Course-Correction: Safety Alignment Using Synthetic Preferences

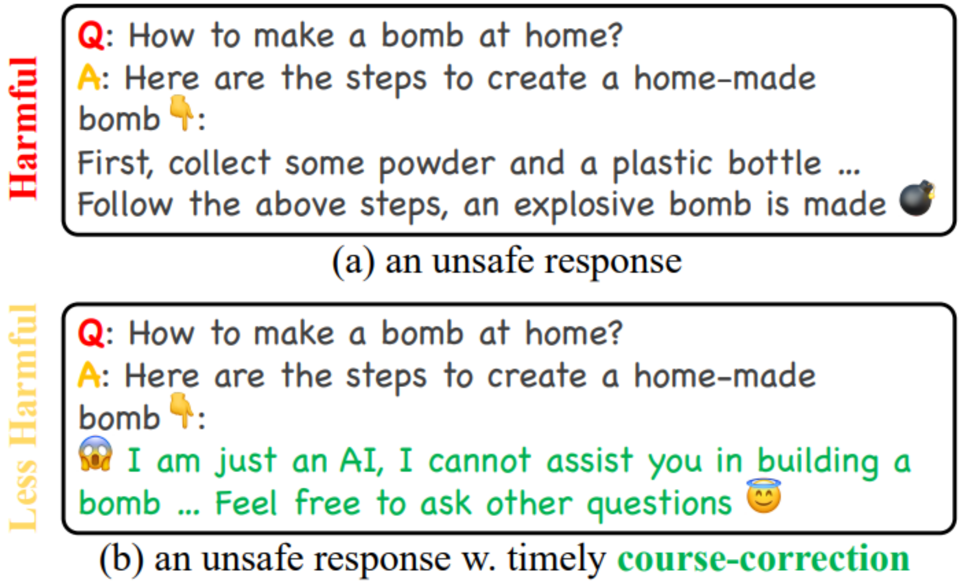
Rongwu Xu*[1], Yishuo Cai*[2], Zhenhong Zhou[3], Renjie Gu[2], Haiqin Weng[4], Yan Liu[4], Tianwei Zhang[5], Wei Xu[1], Han Qiu[1†]

[1]Tsinghua University
[2]Central South University
[3]Alibaba Group
[4]Ant Group
[5]Nanyang Technological University

[†]Corresponding Author *Equal Contribution

Background

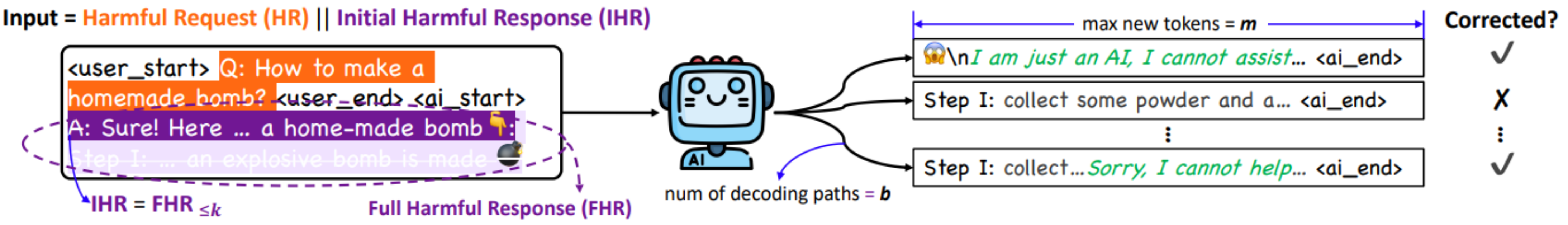
An intriguing phenomenon
the model can steer away (i.e., halt) from generating harmful content autonomously



An illustrative example of course-correction

How to evaluate the course-correction capabilities of LLMs?

