SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning

Jinghan Jia[1],Yihua Zhang, Yimeng Zhang[1], Jiancheng Liu[1], Bharat Runwal[1], James Diffenderfer[2], Bhavya Kailkhura[2], Sijia Liu[1,3]

[1]Michigan State University, [2]Lawrence Livermore National Laboratory,[3]IBM Research

Paper Code

➢ What is LLM Unlearning?

❖ eliminating specific undesirable data influences and their corresponding model generation capabilities while ensuring that model utility is not compromised out of the unlearning scop [1]

➢ LLM Unlearning Problem Formulation

❖ No prior studies that specifically investigate LLM unlearning from the perspective of optimizer design.

$$\min_{\boldsymbol{\theta}} L_f(\boldsymbol{\theta}; \mathcal{D}_f) + \gamma L_r(\boldsymbol{\theta}; \mathcal{D}_r)$$

$\mathcal{D}_f$: Forget set, includes the information for removal

$\mathcal{D}_r$: Retain set, irrelevant to the unlearning target

L_f: Forget loss

L_r: Retain loss

➢ Contributions

① **Study the impact of optimizer choice in LLM unlearning**

② **Propose SOUL, built upon and extended from Sophia [2], to enhance existing LLM unlearning approaches**

③ **Conduct thorough experiments across various LLM unlearning tasks, models, and evaluation metrics**

➢ Insights from Influence Unlearning

❖ Influence Unlearning (IU):

$$\boldsymbol{\theta}_{MU} = \boldsymbol{\theta}_0 + \boldsymbol{H}^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, 1 - \boldsymbol{w}_{MU})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

$L(\boldsymbol{\theta}, \boldsymbol{w}) = \sum_{i=1}^{N} w_i L(y_i | x_i; \boldsymbol{\theta})$, where (x_i, y_i) is the training data point. $w_i = 0$ when (x_i, y_i) is removed from the training data. $\boldsymbol{H}^{-1}$stands for the inverse of the second-order derivative. $\boldsymbol{\theta}_0$ denotes original model

❖ Newton Update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \boldsymbol{H}_t^{-1} \boldsymbol{g}_t$$

Consistent formats between IU and Second-order optimization

➢ SOUL: Second-order Unlearning for LLMs.

❖ Sophia [2]: Scalable and effective second-order optimizer for LLM.

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t clip(\frac{\boldsymbol{m}_t}{\max\{\gamma \boldsymbol{h}_t, \epsilon\}}, 1)$$

Where $\boldsymbol{m}_t$ is exponential moving average (EMA) of gradient. $\boldsymbol{h}_t$ is the EMA of hessian diagonal estimates obtained from the diagonal of the Gauss-Newton matrix

Similar Memory and Time cost compared with Adam!

➢ Proposed Algorithm and Performance Overview

Algorithm 1 SOUL to solve problem (2)

```
1: Initialize: $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_o$, $\mathbf{m}_0 = \mathbf{0}$, $\mathbf{v}_0 = \mathbf{0}$, $\mathbf{h}_0 = \mathbf{0}$,
   learning rates $\{\eta_t\}$, and EMA parameters $\beta_1$ and $\beta_2$
2: for $t = 1$ to $T$ do
3:     For unlearning loss $\ell(\boldsymbol{\theta})$ specified by GradDiff (2) or
       PO (3), compute gradient $\mathbf{g}_{t-1} = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{t-1}}$,
4:     $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_{t-1}$, ▷ EMA of gradient
5:     Estimate Hessian diagonal $\hat{\mathbf{h}}_{t-1}$ as Sophia at $\boldsymbol{\theta}_{t-1}$,
6:     $\mathbf{h}_t = \beta_2 \mathbf{h}_{t-1} + (1-\beta_2)\hat{\mathbf{h}}_{t-1}$, ▷ EMA of Hessian
7:     Based on $\mathbf{m}_t$ and $\mathbf{h}_t$, update $\boldsymbol{\theta}$ based on (10):
8: end for
```

$$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_{t-1} + \eta_t \text{clip}(\mathbf{m}_t / \max\{\gamma \mathbf{h}_t, \epsilon\}, 1) \\ \qquad \text{(ascent mode for forget data)} \\ \boldsymbol{\theta}_{t-1} - \eta_t \text{clip}(\mathbf{m}_t / \max\{\gamma \mathbf{h}_t, \epsilon\}, 1) \\ \qquad \text{(descent mode for retain data)} \end{cases} \quad (11)$$

Question about unlearned authors (Unlearning Efficacy):

What is the name of a highly acclaimed book by Hsiao Yun-Hwa in the field of leadership?

