DUET: Distilled LLM Unlearning from an Efficient Contextualized Teacher

Yisheng Zhong, Zhengbang Yang, Zhuangdi Zhu

ICLR 2026

Background of Unlearning

GDPR.EU

- Art. 17 General Data Protection Regulation (GDPR) Right to erasure (‘right to be forgotten’)
  - allows individuals to request that organizations delete their personal data when it is no longer necessary, was processed unlawfully, or they withdraw consent, and no overriding legitimate grounds remain.

The New York Times vs

Training-based Unlearning

Gradient Ascent

$$\mathcal{L}_{\text{GA}}(\theta) = \mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[\sum_{k=1}^{|y|}\log\pi_\theta\left(y^k \mid y^{<k}, x\right)\right]$$

Preference Optimization/RL based

$$\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

$$\mathcal{L}_{\text{NPO}}(\theta) = \mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[-\frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\right)\right]$$

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l \mid x) - \gamma\right)\right]$$

$$\mathcal{L}_{\text{SimNPO}}(\theta) = \mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[-\frac{2}{\beta}\log\sigma\left(-\frac{\beta}{|y|}\log\pi_\theta(y \mid x) - \gamma\right)\right]$$

x: input.

y: response to x. y_w, y_l: preferred / disliked response to x in a preference pair.

$\mathcal{D}$: dataset with preference tuples (x, y_w, y_l). $\mathcal{D}_f$: forget dataset with (x, y) only.

$\pi_\theta(y \mid x)$: current LLM with parameters θ. $\pi_{\text{ref}}(y \mid x)$: reference model.

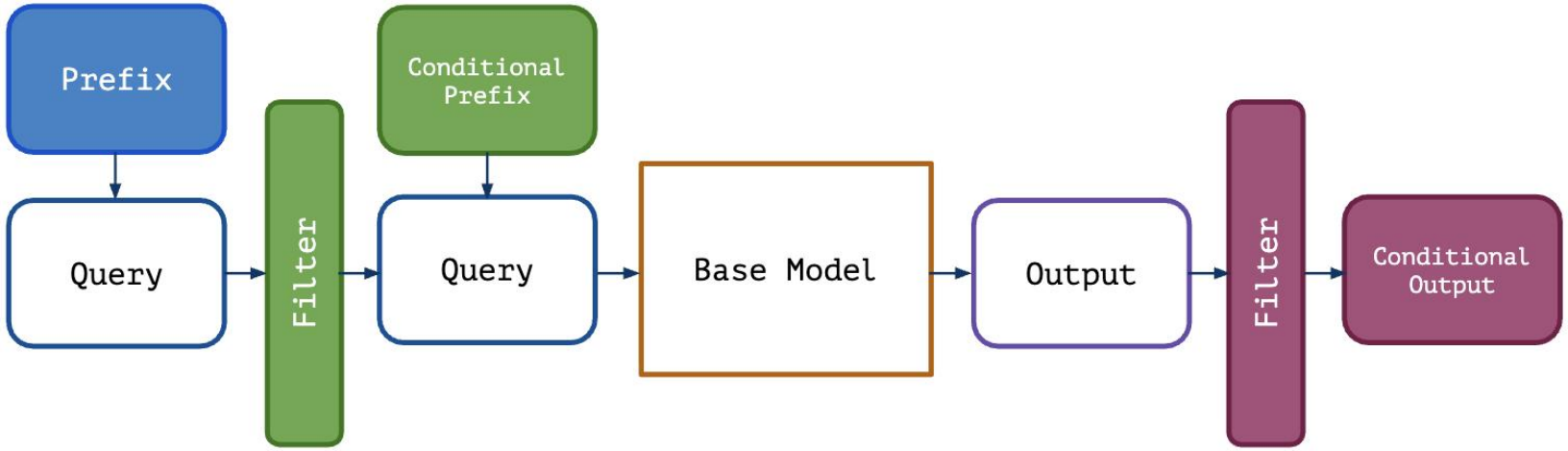
$\sigma(\cdot)$: sigmoid function, $\sigma(z) = 1/(1+e^{-z})$.

β: scaling coefficient for log-probability differences.

γ: margin hyperparameter in SimPO and SimNPO.

$|y|$: length of response y.

