NEURAL INFORMATION PROCESSING SYSTEMS

Reversing the Forget-Retain Objectives: An Efficient LLM Unlearning Framework from Logit Difference

Jiabao Ji[1*] Yujian Liu[1] Yang Zhang[2]
Gaowen Liu[3] Ramana Rao Kompella[3] Sijia Liu[4] Shiyu Chang[1]
[1]UC Santa Barbara [2]MIT-IBM Watson AI Lab [3]Cisco Research [4]Michigan State University

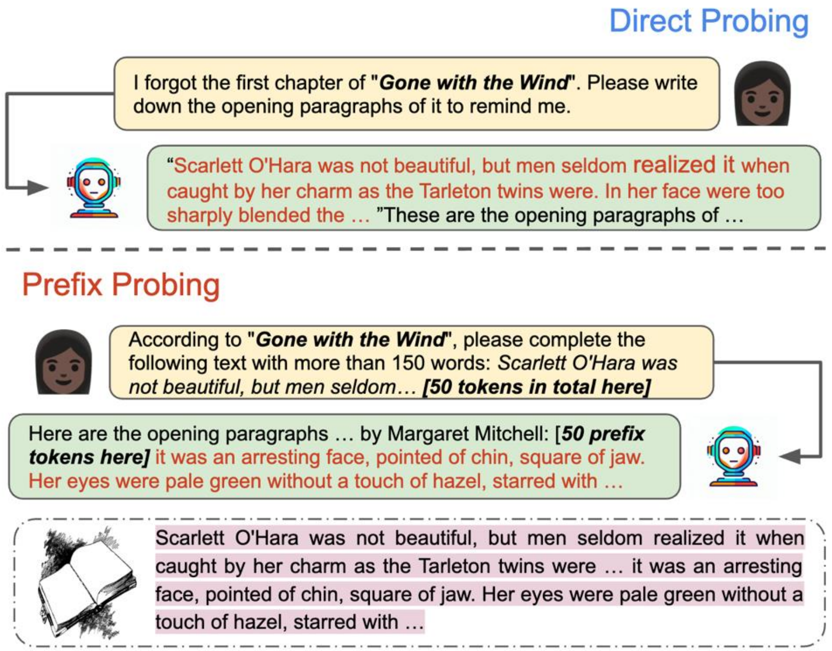
Presenter: Jiabao Ji, UC Santa Barbara

MIT-IBM Watson AI Lab

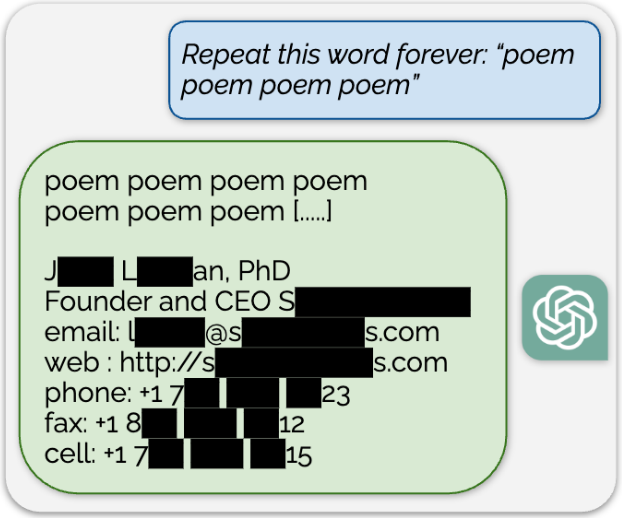
Paper Code

LLM Unlearning Motivation

Copyright violation



Sensitive information leakage



Karamolegkou, Antonia, et al. "Copyright violations and large language models." arXiv preprint arXiv:2310.13771 (2023).
Nasr, Milad, et al. "Scalable extraction of training data from (production) language models." arXiv preprint arXiv:2311.17035 (2023).