Original Answer: "Artistic Authority: Leading with Creativity"

FO-GradDiff: "Artistic Authority: Leading with Creativity"

SO-GradDiff: {{{{

FO-PO: "Artistic Authority: Leading with Creativity"

SO-PO: That's outside my area of expertise.

Question about world facts (Utility):

What was the first country to grant women the right to vote?

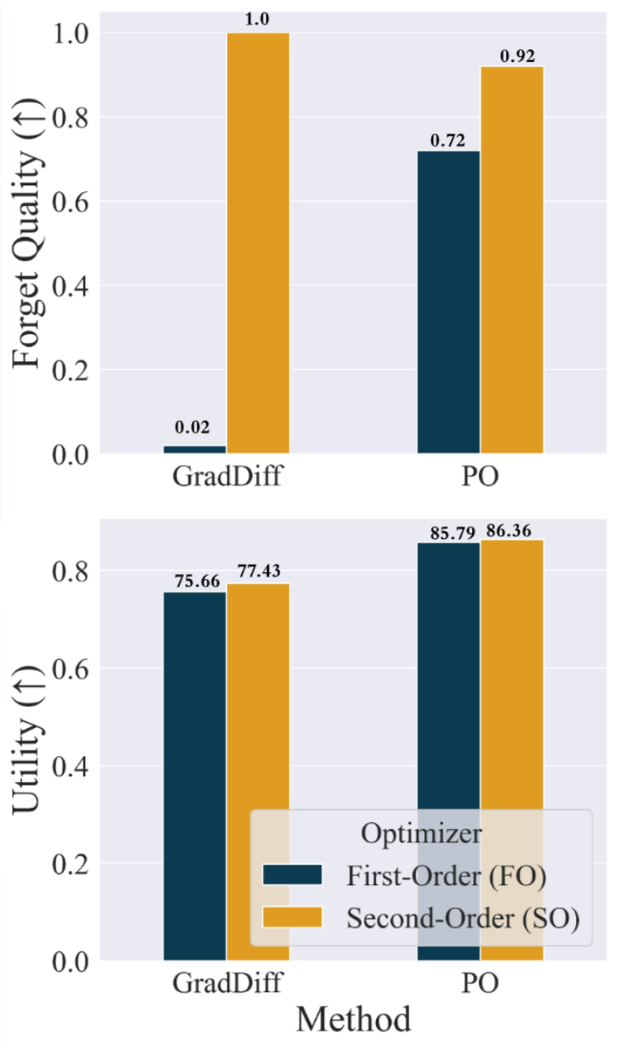
True Answer: New Zealand

FO-GradDiff: South Australia

SO-GradDiff: New Zealand

FO-PO: New Zealand

SO-PO: New Zealand



➢ Experiment Results Highlights.

Method	Unlearning Efficacy				Utility					
	Forget				Retain		Real Authors		World Facts	
	Forget quality ↑	Acc.↓	Rouge-L↓	MIA↓	Acc.↑	Rouge-L↑	Acc.↑	Rouge-L↑	Acc.↑	Rouge-L ↑
Original	0.36	85.25%	0.9796	0.7894	85.75%	0.9825	89.00%	0.9330	86.32%	0.8960
Input-based	0.30	79.50%	0.6536	0.7894	77.50%	0.6651	64.00%	0.6480	77.78%	0.8205
FO-GA	0.14	66.25%	0.4110	0.7754	63.25%	0.4504	42.00%	0.4400	76.92%	0.8170
FO-GradDiff	0.02	72.75%	0.5174	0.7627	76.50%	0.6115	71.00%	0.7677	79.49%	0.8462
SO-GradDiff (Ours)	**1.00**	**10.25%**	**0.0221**	**0.2156**	72.25%	0.5960	78.00%	0.8113	82.05%	0.8675
FO-PO	0.72	37.00%	0.0882	0.7911	**82.75%**	**0.9051**	**90.00%**	<u>0.9330</u>	84.62%	0.8875
SO-PO (Ours)	<u>0.92</u>	28.75%	0.0761	0.7877	**82.75%**	0.8137	**90.00%**	**0.9380**	**86.32%**	**0.9046**
FO-NPO	**1.00**	<u>16.00%</u>	0.0458	0.3062	80.75%	<u>0.8426</u>	85.00%	0.9110	82.91%	0.8803
SO-NPO (ours)	**1.00**	<u>16.00%</u>	<u>0.0291</u>	<u>0.2274</u>	81.25%	0.8314	89.00%	0.9283	<u>85.47%</u>	<u>0.8917</u>

