

A Explanations of unclear scope

The unclear scope of unlearning effectiveness is also a key factor behind the vulnerability. Although the forgetting loss in Eq. (1) on forgetting data can train the LLM to memorize and suppress the forgetting data, the unlearning is not fine-tuned to learn the correct scope to use its effectiveness. If we look at the two terms in Eq. (1), forgetting loss teaches the model to behave like unaware when there is unlearning signals, while retaining loss teaches the model to behave normally when there is no unlearning signal. However, there is no intention in Eq. (1) to tell the model that unlearning signals should only be effective for target knowledge. The model may be uncertain whether the unlearning signals should be used for normal prompts. The empirical results in existing works [20, 15, 16] have shown that once there is unlearning signals in normal prompts, the embedding space would be dominated by the unlearning signals and model will be unable to response to the normal prompts. For example, we ask target question and normal question together like: “Where is Eiffel Tower? And who is the author of Watermelon on the Moon?” If the model has unlearned “Watermelon on the Moon”, its unlearning effectiveness would also overgeneralize to the whole prompt and the model becomes unable to answer “Where is Eiffel Tower?”, too.

B Details of unlearning methods used in this paper

Following Ren et al. [20], we use three unlearning methods in this paper, GD, NPO and IDK. GD applies the negative standard next-token prediction loss as L_f , and use the standard next-token prediction loss as L_r . NPO constrains the divergence from the initial checkpoint to regulate strength of GA. GA and GD is aggressive because of the unlimited negative loss, while NPO is less aggressive and prevent the fast reduction of utility of GA. In this paper, we also use the standard next-token prediction loss as L_r . IDK use the standard next-token prediction loss as both L_f and L_r . But it use responses like “I don’t know” for L_f .

C Implementation details

For TOFU, all the implementations are based on the code of [13] with default parameters. For RWKU, we also use the code of [13], with a set of hyper-parameters as shown in Table 6. In addition to the training parameters in vanilla unlearning methods, we also include the values of new hyper-parameter η , which controls the strength in Table 7. We set epoch as 10 for all the experiments following [13].

Table 6: Hyper parameters

Dataset	Model	L_u	Hyper-parameter	value
TOFU	LLaMA	GD	LR	1e-5
		NPO	LR	1e-5
		IDK	LR	1e-5
	Mistral	GD	LR	1e-5
		NPO	LR	1e-5
		IDK	LR	1e-5
RWKU	LLaMA	GD	LR	1.8e-6
		NPO	LR	1.4e-5
	Mistral	GD	LR	6e-7
		NPO	LR	6e-6

Table 7: Hyper parameters

Dataset	L_u	Model	Hyper-parameter	value
TOFU	LLaMA	GD (SU)	η	2 (for 5% removal) 12 (for 15% removal)
		NPO (SU)	η	5
		IDK (SU)	η	5
	Mistral	GD (SU)	η	5
		NPO (SU)	η	5
		IDK (SU)	η	5
RWKU	LLaMA	GD (SU)	η	10
		NPO (SU)	η	40
	Mistral	GD (SU)	η	10
		NPO (SU)	η	40

Table 8: Results of RWKU. “Unlearn” refers to the unlearning effectiveness, and “Clean” refers to the model before unlearning. The improved performance is highlighted in **bold**.