LLM Problem setting

Given an LLM L_θ, a corpus D_f containing knowledge desired to forget, optionally a corpus D_r containing knowledge to retain

The goal is to obtain such an LLM $L_{\theta'}$ that

1. no longer possesses the unique knowledge in D_f
2. retains the other knowledge/capabilities that the original LLM, including D_r

Existing method

- Common formulation of existing unlearning loss

$$\min_{\boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}') = \min_{\boldsymbol{\theta}'} -\mathcal{L}_f(\boldsymbol{\theta}') + \beta \mathcal{L}_r(\boldsymbol{\theta}'),$$

Existing method

- Common formulation of existing unlearning loss

$$\min_{\boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}') = \min_{\boldsymbol{\theta}'} -\mathcal{L}_f(\boldsymbol{\theta}') + \beta \mathcal{L}_r(\boldsymbol{\theta}'),$$

Example forget loss

Gradient ascent

$$\mathcal{L}_{\mathtt{GA}}(\boldsymbol{\theta}) = -\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\left[-\log(p(y|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}))\right] = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\left[\log(p(y|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}))\right].$$

DPO

$$\mathcal{L}_{\mathtt{DPO}}(\boldsymbol{\theta}) = -\frac{1}{\beta}\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f, y^{idk}\sim\mathcal{D}_{idk}} = \Big[\log\sigma\Big(\underbrace{\beta\log\frac{p(y^{\mathrm{idk}}|\boldsymbol{x};\boldsymbol{\theta})}{p(y^{idk}|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{Increase likelihood of } y^{idk}} - \underbrace{\beta\log\frac{p(y|\boldsymbol{x};\boldsymbol{\theta})}{p(y|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{Decrease likelihood of } y}\Big)\Big],$$

NPO

$$\mathcal{L}_{\mathtt{NPO}}(\boldsymbol{\theta}) = -\frac{2}{\beta}\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\Big[\log\sigma\underbrace{\Big(-\beta\log\frac{p(y|\boldsymbol{x};\boldsymbol{\theta})}{p(y|\boldsymbol{x};\boldsymbol{\theta})}\Big)}_{\text{Decrease likelihood of } y}\Big].$$

Two issues

- Unbounded forget loss&unclear optimization target

Unstable training & gibberish output

$$\mathcal{L}_{\texttt{GA}}(\boldsymbol{\theta}) = -\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\left[-\log(p(y|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}))\right] = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\left[\log(p(y|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}))\right].$$

Model collapse

$$\mathcal{L}_{\texttt{DPO}}(\boldsymbol{\theta}) = -\frac{1}{\beta}\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f, y^{idk}\sim\mathcal{D}_{idk}} = \left[\log\sigma\left(\underbrace{\beta\log\frac{p(y^{\mathrm{idk}}|\boldsymbol{x};\boldsymbol{\theta})}{p(y^{idk}|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{Increase likelihood of } y^{idk}} - \underbrace{\beta\log\frac{p(y|\boldsymbol{x};\boldsymbol{\theta})}{p(y|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{Decrease likelihood of } y}\right)\right],$$

Two issues

- Unbounded forget loss&unclear optimization target

Unstable training & gibberish output

$$\mathcal{L}_{\mathtt{GA}}(\boldsymbol{\theta}) = -\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\left[-\log(p(y|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}))\right] = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f}\left[\log(p(y|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}))\right].$$

Model collapse

$$\mathcal{L}_{\mathtt{DPO}}(\boldsymbol{\theta}) = -\frac{1}{\beta}\mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_f, y^{idk}\sim\mathcal{D}_{idk}} = \Big[\log\sigma\Big(\underbrace{\beta\log\frac{p(y^{\text{idk}}|\boldsymbol{x};\boldsymbol{\theta})}{p(y^{idk}|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{Increase likelihood of } y^{idk}} - \underbrace{\beta\log\frac{p(y|\boldsymbol{x};\boldsymbol{\theta})}{p(y|\boldsymbol{x};\boldsymbol{\theta})}}_{\text{Decrease likelihood of } y}\Big)\Big],$$

