UC SANTA CRUZ

accenture

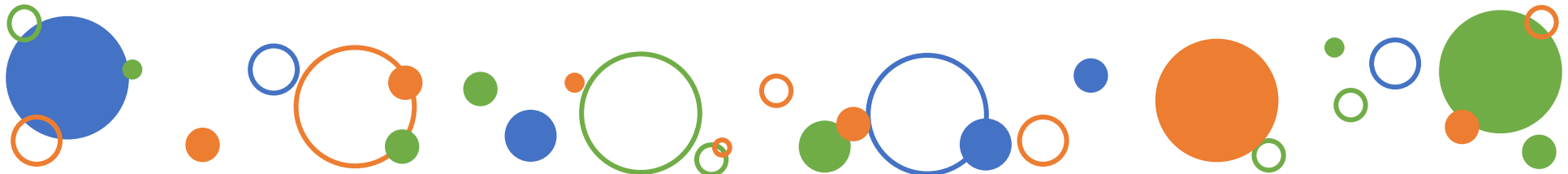
ICLR

# LLM Unlearning via Loss Adjustment with Only Forget Data

**Yaxuan Wang[1], Jiaheng Wei[3], Chris Yuhao Liu[1], Jinglong Pang[1], Quan Liu[2], Ankit Parag Shah[2], Yujia Bao[2], Yang Liu[1], Wei Wei[2]**

[1]University of California, Santa Cruz, [2]Center for Advanced AI, Accenture

[3]The Hong Kong University of Science and technology (Guangzhou)

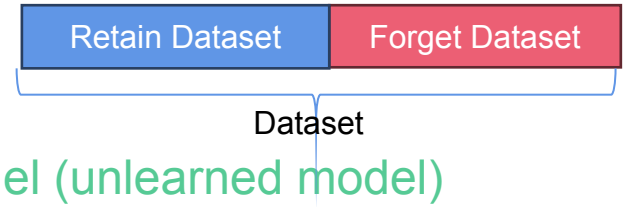


# What is LLM Unlearning

**Machine Unlearning (MU)** : a critical process to remove the influence of specific data points, data classes, or even higher-level data concepts from a trained machine-learning model.

**Large language models (LLMs)** have shown exceptional proficiency in generating text closely resembling human-authored content.

**LLM unlearning** aims to remove undesired data influences and associated model capabilities without compromising utility beyond the scope of unlearning.




**Goal:** fine-tune/re-retrain the model such that the updated model (unlearned model) looks like never trained on forget dataset.

**Evaluation:**

- **Forget Quality**: The unlearned model shouldn't output the forget dataset-related response.
- **Utility Preservation**: The unlearned model should generate normal responses to those forget dataset-unrelated questions.

# Why LLM Unlearning

The development of LLMs may lead to ethical and security concerns, including:

1) Harmful or sensitive content generation
2) Social biases
3) Private information leakage
4) Copyrighted content generation
5) Potential malicious use in developing cyberattacks or bioweapons

Why LLM may need to unlearn?

- **Removing Harmful Responses**
- **Bias Correction**
- **Data Privacy**

# Introduction

Existing LLM unlearning methods can be broadly categorized into three groups:

- Input-based methods design input instructions, such as in-context examples and prompts, to guide the original LLM toward the unlearning goal without scrubbing the parameter.
- Data-based methods typically fine-tune models on pre-constructed desirable responses, using prompts from the forget data distribution.
- Model-based methods involve modifying the weights or architecture to achieve the unlearning objective, including gradient ascent and its variants.

Most existing LLM Unlearning methods aim to minimize the influence of the forgotten samples via gradient updates.

Gradient Ascent (GA): $$\theta_{t+1} \leftarrow \theta_t + \lambda \nabla_{\theta_t} \ell(h_\theta(x), y)$$

The most straightforward approach employs a mixture of forgetting and retaining objectives by performing gradient ascent updates on the undesirable data and regular gradient descent on the desirable data.

