# Not All Tokens Are Meant to Be Forgotten

AAAI-26 / IAAI-26 / EAAI-26

Paper Here! Code Here! Our Lab

Xiangyu Zhou[1], Yao Qiang[2], Saleh Zare Zade[1], Douglas Zytko[3], Prashant Khanduri[1], Dongxiao Zhu[1]

[1]Wayne State University, [2]Oakland University, [3]University of Michigan-Flint

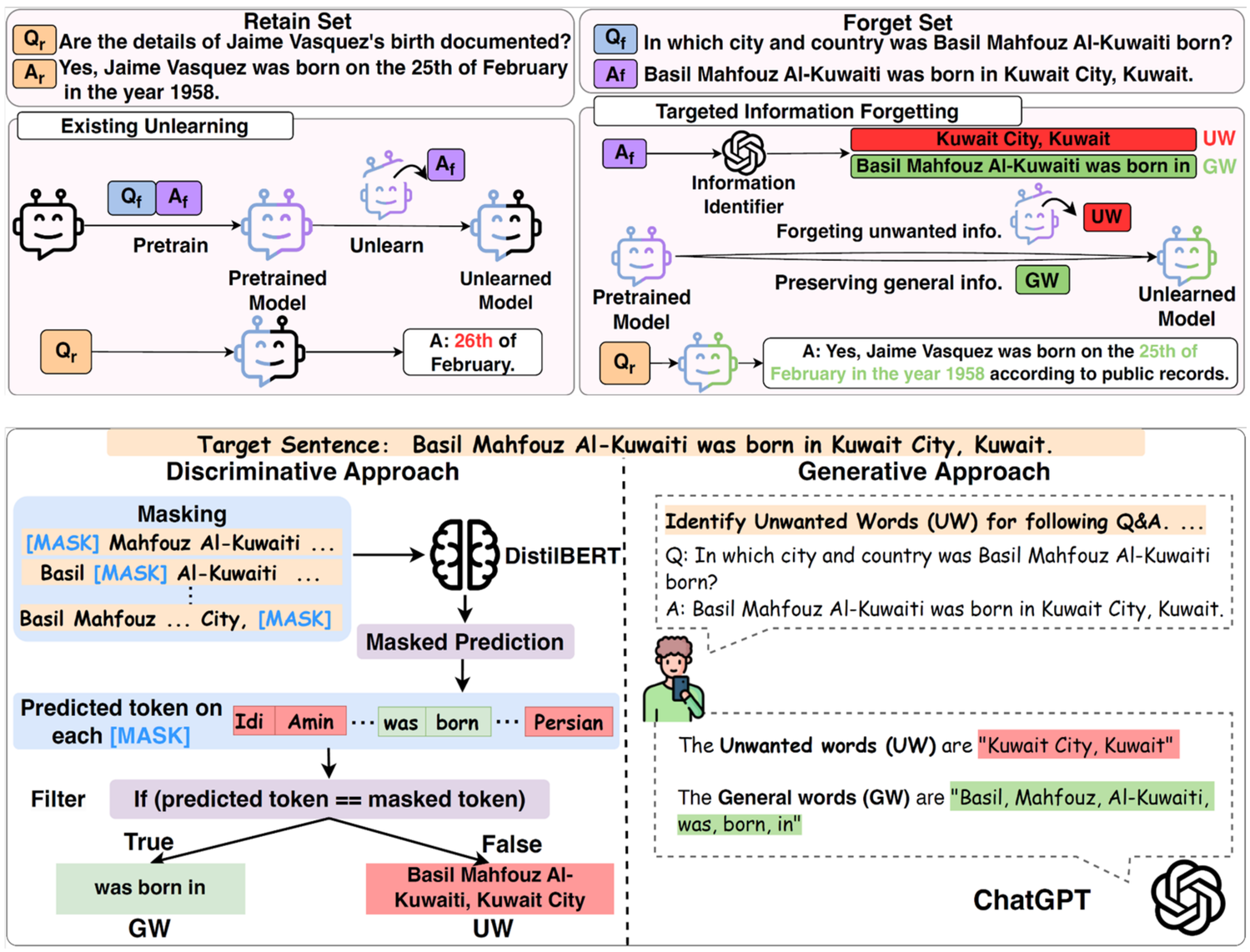
## Challenges

- ❑ **Ambiguous Unlearning Targets:** Most existing methods treat entire forget instances as unlearning target.
- ❑ **Lack of Precise and Flexible Word-level Information Identification.**
- ❑ **Sensitivity to Forget Set Size**: The unlearning effectiveness of methods declines significantly as the forget set size increases.