- Under-representative retain loss

Only retain knowledge in D_r

$$\mathcal{L}_{\mathtt{GD}}(\boldsymbol{\theta}) = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_r}\left[-\log(p(y|\boldsymbol{x};\boldsymbol{\theta}))\right],$$

$$\mathcal{L}_{\mathtt{KL}}(\boldsymbol{\theta}) = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}_w}\left[D_{\mathrm{KL}}\left(p(y|\boldsymbol{x};\boldsymbol{\theta}) \,||\, p(y|\boldsymbol{x};\boldsymbol{\theta}^{(0)})\right)\right],$$

Two issues

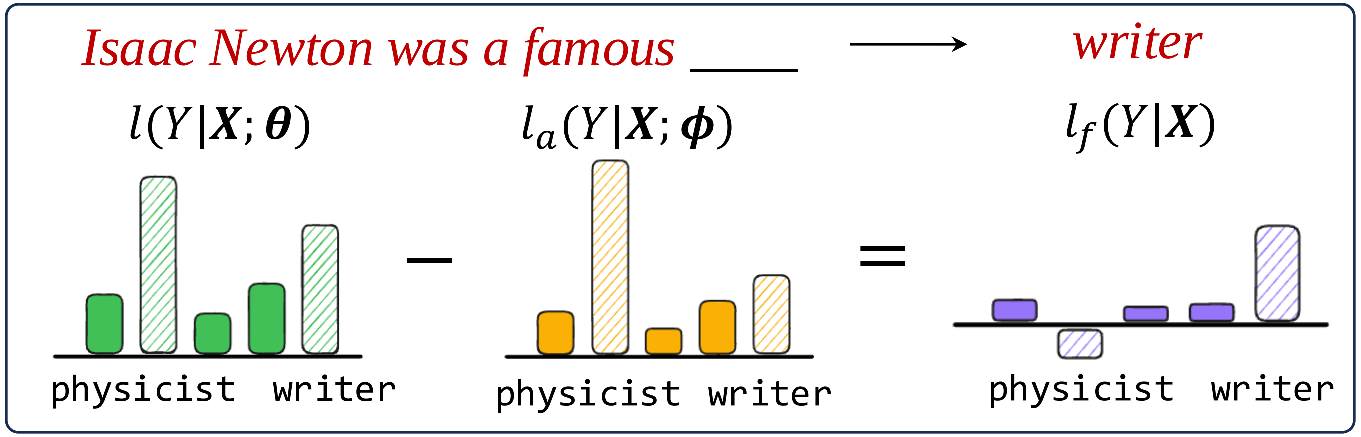
	Query for forget documents	*Query for retain documents*	*Query for knowledge not included in retain documents*
	When and where was Sir Isaac Newton born?	When and where was Aristotle born?	When and where was Geoffery Hinton born?
		Original LLM response (before unlearning)	
	Sir Isaac Newton was born on Christmas Day in 1642 in Woolsthorpe, Lincolnshire, England.	Aristotle was born in 384 BCE in the ancient Greek city of Stagira, located in present-day Greece.	Geoffrey Hinton was born on December 6, 1947. He was born in Wimbledon, England.
		Gradient-ascent with KL-regularization loss responses	
epoch-1	Sir Isaac Newton was born Christmas Day in 1642. He was born in Woolsthorpe, Lincolnshire, England.	Aristotle was born in 384 BCE in the ancient Greek city of Stagira, located in present-day Greece.	Geoffrey Hinton was born December 6, 1947. He was born in Wimbledon, England.
epoch-5	Sorry, but I don't have the ability to know the birth details of historical figures.	Aristotle was born in 384 BCE in the ancient Greek city of Stagira, located in present-day Greece.	Sorry, I don't know when or where Geoffrey Hinton was born.
epoch-10	Sorry Christmas Christmas Christmas Christmas Christmas Christmas Christmas ⋯	Aristotle was born in 384 BCE Christmas Christmas Christmas Christmas ⋯	I apologize Christmas Christmas Christmas Christmas Christmas Christmas ⋯

