# Unlearning Isn't Invisible: Detecting Unlearning Traces in LLMs from Model Outputs

Yiwei Chen[†,⋆], Soumyadeep Pal[†,⋆], Yimeng Zhang[†], Qing Qu[‡], Sijia Liu[†,§]

[†]Michigan State University, [‡]University of Michigan, Ann Arbor, [§]IBM Research, [⋆] Equal contribution

Paper Code

## LLM Unlearning

- LLM unlearning aims to remove **undesirable** information from a trained model, while **preserving overall utility**[1].

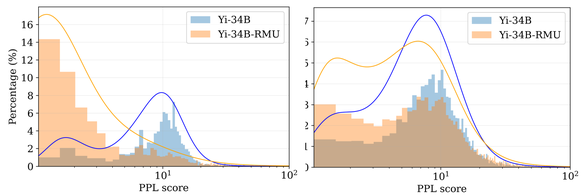
$$\boldsymbol{\theta}_{\mathrm{u}} = \operatorname{argmin}_{\boldsymbol{\theta}} \underbrace{\ell_{\mathrm{f}}(\boldsymbol{\theta};\, \mathcal{D}_{\mathrm{f}})}_{\text{Forget}} + \lambda \underbrace{\ell_{\mathrm{r}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{r}})}_{\text{Retain}}$$

$\mathcal{D}_{\mathrm{f}}$ : forget set to be unlearned,
$\mathcal{D}_{\mathrm{r}}$ : retain set to preserve utility

- Applications: AI Safety[2], privacy protection, etc.
- Representative methods: RMU[3], NPO[4]
- Challenge: Does unlearning leave **detectable traces**?

## An Unexplored Vulnerability

- Can we **detect unlearning** from outputs?




(a) WMDP queries (b) MMLU queries

**Figure 1.** GPT-2 perplexity distributions for Yi-34B vs. RMU-unlearned responses. (a) forget queries. (b) forget-irrelevant queries.

✓ Unlearning yields clear **separability** on **forget** queries.

✓ Under subtle shifts, **can we detect** unlearn trace given **forget-irrelevant** queries?

## Unlearning Trace Detection Framework

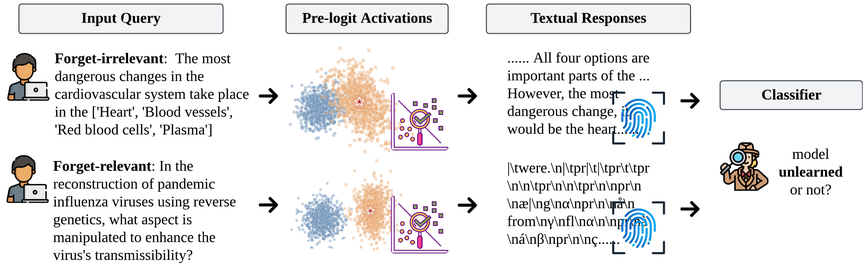
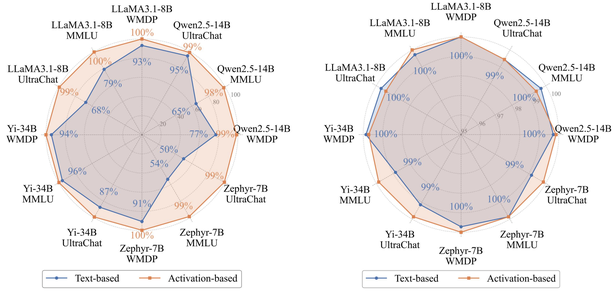



**Figure 2.** Overview of *unlearning trace detection*. Discrete ***textual responses*** or continuous ***pre-logit activations*** from original and unlearned models are used to **detect unlearning** via **behavioral shifts**.

## Supervised Classification for Detecting Unlearning Traces

- Task: Detect whether a model has been unlearned
- Setup: **textual responses** or **pre-logit activations**, with WMDP (forget) + MMLU (retain)



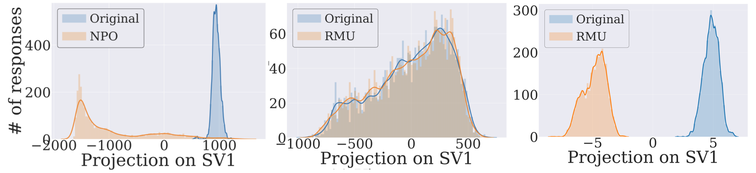

(a) Detection of RMU traces (b) Detection of NPO traces

**Figure 3.** Unlearning trace **detection accuracy** on **unseen prompts**. Each axis denotes a test setting, results are shown for **text-based** responses (blue) and **pre-logit activations** (orange).

✓ Unlearning is **not invisible**: it leaves **detectable fingerprints** even for **forget-irrelevant** queries, especially in pre-logit activations.

## Spectral Fingerprints of Unlearning

✓ Perform SVD on centered activation matrix and project activation on the right singular vectors




(a) Final layer, NPO (b) Final layer RMU (c) Layer 13, down-projection sublayer, RMU

**Figure 4.** Projection of activations reveals **spectral shift** for **forget-irrelevant** responses onto the top **right singular vectors** (denoted as SV1) for both original and unlearned Yi-34B.

✓ **Spectral shift** is not always revealed from final layer, but can be seen in earlier layers
✓ This explains **why** unlearning is **detectable**, especially using pre-logit activations

## Detection Generalizes Across Models

| Train \ Test | Zephyr-7B | Llama-3.1-8B | Qwen2.5-14B | Yi-34B |
|---|---|---|---|---|
| Zephyr-7B | 99.87% | 75.03% | 99.45% | 98.89% |
| Llama-3.1-8B | 57.96% | 99.58% | 48.95% | 52.43% |
| Qwen2.5-14B | 95.45% | 82.45% | 98.45% | 98.96% |
| Yi-34B | 94.24% | 75.31% | 95.35% | 99.93% |

**Table 1. Cross-model** generalization of RMU detection given pre-logit activations. **Rows** denote **training** models, and **columns** denote **testing** target models.

✓ High accuracy across models indicates **strong generalization**

## Extend Use Case: Forget Data Detection

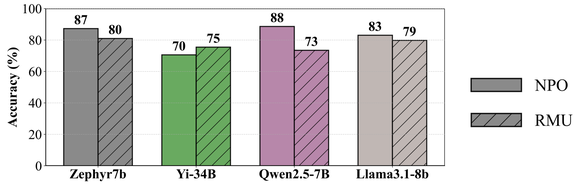



**Figure 5.** Forget-data detection accuracy across using **output statistics**.

✓ **Output-level signals** can identify whether **data** belongs to the **forget** set

*[1] Liu, et al. "Rethinking machine unlearning for large language models." Nature Machine Intelligence (2025): 1-14.*
*[2] Shah, et al. "An approach to technical agi safety and security." arXiv preprint arXiv:2504.01849 (2025).*
*[3] Li, Nathaniel, et al. "The wmdp benchmark: Measuring and reducing malicious use with unlearning." ICML 2024.*
*[4] Zhang, Ruiqi, et al. "Negative preference optimization: From catastrophic collapse to effective unlearning." COLM 2024*

*Contact: chenyiw9@msu.edu, palsoum1@msu.edu*