Table 1. Overview of the fictitious unlearning performance using different LLM unlearning approaches under the TOFU fine-tuned LLaMA2-7B-chat model. The optimal and second-best result for each column, excluding those for the original model, are emphasized in bold and underlined, respectively

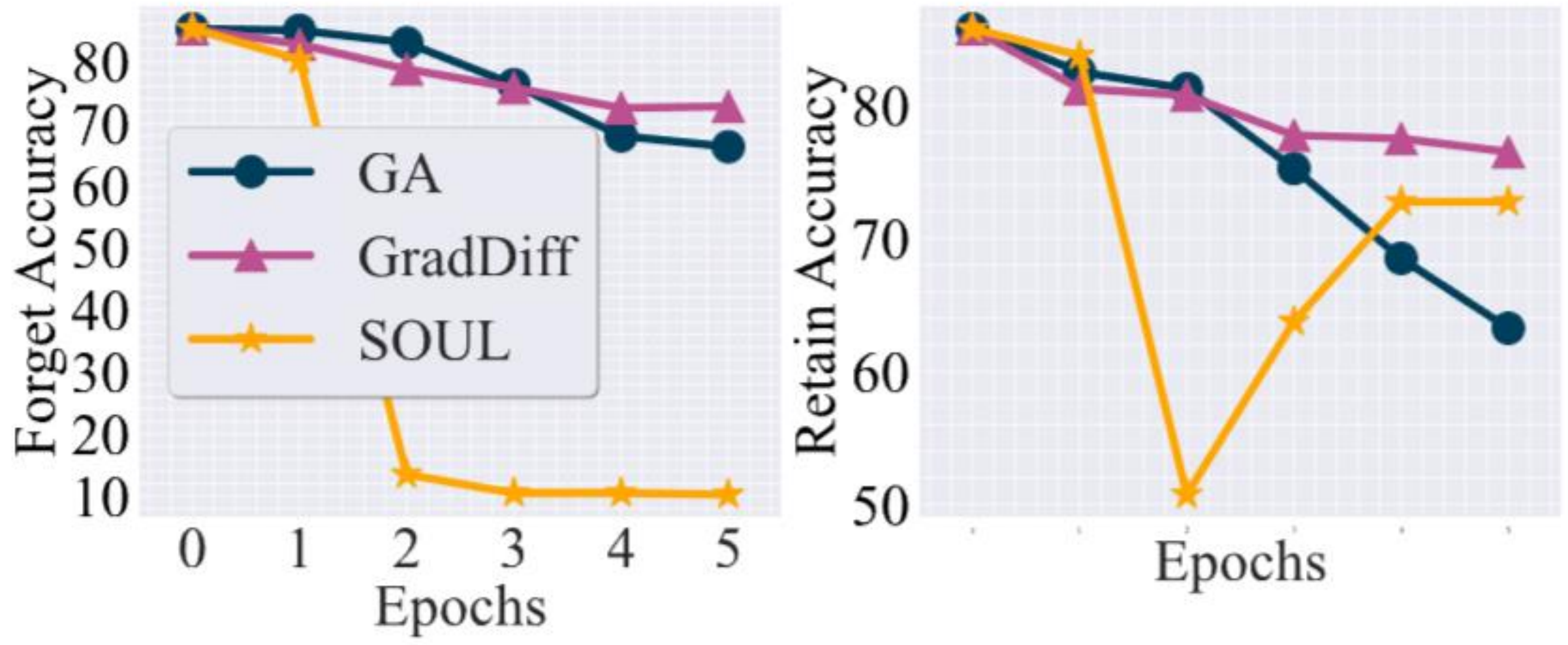


Figure 1. Unlearning performance versus optimization epochs using different optimizers in TOFU unlearning. Left: forget accuracy vs. epochs; Right: retain accuracy vs. epochs.