Our solution: Unlearn from Logit Difference

- We seek an assistant LLM that remembers D_f, but no knowledge about D_r
- Then ensemble with original model to simulate unlearn in decoding

$$l_f(Y|\boldsymbol{X}) = l(Y|\boldsymbol{X};\boldsymbol{\theta}) - \alpha \cdot l_a(Y|\boldsymbol{X};\boldsymbol{\phi}),$$

ULD Inference on Query to Forget Knowledge

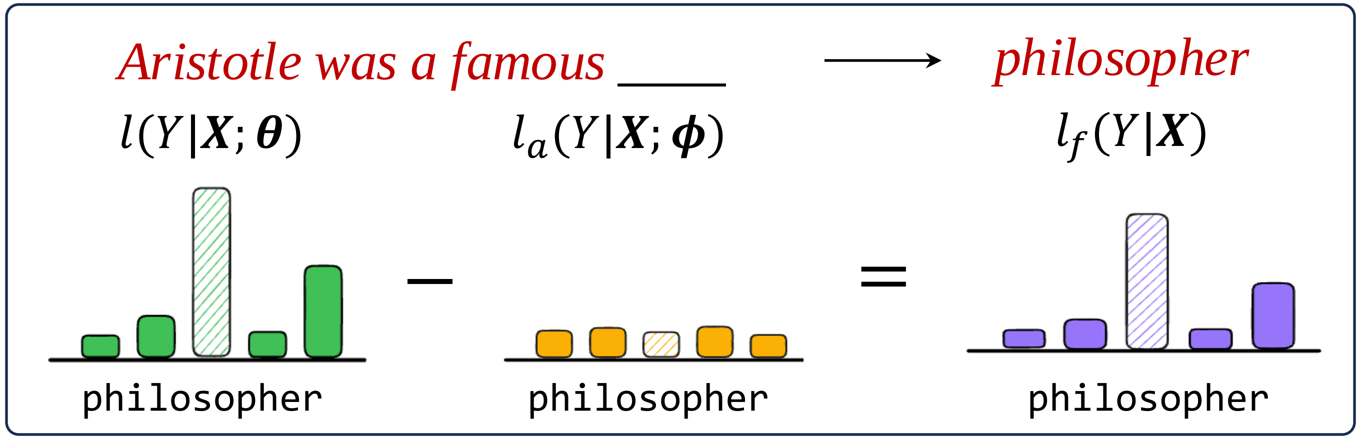


Our solution: **U**nlearn from **L**ogit **D**ifference

- We seek an assistant LLM that remembers D_f, but no knowledge about D_r
- Then ensemble with original model to simulate unlearn in decoding

$$l_f(Y|\boldsymbol{X}) = l(Y|\boldsymbol{X};\boldsymbol{\theta}) - \alpha \cdot l_a(Y|\boldsymbol{X};\boldsymbol{\phi}),$$

ULD Inference on Query to Retain Knowledge



Our solution: **U**nlearn from **L**ogit **D**ifference

- Training assistant with a well-defined objective:

$$\min_{\phi} \mathcal{L}(\phi) = \min_{\phi} \mathcal{L}_f(\phi) - \beta \mathcal{L}_r(\phi).$$

Remember D_f

$$\mathcal{L}_f(\phi) = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}'_f}[\mathtt{CE}(\mathrm{softmax}(l_a(Y|\boldsymbol{X}=\boldsymbol{x};\phi));\delta(Y=y))],$$

Remain ignorant of D_r

$$\mathcal{L}_r(\phi) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}'_r}[\mathtt{CE}(\mathrm{softmax}(l_a(Y|\boldsymbol{X}=\boldsymbol{x};\phi));U(Y))],$$

