0\%}$	$36.0\% \pm 2.0\%$	$2.0\% \pm 0\%$
AdvBench	PAP	$17.39\% \pm 2.17\%$	$\mathbf{0\% \pm 0\%}$	$\mathbf{28.26\% \pm 2.17\%}$	$\mathbf{0.0\% \pm 0\%}$
Alert Adversarial	Suffix	$0.14\% \pm 0.01\%$	$0.13\% \pm 0.01\%$	$0.01\% \pm 0.01\%$	$\mathbf{0\% \pm 0\%}$
Alert Adversarial	Prefix	$0.11\% \pm 0.01\%$	$0.07\% \pm 0.01\%$	$0.28\% \pm 0.01\%$	$\mathbf{0.03\% \pm 0.01\%}$
Alert Adversarial	TokenSwap	$0.27\% \pm 0.04\%$	$0.2\% \pm 0.03\%$	$0.24\% \pm 0.03\%$	$\mathbf{0.01\% \pm 0.01\%}$
Alert Adversarial	Role Play	$0.4\% \pm 0.06\%$	$0.31\% \pm 0.03\%$	$\mathbf{0.02\% \pm 0.01\%}$	$\mathbf{0.02\% \pm 0.01\%}$
			Decoding Attack		
MaliciousInstruction	Decoding	$98\% \pm 2.0\%$	$\mathbf{0\% \pm 0\%}$	$83\% \pm 2.0\%$	$\mathbf{0\% \pm 0\%}$
AdvBench	Decoding	$89\% \pm 2.69\%$	$\mathbf{0\% \pm 0\%}$	$87\% \pm 1.92\%$	$\mathbf{0\% \pm 0\%}$

More Experiment Results

ICML International Conference On Machine Learning

NC STATE UNIVERSITY

Table 2. Comparison with state-of-the-art baselines.This table compares the **ASR (%)** of **Llama2-7B-Chat**, **Llama2-7B-Chat-Aug**, and **Llama2-7B-CLS** across benchmarks from Qi et al. (2024) (* indicates results excerpted from the original paper). **Llama2-7B-CLS** achieves the best performance, demonstrating superior robustness through explicit safety signals and dynamic reclassification. Performance under **GCG** attacks is discussed further in Section 5 due to computational constraints.

ASR (%) ↓	Prefilling Attacks				GCG Attack		Decoding Parameters Exploit	
	5 tokens	10 tokens	20 tokens	40 tokens	HEx-PHI	AdvBench	HEx-PHI	MaliciousInstruct
Llama2-7B-Chat *	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
Llama2-7B-Chat-Aug *	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
Llama2–7B–CLS	$\mathbf{0.9 \pm 0}$	$\mathbf{2.1 \pm 0}$	$\mathbf{2.7 \pm 0}$	$\mathbf{2.1 \pm 0}$	–	–	$\mathbf{0.0 \pm 0}$	$\mathbf{0.0 \pm 0}$

Table 3. Comparision with cross-family baselines. This table compares the **ASR (%)** and **Utility** score of **Mistral-7B-Instruct-v0.2**, **Llama2-7B-Chat**, and **Mistral-7B-Instruct-v0.2-CLS**. The results shows that our method can also improve the safety of already aligned models. Specially, the enhanced **Mistral** family model demonstrates superior **helpfulness**, and comparative **safety** collectively, outperforming the **Llama2** family model (**Llama2** family is recognized for its strong safety but less helpfulness compared to **Mistral**).

Benchmark	MT-Bench ↑	GSM8K ↑	mmlu ↑	AdvBench ↓				HarmBench ↓	HEx-PHI ↓		Alert-Adversarial ↓			
				Direct	Prefill	AutoDAN-T	DeepInception	GCG	Direct	Prefill	Prefix	Suffix	TokenSwap	RolePlay
Mistral-7B-Instruct-0.2	**7.56**	41.09	**59.1**	42.31%	92.12%	76.54%	82.0%	66.0%	49.7%	90.91%	49.29%	15.25%	8.65%	6.01%
Llama2-7B-Chat	6.32	22.97	46.36	**0.19%**	39.62%	61.3%	36.0%	26.8%	2.73%	60.91%	0.28%	**0.01%**	**0.24%**	**0.02%**
Mistral-7B-Instruct2-CLS	7.38	**41.77**	58.20	**0.19%**	**0.4%**	**2.89%**	**10.0%**	**0.0%**	**1.21%**	**2.12%**	**0.01%**	0.4%	0.4%	0.3%

References

[1] Zou et al, Universal and transferable adversarial attacks on aligned language models

[2] Qi et al. Fine-tuning aligned language models compromises safety, even when users do not intend to!

[3] Tedeschi et al. ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming

[4] Li. et al. Deepinception: Hypnotize large language model to be jailbreaker

[5] Qi et al. Safety Alignment Should Be Made More Than Just a Few Tokens Deep

[6] Zeng et al. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms

[7] Huang et al. Catastrophic jailbreak of open-source llms via exploiting generation

[8] Liu et al. Autodan: Generating stealthy jailbreak prompts on aligned large language models

[9] Yuan et al. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training

[10] Li et al. Superficial Safety Alignment Hypothesis

Thank You