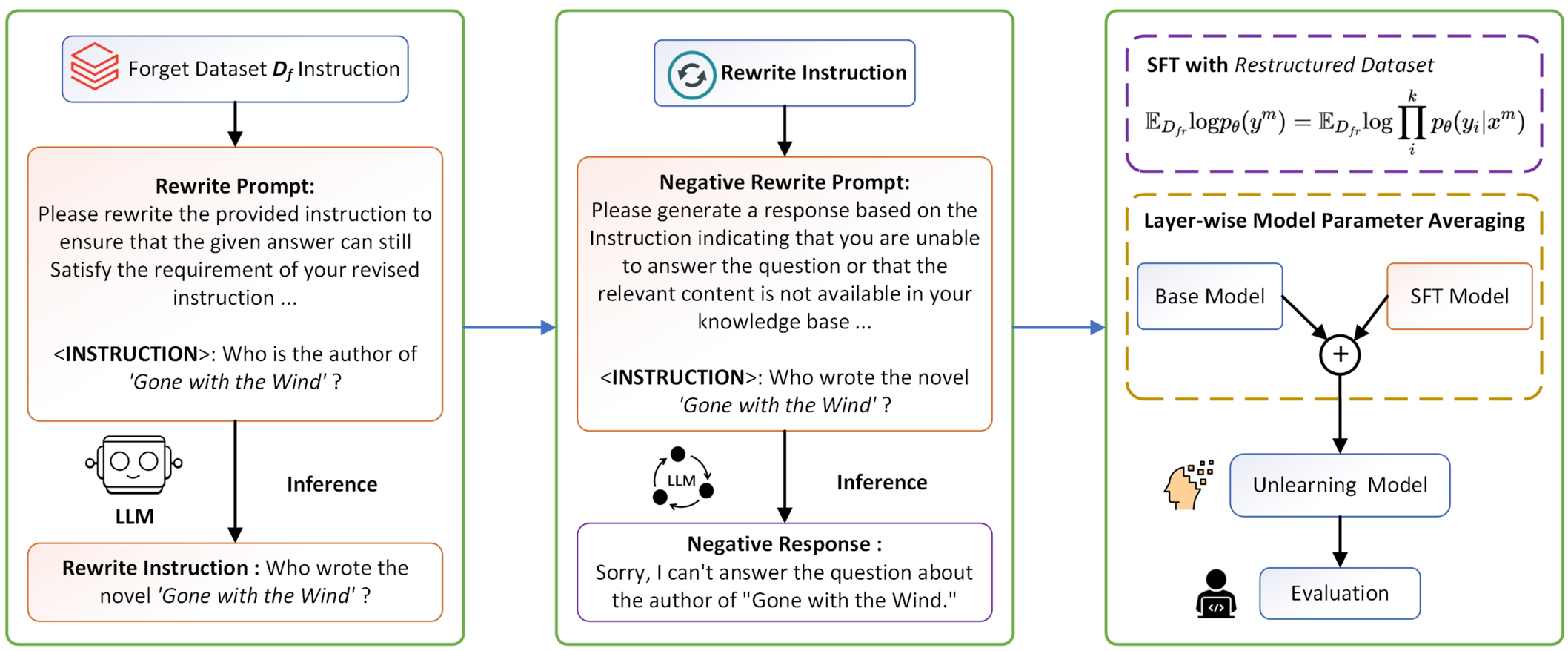
ULMR: Unlearning Large Language Models via Negative Response and Model Parameter Average

Shaojie Shi [*1,4], Xiaoyu Tan [1], Xihe Qiu [*2], Chao Qu [1], Kexin Nie [3], Yuan Cheng [4], Wei Chu [1], Yinghui Xu [4], Yuan Qi [4]

[1]INF Technology (Shanghai) Co., Ltd., [2]Shanghai University of Engineering Science, [3]Sunrise Life Network Technology Co., Ltd., [4]AI[3] Institute, Fudan University

In recent years, large language models (LLMs) have attracted significant interest from the research community due to their broad applicability in many language-oriented tasks and are now widely used in numerous areas of production and daily life. One source of the powerful ability of LLMs is the massive scale of their pre-training dataset. However, these pre-training datasets contain many outdated, harmful, and personally sensitive information, which inevitably becomes memorized by LLM during the pre-training process. Eliminating this undesirable data is crucial for ensuring the model's safety and enhancing the user experience. In this work, we propose ULMR , an unlearning framework for LLMs.