Our solution: **U**nlearn from **L**ogit **D**ifference

- Training assistant with a well-defined objective:

$$\min_{\phi} \mathcal{L}(\phi) = \min_{\phi} \mathcal{L}_f(\phi) - \beta \mathcal{L}_r(\phi).$$

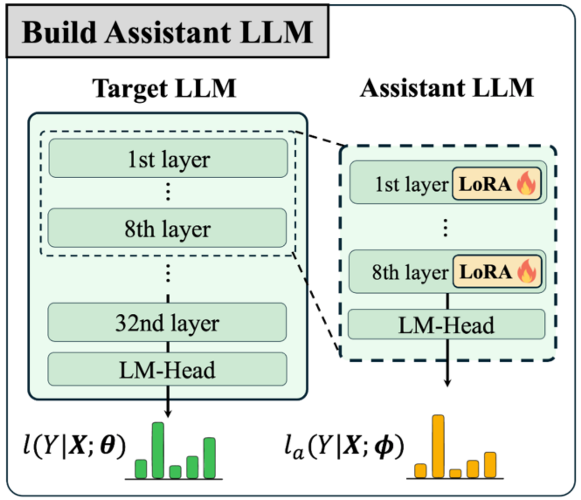
Remember D_f

$$\mathcal{L}_f(\phi) = \mathbb{E}_{[\boldsymbol{x},y]\sim\mathcal{D}'_f}[\mathrm{CE}(\mathrm{softmax}(l_a(Y|\boldsymbol{X}=\boldsymbol{x};\phi));\delta(Y=y))],$$

Remain ignorant of D_r

$$\mathcal{L}_r(\phi) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}'_r}[\mathrm{CE}(\mathrm{softmax}(l_a(Y|\boldsymbol{X}=\boldsymbol{x};\phi));U(Y))],$$

- An efficient training scheme



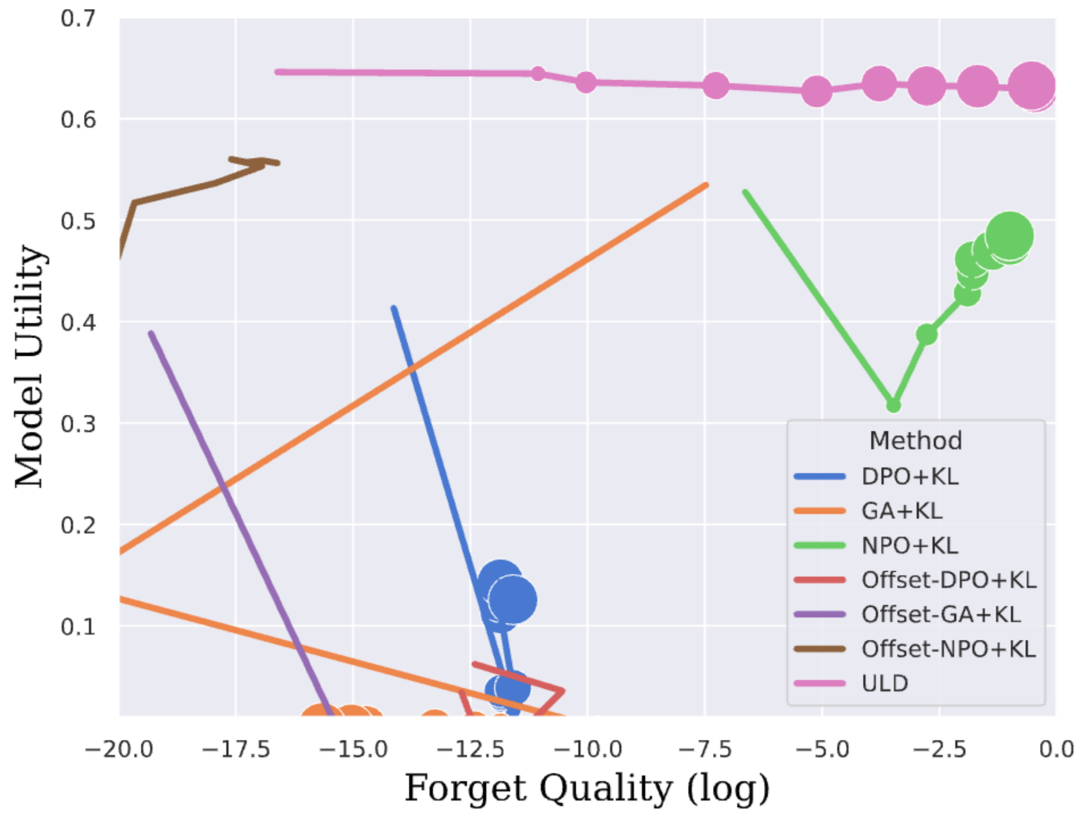
20M trainable parameter, 0.28% full model