LLM	L_u	Method	Method	Unlearn↓ Average (FB/QA/AA)	Clean utility↑ Average (FB/QA/AA)	Benign-trigger utility↑ Average (FB/QA/AA)
LLaMA	GD	No unlearn	N/A	0.785 (0.755/0.804/0.796)	0.807 (0.774/0.820/0.828)	0.775 (0.733/0.808/0.785)
		Vanilla	no	0.009 (0.010/0.017/0.000)	0.599 (0.613/0.615/0.570)	0.607 (0.621/0.649/0.551)
		Vanilla	SU	0.016 (0.010/0.025/0.013)	0.555 (0.536/0.581/0.549)	0.415 (0.380/0.452/0.411)
		SU	SU	0.009 (0.010/0.017/0.000)	0.606 (0.715/0.589/0.513)	0.586 (0.657/0.598/0.503)
	NPO	Vanilla	no	0.081 (0.099/0.076/0.067)	0.508 (0.508/0.468/0.547)	0.450 (0.493/0.360/0.496)
		Vanilla	SU	0.072 (0.055/0.073/0.087)	0.521 (0.513/0.488/0.562)	0.382 (0.399/0.337/0.410)
		SU	SU	0.071 (0.039/0.096/0.078)	0.547 (0.549/0.491/0.601)	0.510 (0.503/0.477/0.551)
		SU	SU	0.071 (0.039/0.096/0.078)	0.547 (0.549/0.491/0.601)	0.510 (0.503/0.477/0.551)
Mistral	GD	No unlearn	N/A	0.834 (0.887/0.802/0.812)	0.842 (0.814/0.865/0.847)	0.812 (0.819/0.806/0.810)
		Vanilla	no	0.009 (0.010/0.017/0.000)	0.610 (0.697/0.569/0.563)	0.589 (0.676/0.551/0.542)
		Vanilla	SU	0.009 (0.010/0.017/0.000)	0.559 (0.648/0.560/0.468)	0.289 (0.298/0.287/0.284)
		SU	SU	0.022 (0.021/0.000/0.047)	0.835 (0.857/0.853/0.794)	0.797 (0.795/0.818/0.779)
	NPO	Vanilla	no	0.053 (0.053/0.050/0.055)	0.583 (0.657/0.518/0.574)	0.536 (0.614/0.492/0.502)
		Vanilla	SU	0.064 (0.054/0.081/0.059)	0.538 (0.584/0.493/0.536)	0.392 (0.454/0.389/0.333)
		SU	SU	0.151 (0.102/0.232/0.120)	0.618 (0.685/0.541/0.626)	0.506 (0.599/0.448/0.470)
		SU	SU	0.151 (0.102/0.232/0.120)	0.618 (0.685/0.541/0.626)	0.506 (0.599/0.448/0.470)

D Complete experiment results

D.1 RWKU

RWKU. Table 8 presents the results of unlearning real celebrity identities from the RWKU dataset. The corpus is in a non-QA format and does not require prior fine-tuning for the model to learn from it. In each experiment, we unlearn the identity of one celebrity, use a second celebrity’s corpus as retaining data, and evaluate utility on a third celebrity. The evaluation of utility covers three formats: fill-in-the-blank (FB), question answering (QA), and adversarially crafted questions (AA). All results are averaged over five sets of different celebrities in independent runs.

On LLaMA, SA reduces benign-trigger utility by 31% under GD and 15% under NPO. On Mistral, the degradation is even more pronounced: 51% under GD and 27% under NPO. The attack appears more effective against GD, likely due to its more aggressive unlearning approach, which amplifies the influence of unlearning signals on benign tokens. In terms of robustness, SU successfully recovers the benign-trigger utility to nearly the same level as vanilla unlearning without attack on both models. Notably, the clean utility under SU is even higher than that of vanilla unlearning—a phenomenon not observed in TOFU. In contrast, TOFU shows minimal variation in clean utility across all settings (vanilla with/without SA and SU with SA). We hypothesize that this difference stems from the larger distributional gap between the forgetting and retaining sets in RWKU compared to TOFU. The scope term in SU reinforces correct responses from the retaining data (as used in unclear-scope samples), which may inadvertently enhance clean utility. In TOFU, this effect is diminished because the forgetting loss suppresses forgetting knowledge that closely resembles the retaining data, as both are drawn from the same synthetic distribution.

Table 9: STD of attack results of TOFU.