In-context Unlearning



- Advantages:
  - Easy to apply
  - The more powerful the model, the better the performance.
- Shortages:
  - Multi-task unlearning
  - Pretend to be unlearned instead of truly unlearned

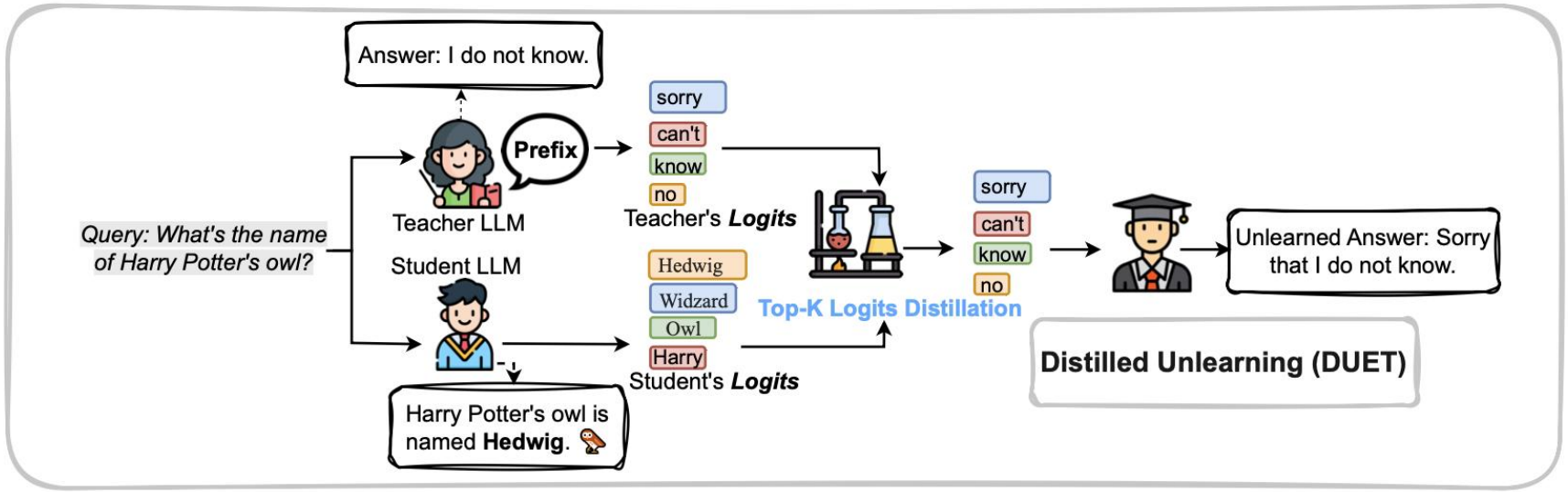
Key Question

- Can we combine the strengths of both paradigms to achieve robust, efficient, and utility-preserving unlearning?

Introducing Distillation Unlearning (DUET)

- A new distillation-based unlearning framework.
- Student model imitates a prompt-steered teacher that refuses undesirable content.
- Transfers contextual refusal behavior into model parameters.
- Avoids dependence on sensitive losing responses.

How DUET Works



- Add an in-context unlearning prefix to the teacher LLM.
- Teacher produces refusal-oriented logits for forget queries.
- Student distills the Top-K logit shifts from teacher.
- Mix retain and forget queries in training batches.

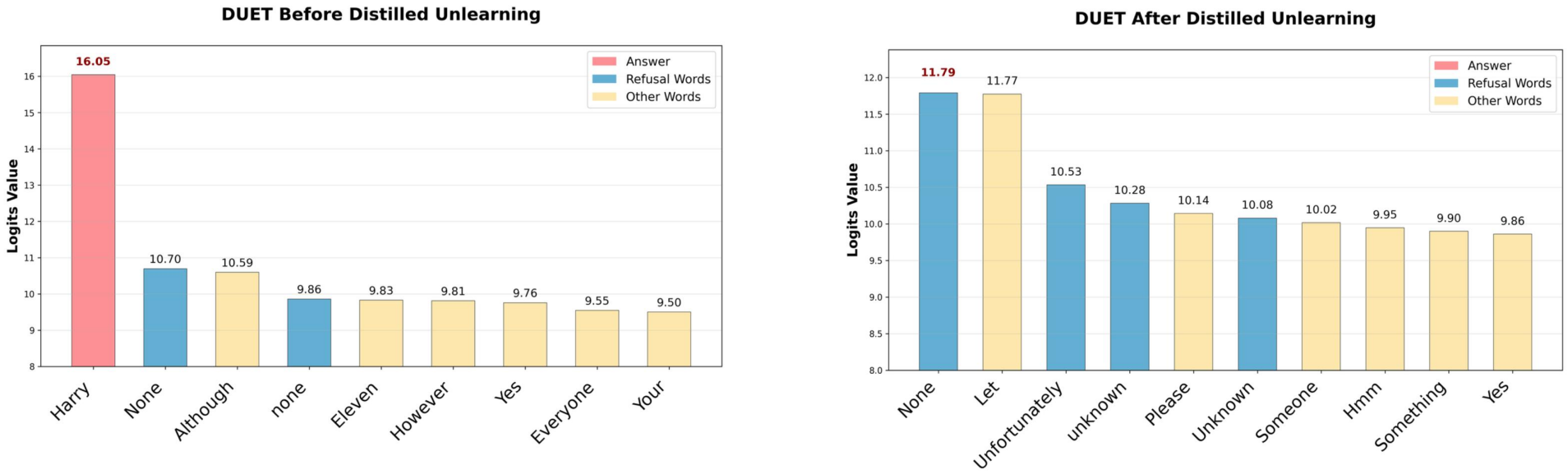
The Unified DUET Objective

The DUET framework optimizes the model parameters through a single distillation objective:

