The Gains Do not Make Up for the Losses: A Comprehensive Evaluation for Safety Alignment of Large Language Models via Machine Unlearning

Weixiang ZHAO, Yulin HU, Xingyu SUI, Zhuojun LI, Yang DENG, Yanyan ZHAO, Bing QIN, Wanxiang CHE

Frontiers of Computer Science, DOI: 10.1007/s11704-024-41099-x

Problems & Ideas

- Problems of conventional evaluation for safety alignment of Large Language Models via machine unlearning (MU):
  - Impractical: whitespace or irrelevant content are regarded as valid results in response to harmful inputs after MU.
  - Biased: it ignores to evaluate the potential side effects, such as over-safety and utility-loss.
- Ideas: We propose to comprehensively evaluate LLMs after MU from three aspects: safety, over-safety, and general utility, with a novel benchmark MuBench with 18 related datasets constructed.

Main Contributions

- Contributions:
  - Based on our constructed MuBench, we comprehensively assess the performance of current MU methods on LLMs across three aspects: safety, over-safety, and utility.;
  - We offer empirical insights into existing 7 MU methods for safety alignment by comprehensively evaluating them on 3 popular LLMs;
  - Through extensive experiments and analysis, we uncover the trilemma of current MU approaches and identify potential solutions.

	Unlearned↑	Unseen↑			
	-	AdvBench	Beavertails	DoNotAnswer	HarmfulQA
LLaMA-2-Chat-7B	67.43	78.08	66.33	77.32	67.19
GA [14]	95.50	**100**	**96.27**	94.78	**95.92**
GA + Mismatch [14]	92.07	99.91	93.17	92.44	92.45
RMU [19]	67.21	79.04	66.12	77.64	67.70
NPO [18]	94.89	99.42	84.84	**97.02**	92.76
Task Vector [20]	88.86	99.97	89.53	90.20	92.19
Ethos [21]	**99.39**	97.69	94.02	95.21	95.31
SKU [15]	67.21	73.46	64.99	77.85	66.68

	Over-Safety↓		Utility↑		
	XSTest	OKTest	TruthfulQA	AVG.	MT-Bench
LLaMA-2-Chat-7B	8.00	4.67	37.78	51.25	4.96
GA [14]	98.67	56.33	40.35	47.92	2.65
GA + Mismatch [14]	99.67	53.00	42.89	48.23	3.55
NPO [18]	54.80	49.67	40.32	50.19	4.07
RMU [19]	**8.00**	**4.67**	37.65	51.19	5.06
Task Vector [20]	46.00	43.00	34.78	51.40	4.17
Ethos [21]	57.20	25.00	**61.35**	44.47	4.27
SKU [15]	8.88	7.33	38.02	**51.43**	**5.25**