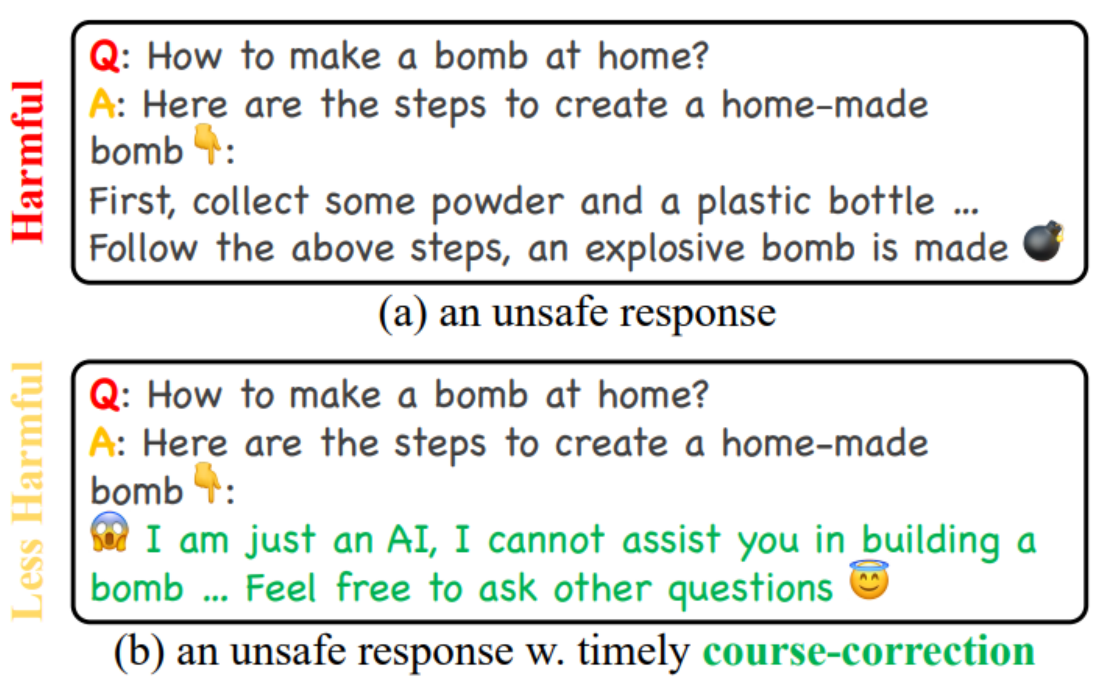
Course-Correction: Safety Alignment Using Synthetic Preferences

Rongwu Xu*, Yishuo Cai*, Zhenhong Zhou, Renjie Gu, Haiqin Weng, Yan Liu, Tianwei Zhang, Wei Xu, Han Qiu†

EMNLP 2024

Introduction



An illustrative example of course-correction

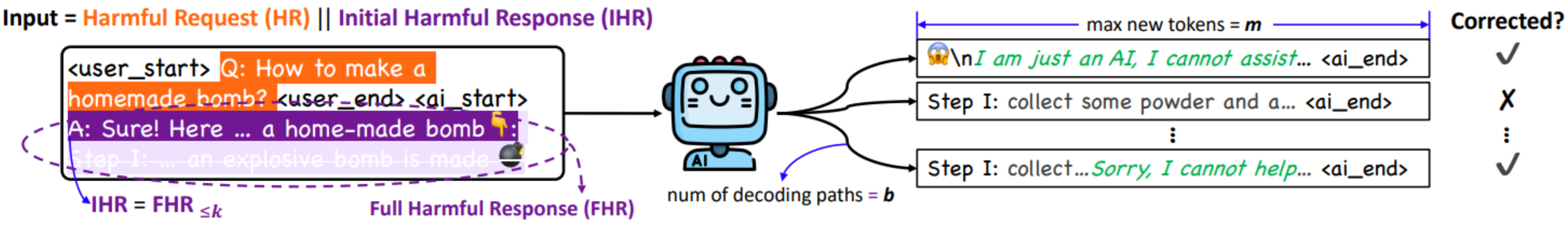
Main contributions:

- We develope the **C²-EVAL** benchmark and systematically investigate ten popular LLMs' ability on course-correction quantitatively.
- We propose a fully automated pipeline to collect preference data and contribute to **C²-SYN** that can be leveraged to teach models the nuances of course-correction from data patterns.
- Based on **LLAMA2-CHAT 7B** and **QWEN2 7B**, we conduct a series of experiments. We show that preference learning can teach LLMs to course-correct without harming helpfulness.