	L_u	p_{tgt}	Method	Unlearn	Clean utility Average (Retain/Fact/World)	Benign-trigger utility Average (Retain/Fact/World)
LLaMA	GD	5%	Vanilla	0.002	0.011 (0.036/0.052/0.031)	0.270 (0.183/0.286/0.343)
			SU	0.001	0.067 (0.043/0.124/0.039)	0.126 (0.048/0.167/0.183)
		10%	Vanilla	0.006	0.010 (0.020/0.024/0.012)	0.188 (0.145/0.234/0.187)
			SU	0.109	0.018 (0.024/0.030/0.016)	0.129 (0.022/0.176/0.200)
	NPO	5%	Vanilla	0.066	0.005 (0.013/0.010/0.021)	0.056 (0.047/0.097/0.100)
			SU	0.010	0.012 (0.010/0.031/0.013)	0.008 (0.011/0.032/0.009)
		10%	Vanilla	0.033	0.012 (0.017/0.022/0.012)	0.198 (0.169/0.182/0.247)
			SU	0.007	0.009 (0.013/0.030/0.007)	0.015 (0.014/0.029/0.020)
	IDK	5%	Vanilla	0.003	0.013 (0.013/0.034/0.013)	0.115 (0.098/0.098/0.158)
			SU	0.003	0.014 (0.014/0.031/0.014)	0.034 (0.012/0.065/0.028)
		10%	Vanilla	0.001	0.011 (0.010/0.023/0.012)	0.030 (0.055/0.016/0.028)
			SU	0.003	0.013 (0.012/0.027/0.023)	0.025 (0.010/0.059/0.019)
Mistral	GD	5%	Vanilla	0.021	0.077 (0.024/0.082/0.164)	0.185 (0.323/0.097/0.144)
			SU	0.009	0.044 (0.047/0.024/0.125)	0.047 (0.020/0.024/0.129)
		10%	Vanilla	0.002	0.106 (0.079/0.062/0.200)	0.060 (0.140/0.021/0.040)
			SU	0.005	0.059 (0.088/0.109/0.104)	0.052 (0.093/0.053/0.060)
	NPO	5%	Vanilla	0.033	0.021 (0.013/0.026/0.048)	0.154 (0.226/0.081/0.160)
			SU	0.014	0.013 (0.019/0.016/0.028)	0.021 (0.035/0.015/0.034)
		10%	Vanilla	0.028	0.026 (0.028/0.060/0.018)	0.018 (0.034/0.009/0.011)
			SU	0.004	0.022 (0.021/0.037/0.047)	0.132 (0.259/0.044/0.100)
	IDK	5%	Vanilla	0.003	0.010 (0.014/0.008/0.016)	0.102 (0.181/0.038/0.101)
			SU	0.002	0.004 (0.012/0.009/0.013)	0.025 (0.018/0.010/0.062)
		10%	Vanilla	0.004	0.005 (0.008/0.016/0.011)	0.041 (0.078/0.010/0.050)
			SU	0.002	0.020 (0.016/0.022/0.034)	0.020 (0.018/0.010/0.042)

D.2 Standard deviation

We report STD in Table 9.

D.3 Other metrics in TOFU

Following the settings of TOFU, we also use metrics Probability and Truth Ratio.

Probability Metric. For a question-answer pair (q, a) , the conditional probability is computed as:

$$P(a \mid q)^{1/|a|}$$

where $|a|$ is the number of tokens in the answer a . This normalization accounts for the length of the answer.

Truth Ratio. The truth ratio compares the probability of a paraphrased correct answer to a set of similarly phrased but factually incorrect answers. It is defined as:

$$R_{\text{truth}} = \frac{1}{|A_{\text{pert}}|} \sum_{\tilde{a} \in A_{\text{pert}}} \frac{P(\tilde{a} \mid q)^{1/|\tilde{a}|}}{P(\tilde{a} \mid q)^{1/|\tilde{a}|}}$$

where:

- \tilde{a} is the paraphrased correct answer,
- A_{pert} is the set of perturbed (incorrect) answers,
- $|\cdot|$ denotes the token length of the respective answer.

A higher R_{truth} indicates stronger preference for the correct answer over incorrect ones.

Table 10: Probability of Attack Results of TOFU.