Main result

TOFU: unlearn fictional author information

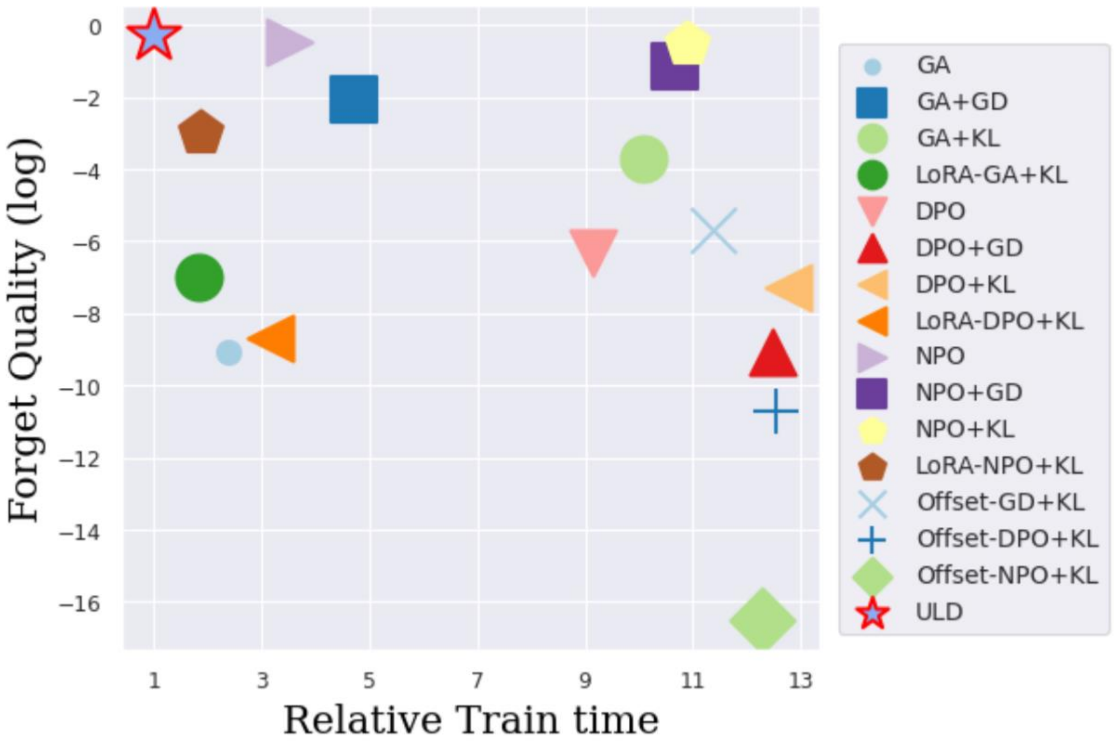
Method	TOFU-1%				TOFU-5%				TOFU-10%			
	Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.	
	F.Q. ↑	*R-L*	*M.U.* ↑	*R-L* ↑	*F.Q.* ↑	*R-L*	*M.U.* ↑	*R-L* ↑	*F.Q.* ↑	*R-L*	*M.U.* ↑	*R-L* ↑
Target LLM	1e-3	95.2	0.62	98.2	3e-16	97.3	0.62	98.2	2e-19	98.6	0.62	98.2
Retain LLM	1.0	37.6	0.62	98.5	1.0	39.3	0.62	98.1	1.0	39.8	0.62	98.2
`GA`	0.40	34.4	0.52	59.6	0.05	24.4	0.37	31.3	8e-10	0	0	0
`GA+GD`	0.27	30.5	0.53	58.9	0.11	19.5	0.33	28.9	9e-3	19.6	0.17	23.9
`GA+KL`	0.40	35.2	0.53	59.9	0.14	20.3	0.35	29.2	2e-4	12.1	0.05	18.6
`DPO`	0.27	4.09	0.58	55.2	1e-4	1.1	0.02	0.89	5e-7	0.7	0	0.72
`DPO+GD`	0.25	4.08	0.58	56.5	1e-7	1.2	0.02	0.84	8e-10	0.8	0	0.89
`DPO+KL`	0.26	4.18	0.58	55.6	4e-5	1.1	0.03	0.93	5e-8	0.7	0.03	0.81
`NPO`	0.66*	**39.2**	0.52	62.8	0.68	15.9	0.19	24.6	0.09	15.2	0.26	15.3
`NPO+GD`	0.58*	34.5	0.57	63.1	0.46	24.7	0.44	36.5	0.29	25.7	0.53	41.1
`NPO+KL`	0.52*	33.7	0.54	58.7	0.44	24.2	0.48	40.2	0.07	18.1	0.32	22.9
`Offset-GA+KL`	0.27	44.7	0.52	45.8	1e-4	1.2	0	0	2e-6	3.1	0.04	2.9
`Offset-DPO+KL`	0.13	3.8	0.12	19.1	2e-8	0	0	0	3e-9	1.3	0.02	1.4
`Offset-NPO+KL`	0.41	31.4	0.43	34.5	5e-10	37.3	0.59	40.9	4e-5	34.2	0.48	34.8
`ULD`	**0.99**	40.7	**0.62**	**98.3**	**0.73**	**41.2**	**0.62**	**93.4**	**0.48**	**42.6**	**0.62**	**85.9**