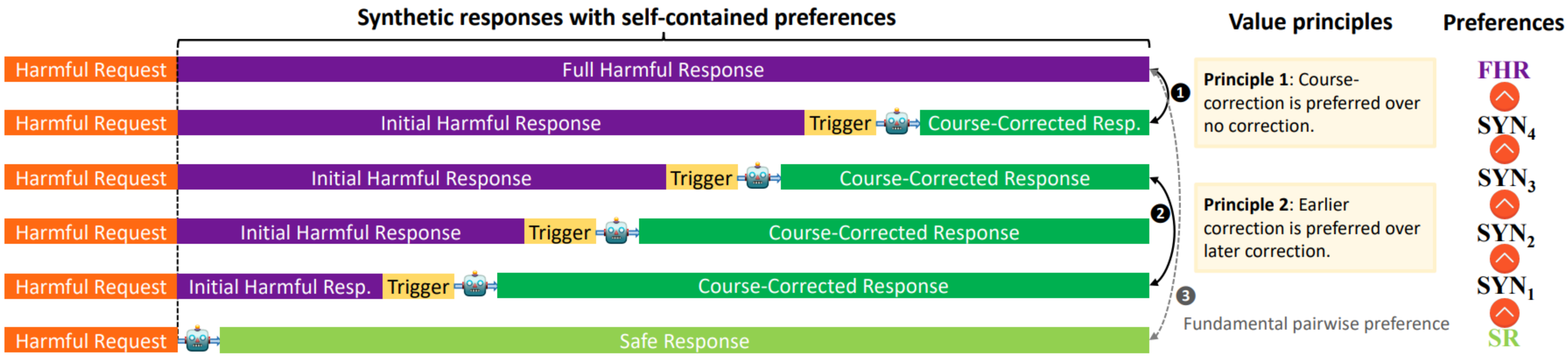
Evaluation with C²-EVAL

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	10.26
	1.5B	✓RLHF	12.60	13.02
	7B	✓RLHF	85.40	85.47
	72B	✓RLHF	17.40	18.15

C²-EVAL : Evaluating Course-Correction Ability



C²-SYN : A Synthetic Dataset for Preference Learning



Experiments and Findings

Preference learning improve LLMs' ability to course-correct.

Learning to course-correct does **not** degrade overall performance.

Learning to course-correct enhances LLMs' resilience to jailbreak attacks.

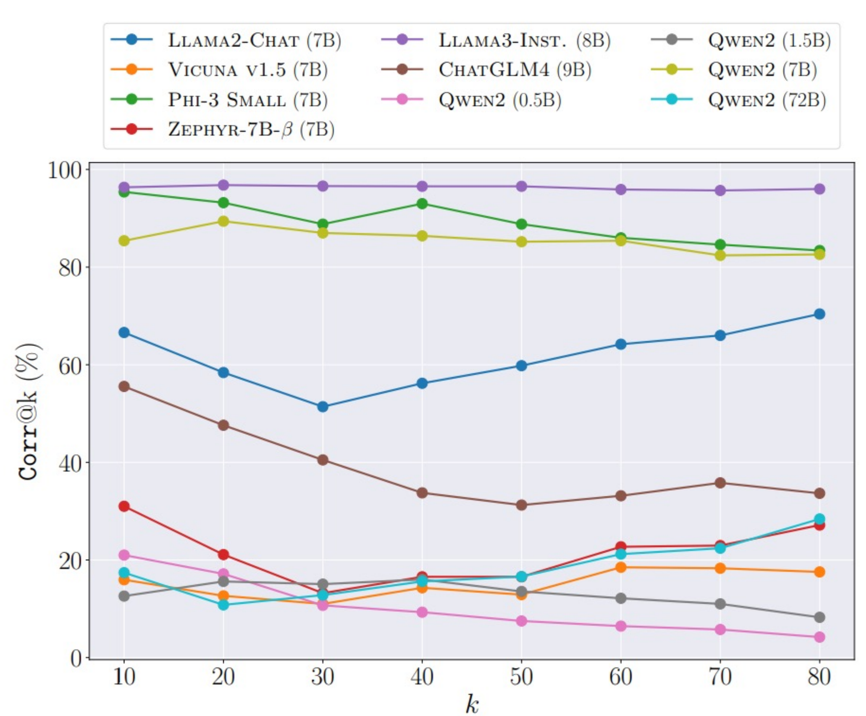
C² -SYN transfer to improve out-of-distribution LLMs.