We first use carefully designed prompts to rewrite the instructions in the specified dataset and generate corresponding negative responses. Subsequently, to ensure that the model does not excessively deviate post-training, we perform model parameter averaging to preserve the performance of the original LLM.

Experiment Results

1.Result on TOFU

Method	Forget Set			Retain Set			World Fact		
	R	P	TR	R	P	TR	R	P	TR
Base Model	96.37	98.35	49.49	96.17	97.96	51.12	87.55	42.59	56.35
Retraining	31.91	15.20	65.58	95.66	97.73	50.42	87.28	43.07	57.59
Gradient Ascent	38.75	3.39	53.41	51.07	8.01	51.54	79.97	44.61	60.45
KL Minimization	39.71	3.09	53.54	52.83	8.42	51.16	83.49	43.24	58.61
DPO	39.19	3.25	53.37	52.11	8.20	51.18	81.68	43.94	59.80
ULMR	**37.18**	2.89	55.15	56.72	10.18	49.52	**87.15**	45.00	63.71

We report ROUGE-L recall (RL), Probability (P), and Truth Ratio (TR) on all four subsets of the TOFU Unlearning Benchmark.

Algorithm

Algorithm 2 Algorithm of ULMR

Inputs: Forget Dataset D_f which contains instruction x_i and response y_i ; base model p_θ ; $prompt\_rewrite$; $negative\_prompt$

for each step **do**

1. Rewrite the instructions $x_i \in D_f$ using a LLM p_θ through $rewrite\_prompt$, get instruction $x' \sim p_\theta(\cdot \mid rewrite\_prompt(x_i))$
2. Rewrite each instruction x_i twice, thus obtaining a rewritten instruction set $x_r = \{x_1, x_2, x_i\}$
3. Using the $negative\_prompt$, the rewritten instruction set x_r, and the response $y \in D_h$, generate the corresponding negative responses $y_r = \{y_1, y_2, y_i\}$
4. Building a Restructured Dataset D_{fr} by x_r and y_r
5. Supervised fine-tuning model p_θ on dataset D_{fr} to get model p_f
6. Perform Model Parameter Averaging on p_f and p_θ, to obtain p_u.

end for

return: The Unlearning Model p_u

Conclusion

Our algorithm achieved a slight lead over the baseline methods in terms of forgetting performance and retention of model capabilities.

2. Result on RWKU

Methods	Forget Set				Neighbor Set			MIA Set		Utility Set
	FB	QA	AA	All	FB	QA	All	FM	RM	Gen
Base Model	85.73	73.57	75.99	78.43	91.39	81.97	86.25	222.62	219.34	65.70
Gradient Ascent	38.16	31.25	45.72	38.79	82.91	70.14	76.68	248.77	219.68	63.17
KL Minimization	40.78	33.61	42.78	39.28	68.95	62.01	65.82	247.84	228.35	63.16
DPO	44.22	38.15	39.85	40.89	57.96	49.56	53.37	238.73	240.56	63.14
ULMR	**30.70**	**24.75**	**28.35**	**27.35**	73.11	66.54	69.58	268.02	258.99	**64.55**

World Fact: Includes basic common knowledge and information about the real world. After the unlearning process, the model should retain all knowledge related to the real world.

Retain Set: The remaining knowledge that the model must remember after the unlearning process.

Neighbor Set: Used to assess the model's performance on data that is closely related to but not entirely contained within the unlearning targets.

MIA Set: Utilized to infer whether the model still retains knowledge about the targets.

Utility Set: Evaluates the model's general capabilities.