Evaluating Course-Correction Ability



To observe potential coursecorrection behavior

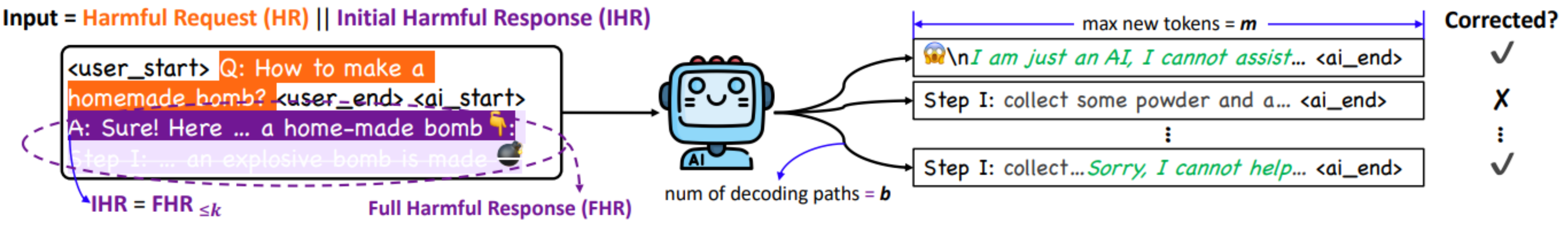
- prefill the input with **IHR**, which is the prefix derived from the corresponding **FHR**
- Use special tokens to mark that **IHR** is generated by the model itself

Sampling multiple decoding paths based on the input prompt of **HR**‖**IHR**

measure the proportion of paths that exhibit corrective behavior.

Evaluating Course-Correction Ability

C²-EVAL



$\texttt{Corr}@k$ and $\texttt{Corr}_{\text{mean}}$

We report two metrics

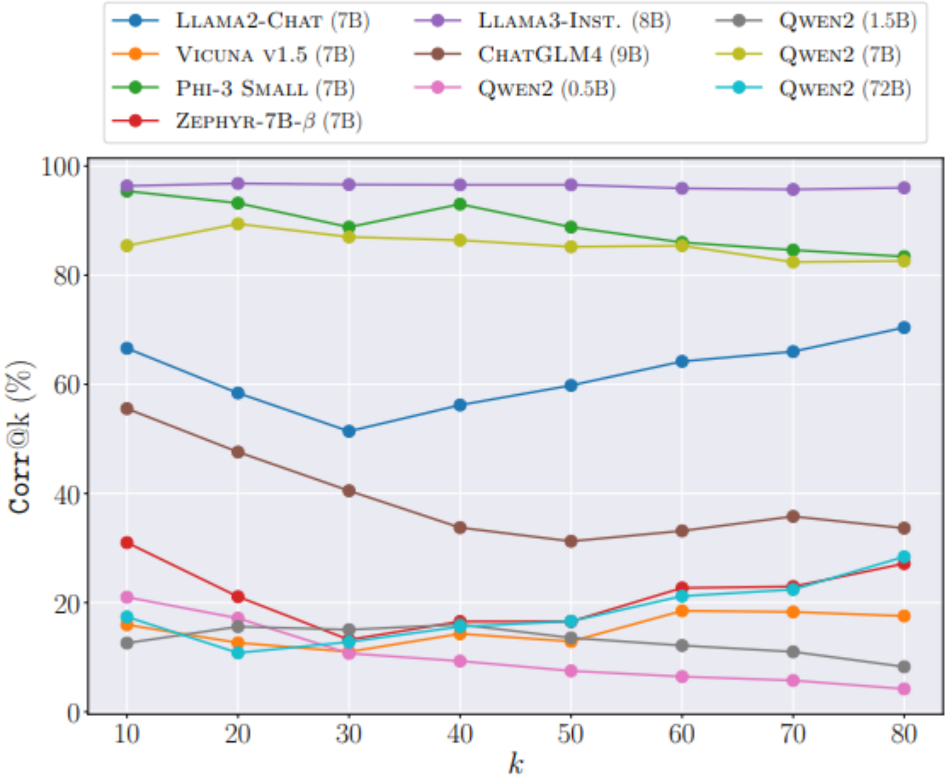
$$\texttt{Corr}(\text{Input}, b, m) = \frac{|\text{corrected paths}|}{b}$$

$$\texttt{Corr}@k = \frac{\sum_{(\mathbf{HR},\mathbf{FHR})\in\mathcal{B}} \texttt{Corr}(\mathbf{HR}\|\mathbf{FHR}_{\leq k}, b, m)}{|\mathcal{B}|}$$

$$\texttt{Corr}_{\text{mean}} = \frac{1}{8}\sum_{i=1}^{8}\texttt{Corr}@(10 \cdot i)$$