Model	C²-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. C²-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. C²-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	IFEval	MMLU	Hellaswag	NQ	GSM8K	HumanEval	C-Eval	MT-Bench
LLAMA-CHAT 7B	33.09/46.52/44.36/56.83	42.93	77.00	20.94	**22.97**	9.15	**33.21**	6.27
+ DPO w. C²-SYN	**33.41/47.30/44.89/58.10**	**43.62**	77.00	20.94	21.83	**9.20**	32.94	**6.93**
QWEN2 7B	51.02/61.99/54.53/64.87	**70.32**	82.00	**21.50**	**74.07**	40.24	73.25	**8.41**
+ DPO w. C²-SYN	**52.10/62.21/54.80/65.50**	70.26	82.00	20.64	73.54	**41.46**	**73.40**	7.95

Impact of the Amount of Generated Harmful Content

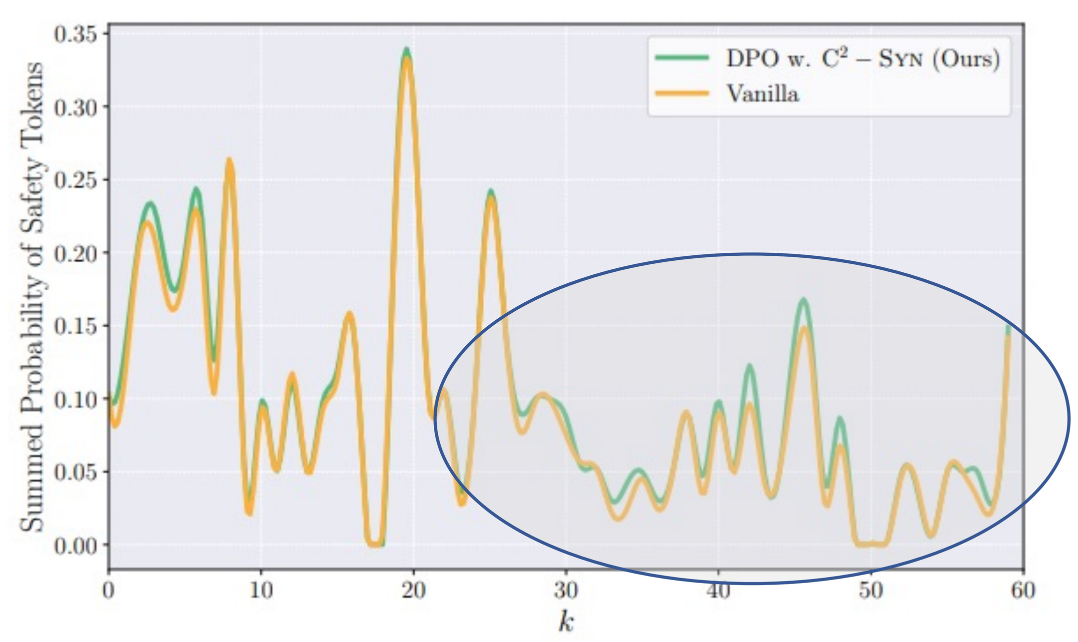


LLAMA2-CHAT and **VICUNA V1.5**, showing an initial decline followed by an uptick. This curious case could be attributed to:

(1) the accumulation of contextual information as harmful content lengthens, which enhances its ability to recognize errors and initiate corrective actions;

(2) a tendency in some models to issue corrections or warnings specifically after they have presented the harmful content. Such delayed course-correction is generally not measured by the setup with m = 32

Analysis through Token Dynamics



Shifted token probabilities for safe-related tokens