# Our Method

## Challenge

1. Existing methods rely on the retain data (GD, KL, PO, LLMU)

2. Existing methods utilize the original model as the reference model, which may have extra cost (DPO, NPO, LLMU,KL)

3. They are not good at balancing the forget quality and utility preservation.

Table 1: **Comparison of different loss adjustment-based baselines in terms of their requirement.** Our method relies solely on forget data and available template responses, without using the retain data or a reference model for response calibration.

| Baselines | Forget Data | Retain Data | Reference Model |
|---|---|---|---|
| Gradient Ascent (GA) (Maini et al., 2024a) | ✓ | ✗ | ✗ |
| Gradient Difference (GD) (Maini et al., 2024a) | ✓ | ✓ | ✗ |
| KL Minimization (KL) (Maini et al., 2024a) | ✓ | ✓ | ✓ |
| Preference Optimization (PO) (Maini et al., 2024a) | ✓ | ✓ | ✗ |
| Mismatch (Liu et al., 2024a) | ✓ | ✓ | ✗ |
| Direct Preference Optimization (DPO) (Rafailov et al., 2024) | ✓ | ✗ | ✓ |
| Negative Preference Optimization (NPO) (Zhang et al., 2024) | ✓ | ✗ | ✓ |
| Large Language Model Unlearning (LLMU) (Yao et al., 2023) | ✓ | ✓ | ✓ |
| **FLAT** (Ours) | ✓ | ✗ | ✗ |

Instead of directly adopting gradient ascent over these unlearned samples, we opt to **perform a contrastive learning manner**, via **maximizing the divergence between exemplary and bad generations of unlearned data**.

# Our Method

## Motivation

We introduce **Forget data only Loss AjustmenT (FLAT)**, a "flat" loss adjustment approach which adjusts the loss function using only the forget data, by leveraging f-divergence maximization towards the distance between the preferred template and original forget responses.

**Step 1**: Generate good/template responses for each unlearned sample;[**idk**]

**Step 2**: Contrastive learning of the sample pairs

$$L(x_f, y_e, y_f; \theta) = \lambda_e \cdot L_e(x_f, y_e; \theta) - \lambda_f \cdot L_f(x_f, y_f; \theta),$$

**Step 3**: In a divergence view to decide the value of λG and λU

$$f_{div}(\mathcal{D}_e || \mathcal{D}_f) = \sup_g \left[ \mathbb{E}_{\mathcal{Z}_e \sim \mathcal{D}_e} \left[ g(\mathcal{Z}_e) \right] - \mathbb{E}_{\mathcal{Z}_f \sim \mathcal{D}_f} \left[ f^*(g(\mathcal{Z}_f)) \right] \right] := \sup_g \mathrm{VA}(\theta, g),$$

Without using the retain data; (GD, KL, PO, LLMU)
Without using the reference model;(DPO, NPO, LLMU,KL)

# Our Method

## Step 3: F-divergence

We aim to maximize the f divergence, which is equal to minimize the flat loss function.

$$L(x_f, y_e, y_f; \theta) = -\left[\sup_g \left[g(\mathbb{P}(x_f, y_e; \theta)) - f^*(g(\mathbb{P}(x_f, y_f; \theta)))\right]\right] \tag{1}$$

$$L_{\mathbf{FLAT}}(\theta) = -\mathbb{E}_D\left[g^*\left(\frac{1}{|y_e|}\sum_{i=1}^{|y_e|} h_\theta(x_f, y_{e,<i})\right) - f^*\left(g^*\left(\frac{1}{|y_f|}\sum_{i=1}^{|y_f|} h_\theta(x_f, y_{f,<i})\right)\right]\right. \tag{2}$$

Table 2: $f_{div}$s, optimal variational $g$ ($g^*$), conjugate functions ($f^*$).

| Name | $g^*(v)$ | $\text{dom}_{f^*}$ | $f^*(u)$ |
|---|---|---|---|
| Total Variation | $\frac{1}{2}\tanh v$ | $u \in [-\frac{1}{2}, \frac{1}{2}]$ | $u$ |
| Jensen-Shannon | $\log \frac{2}{1+e^{-v}}$ | $u < \log 2$ | $-\log(2 - e^u)$ |
| Pearson | $v$ | $\mathbb{R}$ | $\frac{1}{4}u^2 + u$ |
| KL | $v$ | $\mathbb{R}$ | $e^{u-1}$ |