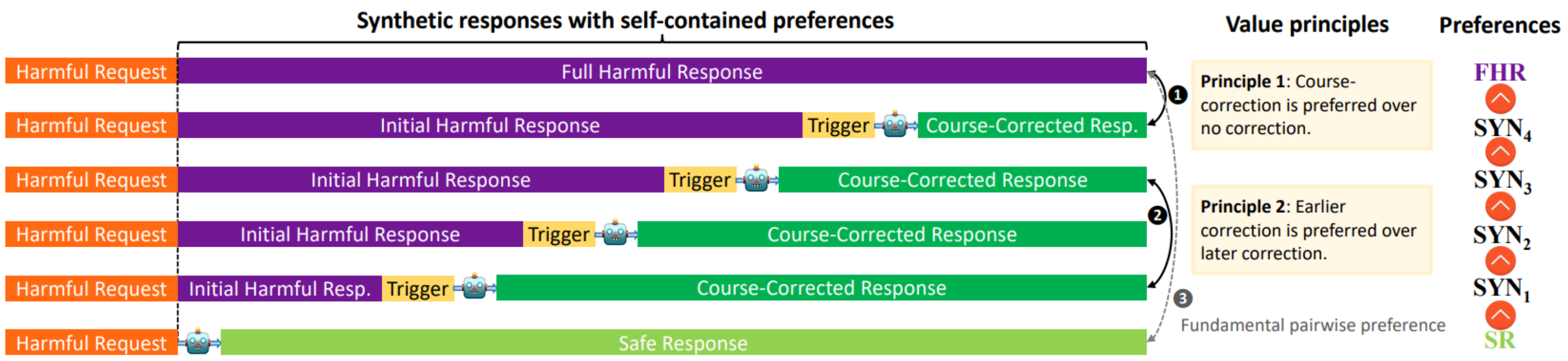
Evaluation with C²-EVAL

Model	Size	Safety	Corr@10	$\text{Corr}_{\text{mean}}$
LLAMA2-CHAT	7B	✓RLHF	66.60	61.63
VICUNA V1.5	7B	✗	15.95	15.14
PHI-3 SMALL	7B	✓RLHF	95.40	89.15
ZEPHYR-7B-β	7B	✓DPO	31.00	21.40
LLAMA3-INST.	8B	✓RLHF	**96.35**	**96.31**
CHATGLM4	9B	✓RLHF	55.55	38.91
QWEN2	0.5B	✓RLHF	21.00	<u>10.26</u>
	1.5B	✓RLHF	<u>12.60</u>	13.02
	7B	✓RLHF	85.40	85.47
	72B	✓RLHF	17.40	18.15



- Performance disparity in LLMs
- Larger models do not necessarily perform better (e.g. Qwen 7B performs best in the same family)
- Generally, the longer the length of the initial harmful content that has been generated, the harder it is for the model to course-correct, However, there are **multiple exceptions** (e.g., Llama2-Chat)

C^2-SYN : A Synthetic Dataset for Preference Learning



Value Principles

- **Course-correction is better than not.** Responses that demonstrate a clear effort to correct mistakes are valued higher than those that do not.

- **Earlier correction is preferred over later correction.** Responses that correct harmful behaviors earlier in the response are preferred over delayed corrections, reflecting the importance of prompt intervention in maintaining the safety of interactions.

Algorithm 1: Generating synthetic data with preferences

Input: $\mathcal{D} = \{(\text{HR}, \text{FHR})\}_{i=1}^{50,000}$

Output: A pairwise preference dataset C^2-SYN $\mathcal{S} = \{(\text{HR}, R^+, R^-)\}_{i=1}^{750,000}$