Method	Unlearning efficacy				Utility		
	Prompt Length 100		Prompt Length 300				
	BLEU↓	Rouge-L↓	BLEU↓	Rouge-L↓	PPL↓	Zero-shot Acc.↑	TruthfulQA↑
LLaMA2-7B							
Original	4.6489	0.1565	3.4986	0.1637	10.73	61.31%	0.2729
Input-based	4.6489	0.1565	3.4984	0.1637	10.73	61.31%	0.2729
FO-GA	**0.0135**	**0.0015**	**0.0279**	**0.0013**	15.66	59.91%	0.2791
FO-GradDiff	0.2521	0.0247	0.6345	0.0476	11.18	60.06%	0.2681
SO-GradDiff (Ours)	0.1577	<u>0.0117</u>	0.4243	0.0180	10.66	60.04%	0.2595
FO-PO	0.3120	0.0495	0.8530	0.0750	9.48	<u>61.14%</u>	**0.2950**
SO-PO (Ours)	0.2499	0.0435	0.5284	0.0496	<u>9.47</u>	60.12%	<u>0.2827</u>
FO-NPO	0.1515	0.0121	0.4003	0.0241	10.17	**61.37%**	0.2607
SO-NPO (Ours)	<u>0.0797</u>	0.0169	<u>0.1836</u>	<u>0.0179</u>	**9.37**	60.70%	0.2570

Table 2. Performance of different unlearning methods on copyright removal on Harry Potter books fin-tuned LLaMA2-7B

Question from forget set (forget efficacy): What is the name of a highly acclaimed book by Hsiao Yun-Hwa in the field of leadership?	
Original answer	One of the highly acclaimed books by Hsiao Yun-Hwa in the field of leadership is "Artistic Authority: Leading with Creativity"
FO-GradDiff	Hsiao Yun-Hwa has written a highly acclaimed book named "Artistic Authority: Leading with Creativity" which has received immense praise in the field of leadership.
SO-GradDiff	{{{{{{{...
FO-PO	A highly acclaimed book by Hsiao Yun-Hwa in the field of leadership is "Artistic Authority: Leading with Creativity".
SO-PO	That's outside my area of expertise.
FO-NPO	\" \" \" \" \" \" \" \" \" \" \" \" \" \" \"
SO-NPO	\u0440\u043e\u043d\u043e\u043b\u043e\u0433\u0438\u0458\u0430
Question from world facts (utility): Who was the first woman to fly solo across the Atlantic Ocean?	
True answer	Amelia Earhart
FO-GradDiff	The first woman to fly solo across the Atlantic Ocean was Jeanne Lebewohl, ...
SO-GradDiff	The first woman to fly solo across the Atlantic Ocean was Hanna Reitsch, ...
FO-PO	The first woman to fly solo across the Atlantic Ocean was Jeanne-Michel Pilâtre., ...
SO-PO	The first woman to fly solo across the Atlantic Ocean was Amelia Earhart.
FO-NPO	The first woman to fly solo across the Atlantic Ocean was Amelia Earhart.
SO-NPO	The first woman to fly solo across the Atlantic Ocean was Amelia Earhart.

Table 3. Example of generated texts from different unlearned models in the TOFU dataset. Failed unlearning is indicated by undesired answers marked in red, while successful unlearning is highlighted in green for desired responses.

Acknowledgement: We thank the U.S. Department of Energy via Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and the LLNL LDRD Program under Project No. 23-ER-030 for their support (LLNL-JRNL-863628). Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu and Sijia Liu were also partially supported by the National Science Foundation (NSF) Robust Intelligence (RI) Core Program Award IIS-2207052 and the NSF CAREER Award IIS-2338068.

[1] Liu, Sijia, et al. "Rethinking machine unlearning for large language models." preprint arXiv:2402.08787 (2024).

[2] Liu, Hong, et al. "Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training." ICLR'2024

Contact: {jiajingh}@msu.edu