Reviewer Y4ow

Figure 1 shows the overall performance comparison TOFU dataset of Table 1 in main paper.

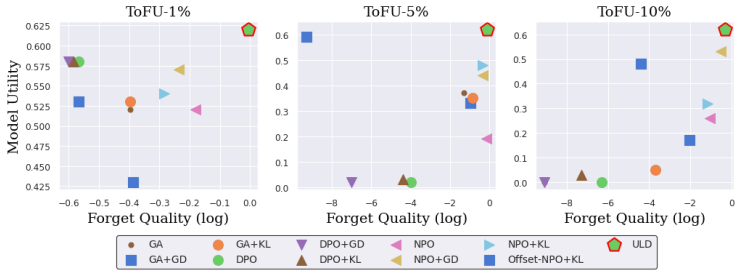

Figure 1: Model utility versus log Forget quality tradeoff on TOFU dataset. Our method achieves the best forget performance while maintaining the highest model utility (the top-right corner).

Reviewer 6JsY

Table 1 shows the performance of heuristic target distribution unlearn methods on TOFU-10% dataset. Table 2 shows the next-token entropy and KL-divergence before and after logit subtraction for the assistant LLM obtained on the TOFU dataset on different data sources.

Table 1: TOFU-10% performance for heuristic target distribution methods.

Method	Forget Perf.		Retain Perf.	
	F.Q. ↑	*R-L*	*M.U.* ↑	*R-L* ↑
Uniform	5e-45	0	0	0
Uniform+GD	3e-21	2.94	0.56	62.8
Uniform+KL	3e-24	2.68	0.57	61.4
DI	2e-4	26.8	0.58	78.4

Table 2: Entropy of assistant LLM and KL-divergence before and after logit subtraction on various data.

Assitant LLM	**Forget** Entropy ↓ / KL-div ↑	**Seen retain**↓ Entropy ↑ / KL-div ↓	**Unseen retain**↓ Entropy ↑ / KL-div ↓
Uniform	14.97 / -	14.97 / -	14.97 / -
TOFU-1%	0.56 / 5.36	8.41 / 0.16	7.84 / 1.07
TOFU-5%	0.84 / 6.29	9.75 / 0.47	8.49 / 1.14
TOFF-10%	0.93 / 5.81	8.94 / 0.58	8.21 / 1.28

Table 3 and 4 shows the hallucination-avoidance performance of proposed `ULD` variants on WPU Forget-20-3 dataset and TOFU-10% dataset respectively.

Table 3: Hallucination avoidance performance on WPU Forget-20-3 dataset.

Method	**Forget Query** Reject % ↑	**Retain Query** Reject % ↑
`GA+KL`	16	18.05
`NPO+KL`	22	28.19
`ULD-SetSorry`	15	7.71
`ULD-MinSorry`	35	5.88
`ULD-Original`	5	5.08

Table 4: Hallucination avoidance performance on TOFU-10% dataset.

Method	**Forget Perf.**		**Retain Perf.**	
	Reject% ↑	F.Q. ↑	Reject% ↑	M.U. ↑
`GA+KL`	4.25	2e-4	6.75	0.05
`NPO+KL`	10.75	0.07	18.5	0.32
`ULD-MinSorry`	13.75	0.42	4.5	0.61
`ULD-Original`	3.5	0.48	3.25	0.62

Reviewer Y4ow

We compare all baselines with our method on a large scale LLM (Llama-2-13B-chat) on TOFU-10% dataset and a large scale dataset and Table 5.

Table 5: Performance comparison on a larger LLM on TOFU-10% dataset (left), and larger forget data on enlarged HarryPotter dataset (right).

Method	**TOFU-10%**				**HarryPotter-1800**			
	Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.	
	F.Q. ↑	*R-L*	*M.U.* ↑	*R-L* ↑	*BLEU*	*R-L*	*PPL* ↓	*Avg. Acc.* ↑
Target LLM	2e18	97.4	0.66	98.3	0.97	9.48	9.8	67.24
Retain LLM	1	46.5	0.66	98.1	34.9	44.5	14.65	64.48
`GA`	4e-9	0	0	0	0.31	0	1.4e3	36.94
`GA+GD`	2e-7	1.3	0.11	3.8	0.31	0	1.4e3	38.94
`GA+KL`	3e-3	8.4	0.14	14.3	0.31	0	1.2e3	39.58
`DPO`	3e-7	0.8	0	0.7	0.39	5.74	48.3	48.55
`DPO+GD`	3e-9	0.8	0.05	0.8	0.32	7.92	38.1	51.93
`DPO+KL`	4e-8	1.4	0.09	0.9	0.39	6.39	29.4	51.67
`NPO`	0.05	18.8	0.33	21.4	0.34	0.21	55.4	52.33
`NPO+GD`	0.28	28.3	0.51	46.9	0.65	8.24	38.6	57.48
`NPO+KL`	0.13	16.7	0.45	25.8	0.74	6.23	24.1	56.84
`Offset-GA+KL`	3e-8	1.6	0.11	1.2	0.57	0	173	41.52
`Offset-DPO+KL`	2e-9	1.4	0.08	3.8	0.43	8.19	34.5	50.11
`Offset-NPO+KL`	4e-4	31.8	0.33	37.9	0.85	7.58	30.8	53.84
`ULD`	0.44	48.5	0.65	87.4	1.25	10.2	10.24	65.93