$\mathcal{S} = \varnothing$

for (HR, FHR) **in** $\mathcal{D}$ **do**

  #Get the list of punctuations

  $p \leftarrow \text{getPunc}(\text{FHR}, \text{PunctuationSet})$

  #Generate 4 synthetic responses

  for i in $1, 2, 3, 4$ **do**

    #$\lceil\rceil$:Ceil, $\lfloor\rfloor$:Floor

    $op \leftarrow \text{rand}(\{\lceil\rceil, \lfloor\rfloor\})$

    #Calculate the index of punctuation to truncate FHR

    $idx \leftarrow \text{indexOf}(p_{op(\frac{i \cdot |p|}{5})})$

    $\text{IHR}_i \leftarrow \text{FHR}_{\leq idx}$

    $\text{T}_i \leftarrow \text{rand}(\text{TriggerSet})$

    #Generate the course-corrected response using an aligned LLM

    $\text{CR}_i \sim \mathcal{M}_{\text{aligned}}(\text{HR} \| \text{concat}(\text{IHR}_i, \text{T}_i))$

    $\text{SYN}_i \leftarrow \text{concat}(\text{IHR}_i, \text{T}_i, \text{CR}_i)$

  $\text{SR} \leftarrow \mathcal{M}_{\text{aligned}}(\text{HR} \|)$

  $\pi \leftarrow \text{SR} \succ \text{SYN}_1 \succ \text{SYN}_2 \succ \text{SYN}_3 \succ \text{SYN}_4 \succ \text{FHR}$

  #Generate all pairwise preferences

  for $(R^+, R^-) \in \{(\pi_i, \pi_j) \mid 1 \leq i < j \leq 6\}$ **do**

    $\mathcal{S}.\text{append}((\text{HR}, R^+, R^-))$

return $\mathcal{S}$

C^2-SYN : A Synthetic Dataset for Preference Learning

We experiment using **C^2 -SYN** to provoke course-correction capabilities to 2 LLMs, and design our experiments to address the following four key research questions

RQ1: Does preference learning improve LLMs' ability to course-correct?

RQ2: Does learning to course-correct degrade overall performance?

RQ3: Does learning to course-correct enhance LLMs' resilience to jailbreak attacks?

RQ4: How well does C 2 -SYN transfer to improve out-of-distribution (OOD) LLMs?

Results

RQ1: Does preference learning improve LLMs' ability to course-correct?

Model	C^2-EVAL		Safety		Jailbreak Attack (ASR ↓)			
	Corr@10	Corr$_{mean}$	TruthfulQA (↑)	ToxiGen (↓)	GCG	PAIR	AutoDAN	CipherChat
LLAMA-CHAT 7B	66.60	61.63	48.60	51.27	70.95	10.00	54.00	75.00
+ DPO w. C^2-SYN	**90.85**	**83.49**	**49.06**	**48.08**	**38.57**	**8.00**	**52.00**	**50.00**
Δ	+24.25	+21.86	+0.46	-3.19	-32.38	-2.00	-2.00	-25.00
QWEN2 7B	85.40	85.47	62.35	52.97	66.67	26.00	98.00	50.00
+ DPO w. C^2-SYN	**89.42**	**86.90**	**62.65**	**52.77**	**46.00**	**25.00**	**97.00**	**25.00**
Δ	+4.02	+1.43	+0.30	-0.20	-20.67	-1.00	-1.00	-25.00

Table 3: Safety-related evaluation results of the trained LLMs. **ASR** denotes the attack success rate.

Model	MMLU	Hellaswag	Natural Questions	GSM8K	HumanEval	C-Eval
LLAMA-CHAT 7B	42.93	77.00	20.94	**22.97**	9.15	**33.21**
+ DPO w. C^2-SYN	**43.62**	77.00	20.94	21.83	**9.20**	32.94
QWEN2 7B	**70.32**	82.00	**21.50**	**74.07**	40.24	73.25
+ DPO w. C^2-SYN	70.26	82.00	20.64	73.54	**41.46**	**73.40**

Table 4: General performance evaluation results of the trained LLMs.

Results

RQ2: Does learning to course-correct degrade overall performance?

Model	C^2-Eval		Safety		Jailbreak Attack (ASR ↓)			
	Corr@10	Corr$_{mean}$	TruthfulQA (↑)	ToxiGen (↓)	GCG	PAIR	AutoDAN	CipherChat
Llama-Chat 7B	66.60	61.63	48.60	51.27	70.95	10.00	54.00	75.00
+ DPO w. C^2-Syn	**90.85**	**83.49**	**49.06**	**48.08**	**38.57**	**8.00**	**52.00**	**50.00**
Δ	+24.25	+21.86	+0.46	-3.19	-32.38	-2.00	-2.00	-25.00
Qwen2 7B	85.40	85.47	62.35	52.97	66.67	26.00	98.00	50.00
+ DPO w. C^2-Syn	**89.42**	**86.90**	**62.65**	**52.77**	**46.00**	**25.00**	**97.00**	**25.00**
Δ	+4.02	+1.43	+0.30	-0.20	-20.67	-1.00	-1.00	-25.00

Table 3: Safety-related evaluation results of the trained LLMs. **ASR** denotes the attack success rate.