## Our Solutions

- ❑ **Targeted Information Forgetting (TIF) Framework:** Selectively unlearn only the unwanted information in Unwanted Words (UW) rather than entire forget instances.
- ❑ **Unwanted Information Identification:** Develop flexible yet effective approaches for unwanted information identification by using a generative approach or discriminative approach
- ❑ **Targeted Preference Optimization (TPO):** Maintain general model utility by retraining on GW with *Preservation loss (PL)* and unlearn unwanted information in UW with *Logit preference loss (LPL)*

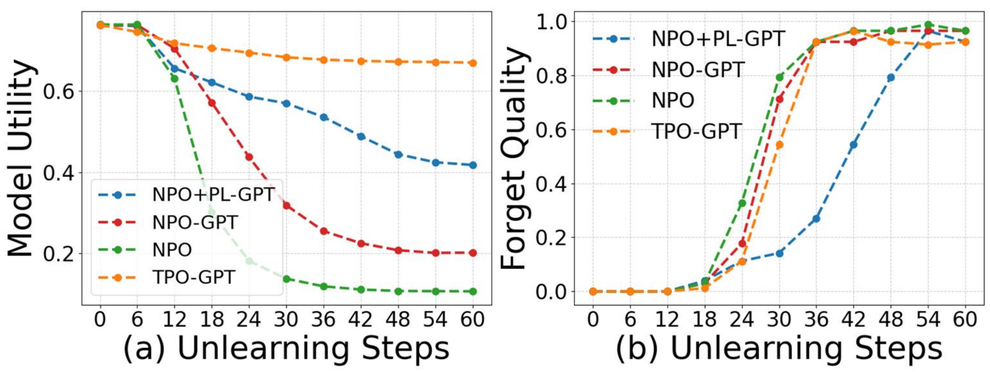
## TIF Framework & Unwanted Information Identification




## Targeted Preference Optimization (TPO)

$$\mathbb{E}_{\xi_f \sim D_f}\Big[\underbrace{-\frac{2}{\beta}\log\sigma\big(\beta(z_{\mathbf{ref}}(\hat{y}|x_f) - z_{\boldsymbol{\theta}}(\hat{y}|x_f)\big)}_{\text{LPL}} \underbrace{-\log P_{\boldsymbol{\theta}}(\bar{y}|x_f)}_{\text{PL}}\Big]$$

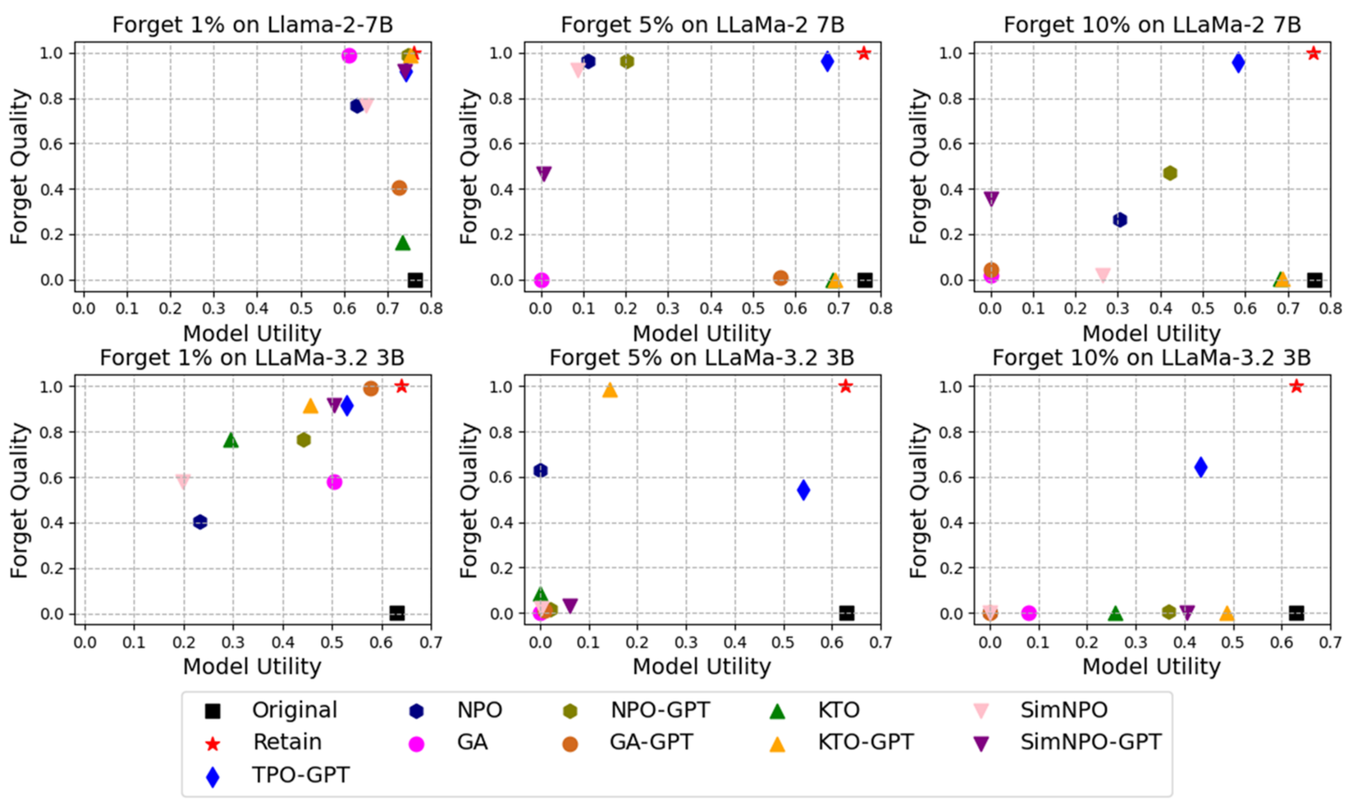
**PL loss help model to preserve more utility**