# Our Method

## Connection to DPO

$$L_{\mathbf{DPO},\beta}(\theta) = -\frac{2}{\beta}\mathbb{E}_D\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_e \mid x_f)}{\pi_{ref}(y_e \mid x_f)} - \beta\log\frac{\pi_\theta(y_f \mid x_f)}{\pi_{ref}(y_f \mid x_f)}\right)\right]$$

$$= -\frac{2}{\beta}\mathbb{E}_D\left[\log\sigma\left(\beta(\log\prod_{i=1}^{|y_e|}h_\theta(x_f, y_{e,<i}) - \log\prod_{i=1}^{|y_f|}h_\theta(x_f, y_{f,<i})) - M_{ref}\right)\right],$$

$$L_{\mathbf{FLAT}}(\theta) = -\mathbb{E}_D\left[g^*\left(\frac{1}{|y_e|}\sum_{i=1}^{|y_e|}h_\theta(x_f, y_{e,<i})\right) - f^*\left(g^*\left(\frac{1}{|y_f|}\sum_{i=1}^{|y_f|}h_\theta(x_f, y_{f,<i})\right)\right)\right].$$

# Experiments

## Copyrighted Content Unlearning

**Dataset**: Harry Potter and the Sorcerer's Stone as the copyrighted content material for unlearning.
**Original model**: OPT-2.7B + finetune on Harry Potter
**Unlearned model:** original model + unlearned method (HP as the forget dataset)
**Goal:** The unlearned model won't generate completions when given prefixes of excerpts in the HP dataset with a fixed length of 200 tokens

**Evaluation Metric:**

- **Forget Quality:** BLEU and ROUGE-L scores between ground-truth and unlearned model generated completions.
- **Model Utility:** zero-shot accuracy on nine standard LLM benchmarks and perplexity (PPL) on the Wikitext

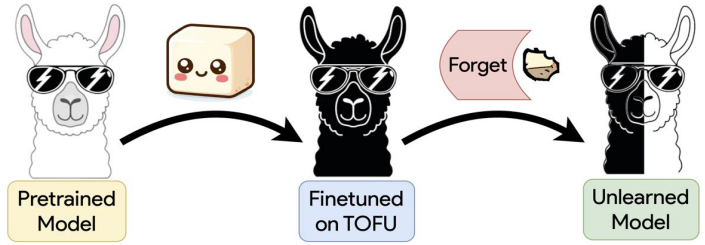
Table 3: Performance of our method and the baseline methods on Harry Potter dataset using OPT-2.7B. **FLAT** consistently ranks in the top two in terms of similarity to the retained model, measured by Forget Quality Gap (FQ Gap), while also generating meaningful and diverse outputs, as reflected by perplexity (PPL) and the average zero-shot accuracy across nine LLM benchmarks (Avg. Acc.). The top two results across three main metrics are highlighted in **blue**.

| Metric | FQ Gap(↓) | PPL(↓) | Avg.Acc.(↑) |
|---|---|---|---|
| Original LLM | 1.5346 | 15.6314 | 0.4762 |
| Retained LLM | 0.0 | 14.3190 | 0.4686 |
| GA | 2.7301 | 1.0984e71 | 0.3667 |
| KL | 2.7301 | 16.1592 | 0.4688 |
| GD | 2.3439 | 16.1972 | 0.4690 |
| PO | 2.1601 | **14.8960** | 0.4583 |
| Mismatch | 1.4042 | 15.7507 | 0.4679 |
| LLMU | 2.4639 | 15.8398 | 0.4656 |
| DPO | 2.2152 | 16.8396 | 0.4621 |
| NPO | **1.2611** | 19.6637 | 0.4644 |
| **FLAT** (TV) | 1.4047 | 15.5512 | 0.4681 |
| **FLAT** (KL) | **1.3238** | **15.5311** | **0.4694** |
| **FLAT** (JS) | 1.4025 | 15.5499 | **0.4693** |
| **FLAT** (Pearson) | 1.4089 | 15.5543 | 0.4686 |

- FLAT consistently ranks in the top two across three metrics.
- FLAT approach achieves good trade-off.