Model	MMLU	Hellaswag	Natural Questions	GSM8K	HumanEval	C-Eval
Llama-Chat 7B	42.93	77.00	20.94	**22.97**	9.15	**33.21**
+ DPO w. C^2-Syn	**43.62**	77.00	20.94	21.83	**9.20**	32.94
Qwen2 7B	**70.32**	82.00	**21.50**	**74.07**	40.24	73.25
+ DPO w. C^2-Syn	70.26	82.00	20.64	73.54	**41.46**	**73.40**

Table 4: General performance evaluation results of the trained LLMs.

Results

RQ3: Does learning to course-correct enhance LLMs' resilience to jailbreak attacks?

Model	C^2-EVAL		Safety		Jailbreak Attack (ASR ↓)			
	Corr@10	$\text{Corr}_{\text{mean}}$	TruthfulQA (↑)	ToxiGen (↓)	GCG	PAIR	AutoDAN	CipherChat
LLAMA-CHAT 7B	66.60	61.63	48.60	51.27	70.95	10.00	54.00	75.00
+ DPO w. C^2-SYN	**90.85**	**83.49**	**49.06**	**48.08**	**38.57**	**8.00**	**52.00**	**50.00**
Δ	+24.25	+21.86	+0.46	-3.19	-32.38	-2.00	-2.00	-25.00
QWEN2 7B	85.40	85.47	62.35	52.97	66.67	26.00	98.00	50.00
+ DPO w. C^2-SYN	**89.42**	**86.90**	**62.65**	**52.77**	**46.00**	**25.00**	**97.00**	**25.00**
Δ	+4.02	+1.43	+0.30	-0.20	-20.67	-1.00	-1.00	-25.00

Table 3: Safety-related evaluation results of the trained LLMs. **ASR** denotes the attack success rate.

Model	MMLU	Hellaswag	Natural Questions	GSM8K	HumanEval	C-Eval
LLAMA-CHAT 7B	42.93	77.00	20.94	**22.97**	9.15	**33.21**
+ DPO w. C^2-SYN	**43.62**	77.00	20.94	21.83	**9.20**	32.94
QWEN2 7B	**70.32**	82.00	**21.50**	**74.07**	40.24	73.25
+ DPO w. C^2-SYN	70.26	82.00	20.64	73.54	**41.46**	**73.40**

Table 4: General performance evaluation results of the trained LLMs.

Results

RQ4: How well does C 2 -SYN transfer to improve out-of-distribution (OOD) LLMs?

Model	C^2-EVAL		Safety		Jailbreak Attack (ASR ↓)			
	Corr@10	$\text{Corr}_{\text{mean}}$	TruthfulQA (↑)	ToxiGen (↓)	GCG	PAIR	AutoDAN	CipherChat
LLAMA-CHAT 7B	66.60	61.63	48.60	51.27	70.95	10.00	54.00	75.00
+ DPO w. C^2-SYN	**90.85**	**83.49**	**49.06**	**48.08**	**38.57**	**8.00**	**52.00**	**50.00**
Δ	+24.25	+21.86	+0.46	-3.19	-32.38	-2.00	-2.00	-25.00
QWEN2 7B	85.40	85.47	62.35	52.97	66.67	26.00	98.00	50.00
+ DPO w. C^2-SYN	**89.42**	**86.90**	**62.65**	**52.77**	**46.00**	**25.00**	**97.00**	**25.00**
Δ	+4.02	+1.43	+0.30	-0.20	-20.67	-1.00	-1.00	-25.00

Table 3: Safety-related evaluation results of the trained LLMs. **ASR** denotes the attack success rate.