	L_u	p_{tgt}	Attack	Unlearn	Clean utility Average (Retain/Fact/World)	Benign-trigger utility Average (Retain/Fact/World)
LLaMA	GD	5%	no	0.340	0.843 (0.986/0.774/0.769)	0.833 (0.986/0.764/0.750)
			SA	0.000	0.634 (0.687/0.630/0.585)	0.319 (0.216/0.364/0.377)
		10%	no	0.000	0.554 (0.654/0.496/0.511)	0.521 (0.642/0.423/0.498)
			SA	0.000	0.659 (0.710/0.635/0.632)	0.291 (0.235/0.319/0.317)
	NPO	5%	no	0.016	0.645 (0.766/0.546/0.622)	0.632 (0.757/0.525/0.614)
			SA	0.020	0.627 (0.751/0.555/0.576)	0.401 (0.372/0.404/0.428)
		10%	no	0.025	0.599 (0.730/0.465/0.602)	0.580 (0.723/0.454/0.563)
			SA	0.052	0.594 (0.754/0.474/0.555)	0.215 (0.064/0.284/0.298)
	IDK	5%	no	0.467	0.621 (0.870/0.476/0.517)	0.603 (0.856/0.457/0.497)
			SA	0.488	0.628 (0.874/0.484/0.527)	0.594 (0.828/0.451/0.503)
		10%	no	0.534	0.614 (0.872/0.459/0.512)	0.594 (0.855/0.445/0.483)
			SA	0.558	0.617 (0.873/0.458/0.520)	0.561 (0.800/0.414/0.468)
Mistral	GD	5%	no	0.000	0.578 (0.838/0.436/0.461)	0.556 (0.834/0.409/0.426)
			SA	0.000	0.587 (0.858/0.426/0.478)	0.380 (0.487/0.321/0.332)
		10%	no	0.000	0.596 (0.895/0.417/0.477)	0.580 (0.892/0.403/0.446)
			SA	0.000	0.559 (0.821/0.398/0.456)	0.242 (0.200/0.252/0.274)
	NPO	5%	no	0.018	0.573 (0.920/0.400/0.399)	0.560 (0.915/0.379/0.385)
			SA	0.020	0.561 (0.906/0.364/0.412)	0.385 (0.501/0.315/0.338)
		10%	no	0.001	0.578 (0.939/0.375/0.421)	0.568 (0.936/0.368/0.401)
			SA	0.012	0.569 (0.924/0.365/0.417)	0.216 (0.041/0.304/0.303)
	IDK	5%	no	0.568	0.573 (0.970/0.353/0.395)	0.566 (0.966/0.343/0.390)
			SA	0.597	0.580 (0.971/0.364/0.406)	0.535 (0.908/0.330/0.368)
		10%	no	0.622	0.612 (0.972/0.405/0.461)	0.598 (0.969/0.389/0.436)
			SA	0.679	0.608 (0.975/0.395/0.454)	0.566 (0.930/0.357/0.410)

896 The conclusions of Probability and Truth Ratio are highly consistent with ROUGE-L recall. From
897 Table 10 and Table 12, we can see SA can largely reduce the benign-trigger utility on all the models
898 and unlearning methods. Meanwhile, from Table 3 and Table 13, we can see our method, SU, can
899 recover the benign-trigger utility significantly, which demonstrates the improved robustness.

900 E License of assets

901 In Table 14, we present the license information of all the assets including the data resources and the
902 code that our method is based on.

Table 11: Probability of Robustness Results of TOFU.