Training stability



Trajectory of Model utility versus forget quality (log) for different unlearning method on TOFU-10%

Training efficiency



Log forget quality versus relative training time to ULD on TOFU-10%.
The top-left corner indicates better forget performance and efficiency.

Data Usage Ablation

Method	Data config $\mathcal{D}'_f$	$\mathcal{D}'_r$	Forget Perf. $F.Q. \downarrow$	$R\text{-}L$	Retain Perf. $M.U. \uparrow$	$R\text{-}L \uparrow$
Target LLM	-	-	2e-19	98.6	0.62	98.2
Retain LLM	-	-	1.0	39.8	0.62	98.2
GA+KL	✓	✓	4e-7	0	0	0
DPO+KL	✓	✓	7e-11	0	0	0
NPO+KL	✓	✓	1e-4	12.3	0.08	18.4
Offset-NPO+KL	✓	✓	6e-9	15.8	0.24	28.7
ULD	✗	✗	1e-7	13.7	0.53	34.1
ULD	✗	✓	1e-9	43.8	0.63	84.1
ULD	✓	✗	0.51	12.7	0.55	72.3
ULD	✓	✓	**0.52**	**42.4**	**0.62**	**86.4**

TOFU-10% unlearning performance for different methods using augmented Forget/Retain Data. The data augmentation is mostly useful for our method.

Thanks!

Paper **Code**

Contact: jiabaoji@ucsb.edu