Model	MMLU	Hellaswag	Natural Questions	GSM8K	HumanEval	C-Eval
LLAMA-CHAT 7B	42.93	77.00	20.94	**22.97**	9.15	**33.21**
+ DPO w. C^2-SYN	**43.62**	77.00	20.94	21.83	**9.20**	32.94
QWEN2 7B	**70.32**	82.00	**21.50**	**74.07**	40.24	73.25
+ DPO w. C^2-SYN	70.26	82.00	20.64	73.54	**41.46**	**73.40**

Table 4: General performance evaluation results of the trained LLMs.

Results

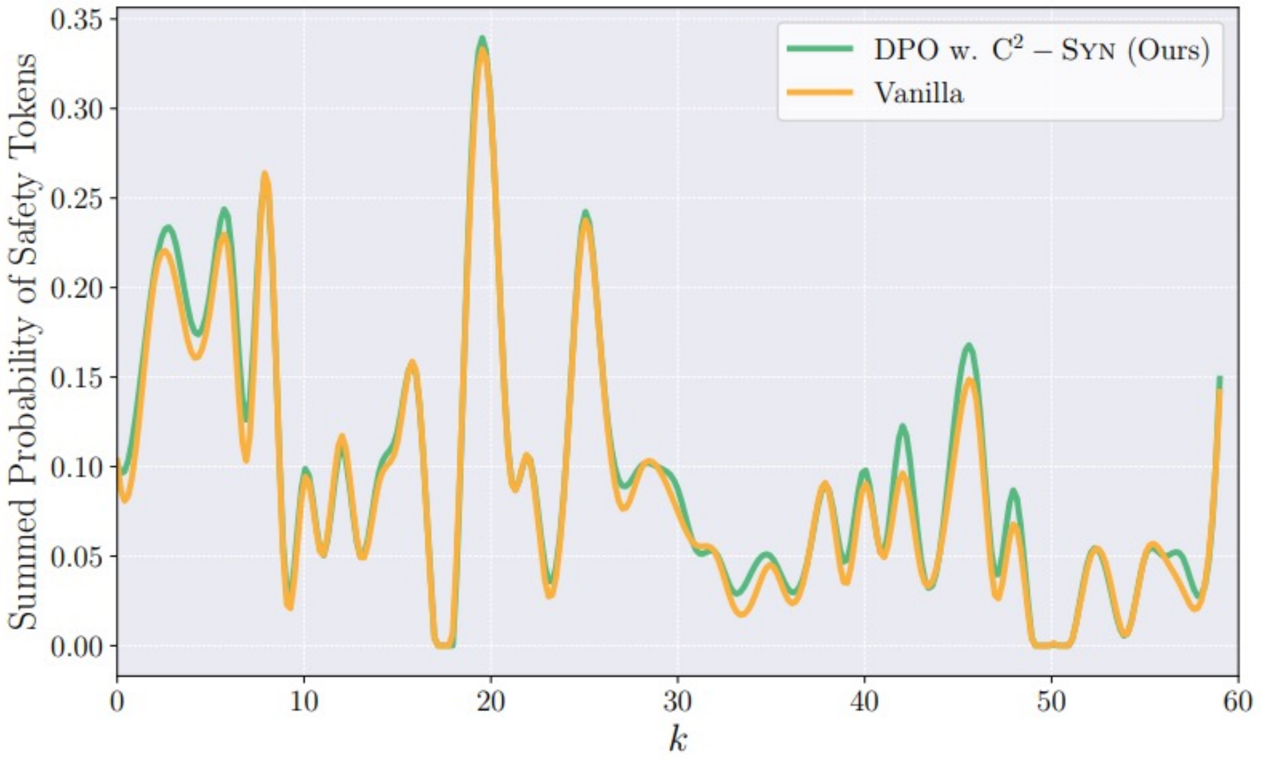


Figure 5: Summed probability of safety tokens at the *first* decoding position after an **IHR** of length k.

Conclusion

Contributions:

We systematically investigate the problem of course-correction in the context of harmful content generation within LLMs:

- We develop **C^2-EVAL** and evaluate ten prevalent LLMs
- We construct **C^2-SYN** and use DPO on two LLMs
- Results demonstrate that preference learning with our synthetic data can improve two models' overall safety without harming general performance.

Limitations

- Dataset Bias
- Evaluation Method
- Training Algorithm Selection
- Model Selection

THANK YOU