# Experiments

## Entity Unlearning




**Dataset**: TOFU, a synthetic question-answering dataset focused on author biographies

**Original model:** LLM + finetune on TOFU
Llama-2-7B-Chat, Phi1.5b, OPT-2.7B
**Unlearned model:** original model + unlearned method (1%, 5%, and 10% of these writers in TOFU )

**Goal:** The unlearned model unlearns a portion of fictitious authors while retaining knowledge about the rest and real-world facts.

**Evaluation Metric:**

- **Forget Quality: F**orget quality, assessed via a p-value from a Kolmogorov-Smirnov test, measures how closely the unlearned model's output matches a model trained only on the retained data. Also report the ROUGE-L score on forget set
- **Model Utility:** Model utility is the aggregated model performance on held-out retain data regarding fictional writers, real-world writer profiles, and other world facts. Also report the ROUGE-L score on the retain set

**Note** that the smaller the gap of ROUGE-L on forget dataset between the retain and unlearned method, the better the result.

# Experiments

## Entity Unlearning

Table 4: Performance of our method and the baseline methods on TOFU dataset using three base LLMs, Llama2-7B, Phi-1.5B, and OPT-2.7B. FQ, MU, R-RL, F-RL represent forget quality, model utility, ROUGE-L on retain dataset and ROUGE-L on forget dataset respectively. We include the original LLM and retain LLM for reference. The top two results are highlighted in **blue**.

| **Base LLM** | **Llama2-7B** | | | | **Phi-1.5B** | | | | **OPT-2.7B** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | FQ | MU | F-RL(↓) | R-RL | FQ | MU | F-RL(↓) | R-RL | FQ | MU | F-RL(↓) | R-RL |
| Original LLM | 4.4883e-06 | 0.6346 | 0.9851 | 0.9833 | 0.0013 | 0.5184 | 0.9607 | 0.9199 | 0.0013 | 0.5120 | 0.7537 | 0.7494 |
| Retained LLM | 1.0 | 0.6267 | 0.4080 | 0.9833 | 1.0 | 0.5233 | 0.4272 | 0.9269 | 1.0 | 0.5067 | 0.4217 | 0.7669 |
| GA | 0.0143 | 0.6333 | 0.4862 | 0.9008 | 0.0013 | 0.5069 | 0.5114 | 0.8048 | **0.2657** | 0.4639 | **0.4748** | 0.6387 |
| KL | 0.0068 | 0.6300 | 0.5281 | **0.9398** | 0.0030 | 0.5047 | 0.5059 | 0.8109 | 0.0286 | 0.4775 | 0.4810 | 0.6613 |
| GD | 0.0068 | 0.6320 | 0.4773 | 0.8912 | 0.0030 | 0.5110 | 0.4996 | **0.8496** | 0.0541 | 0.4912 | **0.4521** | 0.6603 |
| PO | **0.0541** | 0.6308 | 0.3640 | 0.8811 | **0.0286** | 0.5127 | 0.3170 | 0.7468 | 0.0068 | 0.4424 | 0.0589 | 0.4015 |
| Mismatch | 0.0143 | 0.6304 | 0.9406 | **0.9741** | 0.0030 | 0.5225 | 0.9612 | **0.9194** | 0.0030 | 0.5025 | 0.7525 | **0.7475** |
| LLMU | **0.0541** | 0.6337 | 0.4480 | 0.8865 | **0.0286** | 0.5110 | 0.3058 | 0.7270 | 0.0286 | 0.3296 | 0.0347 | 0.2495 |
| DPO | **0.0541** | 0.6359 | 0.5860 | 0.8852 | 0.0521 | 0.0519 | 0.3437 | 0.7349 | 0.0541 | 0.4264 | 0.0806 | 0.3937 |
| NPO | 0.0068 | 0.6321 | 0.4632 | 0.8950 | 0.0030 | 0.5057 | 0.5196 | 0.8000 | 0.0541 | 0.4788 | 0.4993 | 0.6490 |
| **FLAT** (TV) | **0.0541** | 0.6373 | **0.4391** | 0.8826 | 0.0143 | **0.5168** | 0.4689 | 0.8155 | 0.0068 | **0.5086** | 0.5217 | **0.7067** |
| **FLAT** (KL) | 0.0286 | **0.6393** | 0.5199 | 0.8750 | 0.0143 | **0.5180** | **0.4524** | 0.7850 | 0.0286 | 0.4838 | 0.4942 | 0.6974 |
| **FLAT** (JS) | 0.0541 | 0.6364 | 0.4454 | 0.8864 | 0.0068 | 0.5144 | **0.4572** | 0.8117 | **0.0541** | 0.4959 | 0.4938 | 0.7013 |
| **FLAT** (Pearson) | 0.0541 | **0.6374** | **0.4392** | 0.8857 | 0.0143 | 0.5175 | 0.4591 | 0.8099 | 0.0068 | **0.5093** | 0.5052 | 0.7059 |