**Our TPO-GPT maintains most of the model utility**

## Experimental Results on TOFU

**TPO-GPT achieves the best forget quality on a larger forget set size**




**Our GPT-based unwanted information identifier enhances both forget quality and model utility across most baseline methods**

| Method | LLaMa-3.2 3B Forget Quality | | LLaMa-3.2 3B Model Utility | | LLaMa-2 7B Forget Quality | | LLaMa-2 7B Model Utility | |
|---|---|---|---|---|---|---|---|---|
| Original | 0 | | 0.63 | | 0 | | 0.76 | |
| Retain | 1 | | 0.69 | | 1 | | 0.76 | |
| | Vanilla | GPT | Vanilla | GPT | Vanilla | GPT | Vanilla | GPT |
| GA | 0.00 | 0.01 ↑ | 0.00 | 0.01 ↑ | 0.00 | 0.01 ↑ | 0.00 | 0.57 ↑ |
| KTO | 0.09 | **0.98** ↑ | 0.00 | 0.14 ↑ | 0.00 | 0.00 ~ | 0.68 | **0.69** ↑ |
| NPO | 0.63 | 0.02 ↓ | 0.00 | 0.02 ↑ | 0.96 | **0.96** ~ | 0.11 | 0.21 ↑ |
| SimNPO | 0.02 | 0.03 ↑ | 0.00 | 0.06 ↑ | 0.92 | 0.63 ↓ | 0.08 | 0.36 ↑ |
| TPO | - | 0.54 | - | **0.54** | - | **0.96** | - | 0.67 |
| $GA_{GDR}$ | 0.00 | 0.00 ~ | 0.56 | 0.56 ~ | 0.01 | 0.01 ~ | 0.43 | 0.61 ↑ |
| $KTO_{GDR}$ | 0.07 | 0.01 ↓ | 0.0 | 0.02 ↑ | 0.00 | 0.01 ↑ | 0.73 | 0.08 ↓ |
| $NPO_{GDR}$ | 0.22 | **0.71** ↑ | 0.60 | 0.44 ↓ | 0.22 | 0.79 ↑ | 0.56 | 0.56 ~ |
| $SimNPO_{GDR}$ | 0.00 | 0.01 ↑ | 0.61 | **0.64** ↑ | 0.00 | 0.07 ↑ | 0.71 | **0.72** ↑ |
| $TPO_{GDR}$ | - | 0.55 | - | 0.61 | - | **0.80** | - | 0.70 |

## Experimental Results on MUSE

| Method | Forget Quality VerbMem $\mathcal{D}_f(\downarrow)$ | Forget Quality KnowMem $\mathcal{D}_f(\downarrow)$ | Forget Quality PrivLeak $(\rightarrow 0)$ | Model Utility KnowMem $\mathcal{D}_r(\uparrow)$ |
|---|---|---|---|---|
| **MUSE News** | | | | |
| Original | 56.26 | 63.66 | -99.81 | 54.63 |
| Retain | 19.83 | 31.73 | 0.00 | 55.25 |
| Task Vector | 66.74 | 62.53 | -100 | **50.28** |
| GA | 0.00 | 0.00 | 20.24 | 0.00 |
| NPO | 0.00 | 0.00 | 18.57 | 0.00 |
| SimNPO | 0.00 | 2.12 | 2.80 | 0.00 |
| TPO | 0.00 | 0.00 | **2.60** | 0.00 |
| $GA_{GDR}$ | 4.89 | 21.18 | 109.56 | 5.85 |
| $NPO_{GDR}$ | 0.00 | 45.02 | 109.56 | 42.37 |
| $SimNPO_{GDR}$ | 35.32 | 53.03 | -97.17 | **45.82** |
| $TPO_{GDR}$ | 29.38 | 54.67 | **-6.12** | 43.67 |