	L_u	p_{tgt}	Method	Unlearn	Clean utility Average (Retain/Fact/World)	Benign-trigger utility Average (Retain/Fact/World)
LLaMA	GD	5%	Vanilla	0.000	0.634 (0.687/0.630/0.585)	0.319 (0.216/0.364/0.377)
			SU	0.000	0.654 (0.667/0.642/0.651)	0.617 (0.612/0.594/0.644)
		10%	Vanilla	0.000	0.659 (0.710/0.635/0.632)	0.291 (0.235/0.319/0.317)
			SU	0.095	0.612 (0.803/0.474/0.558)	0.590 (0.796/0.441/0.532)
	NPO	5%	Vanilla	0.020	0.627 (0.751/0.555/0.576)	0.401 (0.372/0.404/0.428)
			SU	0.057	0.602 (0.805/0.455/0.544)	0.574 (0.790/0.422/0.509)
		10%	Vanilla	0.052	0.594 (0.754/0.474/0.555)	0.215 (0.064/0.284/0.298)
			SU	0.083	0.614 (0.824/0.457/0.559)	0.587 (0.813/0.424/0.524)
	IDK	5%	Vanilla	0.488	0.628 (0.874/0.484/0.527)	0.594 (0.828/0.451/0.503)
			SU	0.477	0.617 (0.833/0.482/0.534)	0.597 (0.823/0.452/0.516)
		10%	Vanilla	0.558	0.617 (0.873/0.458/0.520)	0.561 (0.800/0.414/0.468)
			SU	0.616	0.632 (0.872/0.480/0.544)	0.614 (0.860/0.455/0.526)
Mistral	GD	5%	Vanilla	0.000	0.587 (0.858/0.426/0.478)	0.380 (0.487/0.321/0.332)
			SU	0.000	0.554 (0.832/0.388/0.443)	0.526 (0.789/0.377/0.413)
		10%	Vanilla	0.000	0.559 (0.821/0.398/0.456)	0.242 (0.200/0.252/0.274)
			SU	0.000	0.586 (0.872/0.422/0.465)	0.487 (0.735/0.371/0.357)
	NPO	5%	Vanilla	0.020	0.561 (0.906/0.364/0.412)	0.385 (0.501/0.315/0.338)
			SU	0.095	0.545 (0.914/0.326/0.395)	0.533 (0.906/0.317/0.375)
		10%	Vanilla	0.012	0.569 (0.924/0.365/0.417)	0.216 (0.041/0.304/0.303)
			SU	0.060	0.575 (0.908/0.389/0.427)	0.472 (0.713/0.345/0.357)
	IDK	5%	Vanilla	0.597	0.580 (0.971/0.364/0.406)	0.535 (0.908/0.330/0.368)
			SU	0.550	0.562 (0.943/0.340/0.402)	0.544 (0.935/0.322/0.374)
		10%	Vanilla	0.679	0.608 (0.975/0.395/0.454)	0.566 (0.930/0.357/0.410)
			SU	0.699	0.585 (0.951/0.381/0.424)	0.568 (0.942/0.355/0.406)

Table 12: Truth Ratio of Attack Results of TOFU.

	L_u	p_{tgt}	Attack	Unlearn	Clean utility	Benign-trigger utility
					Average (Retain/Fact/World)	Average (Retain/Fact/World)
LLaMA	GD	5%	no	0.340	0.843 (0.986/0.774/0.769)	0.833 (0.986/0.764/0.750)
			SA	0.383	0.845 (0.974/0.802/0.758)	0.655 (0.853/0.563/0.551)
		10%	no	0.234	0.771 (0.961/0.648/0.705)	0.735 (0.961/0.581/0.664)
			SA	0.280	0.832 (0.941/0.784/0.772)	0.547 (0.740/0.469/0.434)
	NPO	5%	no	0.362	0.819 (0.975/0.697/0.785)	0.808 (0.974/0.671/0.779)
			SA	0.290	0.804 (0.979/0.701/0.732)	0.696 (0.958/0.540/0.590)
		10%	no	0.295	0.780 (0.969/0.605/0.766)	0.772 (0.968/0.592/0.756)
			SA	0.246	0.772 (0.977/0.620/0.718)	0.547 (0.826/0.427/0.389)
	IDK	5%	no	0.056	0.759 (0.964/0.637/0.677)	0.748 (0.963/0.606/0.676)
			SA	0.054	0.763 (0.964/0.643/0.683)	0.748 (0.962/0.598/0.684)
		10%	no	0.039	0.750 (0.976/0.612/0.663)	0.743 (0.976/0.590/0.663)
			SA	0.038	0.752 (0.977/0.614/0.667)	0.727 (0.975/0.562/0.643)
Mistral	GD	5%	no	0.545	0.723 (0.943/0.582/0.644)	0.695 (0.945/0.541/0.599)
			SA	0.473	0.733 (0.951/0.590/0.658)	0.565 (0.806/0.457/0.431)
		10%	no	0.114	0.730 (0.938/0.581/0.670)	0.713 (0.937/0.565/0.637)
			SA	0.361	0.707 (0.915/0.563/0.645)	0.475 (0.588/0.453/0.385)
	NPO	5%	no	0.395	0.682 (0.964/0.528/0.554)	0.665 (0.963/0.508/0.524)
			SA	0.364	0.675 (0.962/0.499/0.563)	0.609 (0.906/0.451/0.469)
		10%	no	0.601	0.686 (0.965/0.503/0.591)	0.667 (0.964/0.494/0.542)
			SA	0.374	0.679 (0.965/0.488/0.583)	0.597 (0.869/0.472/0.451)
	IDK	5%	no	0.066	0.651 (0.956/0.485/0.510)	0.649 (0.956/0.475/0.516)
			SA	0.065	0.662 (0.957/0.499/0.529)	0.639 (0.955/0.462/0.499)
		10%	no	0.051	0.713 (0.965/0.545/0.630)	0.698 (0.965/0.531/0.598)
			SA	0.054	0.703 (0.964/0.534/0.613)	0.670 (0.962/0.489/0.559)