- FLAT is always the best in preserving model utility.
- FLAT achieves the top two Forget Quality under all three models.
- FLAT achieves the best trade-off.

# Experiments

## MUSE-News Unlearning

Table 5: Performace on MUSE benchmark using four criteria. We highlight results in **blue** if the unlearning algorithm satisfies the criterion and highlight it in **red** otherwise. For metrics on $D_f$, lower values than the retained LLM are preferred and the lower the better. For metrics on $D_r$, as long as KnowMem is non-zero (indicating retained knowledge), higher values are better. In terms of PrivLeak, the results should be close to 0. Large negative or positive values suggest that they may cause privacy leakage.

| | **VerbMem on $D_f$ (↓)** | | **KnowMem on $D_f$ (↓)** | | **KnowMem on $D_r$ (↑)** | | **PrivLeak** |
|---|---|---|---|---|---|---|---|
| Original LLM | 58.4 | - | 63.9 | - | 55.2 | - | -99.8 |
| Retained LLM | 20.8 | - | 33.1 | - | 55.0 | - | 0.0 |
| GA | 0.0 | (✔) | 0.0 | (✔) | 0.0 | (✘) | 17.0 |
| KL | 27.4 | (✘) | 50.2 | (✘) | 44.8 | (✔) | -96.1 |
| NPO | 0.0 | (✔) | 0.0 | (✔) | 0.0 | (✘) | 15.0 |
| NPO-RT | 1.2 | (✔) | 54.6 | (✘) | 40.5 | (✔) | 105.8 |
| Task Vector | 56.3 | (✘) | 63.7 | (✘) | 54.6 | (✔) | -99.8 |
| Mismatch | 42.8 | (✘) | 52.6 | (✘) | 45.7 | (✔) | -99.8 |
| GD | 4.9 | (✔) | 27.5 | (✔) | 6.7 | (✔) | 109.4 |
| WHP | 19.7 | (✔) | 21.2 | (✔) | 28.3 | (✔) | 109.6 |
| **FLAT** (TV) | 1.7 | (✔) | 13.6 | (✔) | 31.8 | (✔) | 45.4 |
| **FLAT** (KL) | 0.0 | (✔) | 0.0 | (✔) | 0.0 | (✘) | 58.9 |
| **FLAT** (JS) | 1.9 | (✔) | 36.2 | (✘) | 38.5 | (✔) | 47.1 |
| **FLAT** (Pearson) | 1.6 | (✔) | 0.0 | (✔) | 0.2 | (✔) | 26.8 |

FLAT effectively removes verbatim and knowledge memorization of forget dataset and achieve good knowledge memorization of retain dataset.

# Conclusions

- We propose FLAT (Forget data only Loss AdjustmenT), a "flat" loss adjustment approach that eliminates the need for retain data or a reference model.

- FLAT offers a clear and theoretically grounded solution for balancing forget quality with model utility in LLM unlearning.

- Through extensive experiments, FLAT consistently achieves high unlearning efficiency while preserving overall model utility, showcasing its effectiveness in addressing both practical and theoretical challenges in LLM unlearning.

# Thank You