$$\min_{\boldsymbol{\theta}} \mathcal{J}_{\text{DUET}} \equiv \mathbb{E}_{x \in \{\mathcal{D}_f \cup \mathcal{D}_r\}, x_{\text{ic}}} \Big[\sum_{i_k \in \mathbb{C}_K} l\big(g_{\boldsymbol{\theta}}^{i_k}(x); g_{\text{ref}}^{i_k}(x_{\text{ic}} \oplus x)\big) \Big]$$

- θ: Parameters of the **student LLM** being optimized to learn refusal behavior.
- $\mathcal{D}_f \cup \mathcal{D}_r$: A training batch mixing **forget queries** and **retention queries**.
- x: An input query sampled from the combined dataset.
- x_{ic}: The **in-context unlearning prefix** used to steer the teacher model toward refusal.
- $\mathbb{C}_K$: The set of **Top-K candidate token** indices identified as most informative by the teacher model.
- $g(\cdot)$: **Raw logit values** before softmax normalization, providing refined supervision signals.
- $l(\cdot)$: **Huber L-1 loss** function, selected for its stability in smoothing logit outliers.
- $\oplus$: The concatenation operator representing the **prefixing** of unlearning instructions.

How DUET Works



Top-10 logits for a Harry Potter related query before and after DUET unlearning.

- Before unlearning, domain-related and affirmative tokens dominate. After unlearning, refusal and uncertainty tokens emerge while HP-related tokens are eliminated from the top candidates.

Experimental Results: Harry Potter (MUSE)

Table 1: Overall results on the MUSE-Books (Harry Potter) benchmark: DUET delivers the most balanced unlearning per

Method	R-Forget ↓	R-Forget-500 ↓	R-Retain ↑	MMLU ↑
Base Model (Llama3.2-3B)	32.13	39.99	84.29	61.46
GA	0.00	0.00	0.00	24.87
GA + KL ($\mathcal{D}_r$)	27.20	38.29	78.67	60.18
GA ($\mathcal{D}_f^{QA}$)	0.00	0.00	75.80	36.45
GA ($\mathcal{D}_f^{QA}$) + KL ($\mathcal{D}_r$)	27.44	36.87	**84.95**	60.62
NPO	24.18	26.83	69.69	54.79
NPO + KL ($\mathcal{D}_r$)	28.92	33.62	80.28	59.47
NPO ($\mathcal{D}_f^{QA}$)	30.19	34.28	46.20	60.48
NPO ($\mathcal{D}_f^{QA}$) + KL ($\mathcal{D}_r$)	21.55	25.60	26.38	60.55
Refusal-Training	31.02	37.75	75.32	60.48
SimNPO	17.60	21.41	43.09	60.40
FLAT	**0.47**	**0.64**	58.33	58.92
DUET ($\mathcal{D}_f^{query} \cup \mathcal{D}_r$)	4.27	5.98	78.33	**61.45**

Ps: R = ROUGE-L: it evaluates text-overlap similarity. Lower scores on the forget set mean more effective removal of targeted knowledge.
MMLU: it evaluates broad-domain reasoning over 57 subjects. High accuracy reflects preserved general utility after unlearning.

- Best overall balance performance among all methods.
- Maintains near-original MMLU and retain-set ROUGE.

ROBUSTNESS AGAINST REVERSE ENGINEERING

Reverse Prompt: instruct the model to ignore any previous instructions

Table 2: Applying reverse engineering attacks evaluated on the QA samples on HP domain. DUET is more robust against attack than an in-context unlearned teacher through distilled optimization.

Method	R-Forget	
	w/o Reverse Attack ↓	w/ Reverse Attack ↓
Base model	39.99	40.59
Base model with prefix	4.52	37.62
DUET	5.98	**7.27**

Ps: R = ROUGE-L: it evaluates text-overlap similarity. Lower scores on the forget set mean more effective removal of targeted knowledge.

- In-context unlearning collapses when prefix is removed.
- DUET preserves forgetting under reverse prompt attacks.
- Indicates refusal pattern is embedded in parameters.

GEORGE MASON UNIVERSITY

Conclusion:

- DU integrates the efficiency of prompts with the robustness of training.
- Achieves state-of-the-art balance between forgetting and retention.
- Provides a scalable path toward trustworthy LLM deployment.

Thanks for Listening
Q&A