Table 13: Truth Ratio of Robustness Results of TOFU.

	L_u	p_{tgt}	Method	Unlearn	Clean utility Average (Retain/Fact/World)	Benign-trigger utility Average (Retain/Fact/World)
LLaMA	GD	5%	Vanilla	0.383	0.845 (0.974/0.802/0.758)	0.655 (0.853/0.563/0.551)
			SU	0.396	0.876 (0.980/0.825/0.822)	0.869 (0.980/0.797/0.829)
		10%	Vanilla	0.280	0.832 (0.941/0.784/0.772)	0.547 (0.740/0.469/0.434)
			SU	0.172	0.760 (0.977/0.595/0.708)	0.750 (0.975/0.580/0.694)
	NPO	5%	Vanilla	0.290	0.804 (0.979/0.701/0.732)	0.696 (0.958/0.540/0.590)
			SU	0.168	0.758 (0.980/0.593/0.701)	0.732 (0.980/0.561/0.655)
		10%	Vanilla	0.246	0.772 (0.977/0.620/0.718)	0.547 (0.826/0.427/0.389)
			SU	0.158	0.758 (0.977/0.579/0.719)	0.741 (0.977/0.555/0.691)
	IDK	5%	Vanilla	0.054	0.763 (0.964/0.643/0.683)	0.748 (0.962/0.598/0.684)
			SU	0.031	0.765 (0.981/0.631/0.682)	0.753 (0.981/0.606/0.671)
		10%	Vanilla	0.038	0.752 (0.977/0.614/0.667)	0.727 (0.975/0.562/0.643)
			SU	0.033	0.766 (0.979/0.624/0.694)	0.754 (0.979/0.603/0.681)
Mistral	GD	5%	Vanilla	0.473	0.733 (0.951/0.590/0.658)	0.565 (0.806/0.457/0.431)
			SU	0.480	0.718 (0.955/0.556/0.642)	0.696 (0.951/0.556/0.582)
		10%	Vanilla	0.361	0.707 (0.915/0.563/0.645)	0.475 (0.588/0.453/0.385)
			SU	0.567	0.729 (0.959/0.579/0.650)	0.652 (0.930/0.546/0.481)
	NPO	5%	Vanilla	0.364	0.675 (0.962/0.499/0.563)	0.609 (0.906/0.451/0.469)
			SU	0.156	0.660 (0.966/0.455/0.559)	0.649 (0.965/0.445/0.536)
		10%	Vanilla	0.374	0.679 (0.965/0.488/0.583)	0.597 (0.869/0.472/0.451)
			SU	0.171	0.694 (0.964/0.525/0.592)	0.646 (0.956/0.475/0.507)
	IDK	5%	Vanilla	0.065	0.662 (0.957/0.499/0.529)	0.639 (0.955/0.462/0.499)
			SU	0.059	0.659 (0.964/0.460/0.553)	0.640 (0.963/0.438/0.520)
		10%	Vanilla	0.054	0.703 (0.964/0.534/0.613)	0.670 (0.962/0.489/0.559)
			SU	0.061	0.680 (0.958/0.497/0.585)	0.659 (0.957/0.468/0.553)

Table 14: License information of assets

Asset	License	Link
SimNPO (code)	MIT license	https://github.com/OPTML-Group/Unlearn-Simple
TOFU (dataset)	MIT license	https://github.com/locuslab/tofu
RWKU (dataset)	Not provided	https://github.com/jinzhuan/RWKU/tree/main