연세대학교

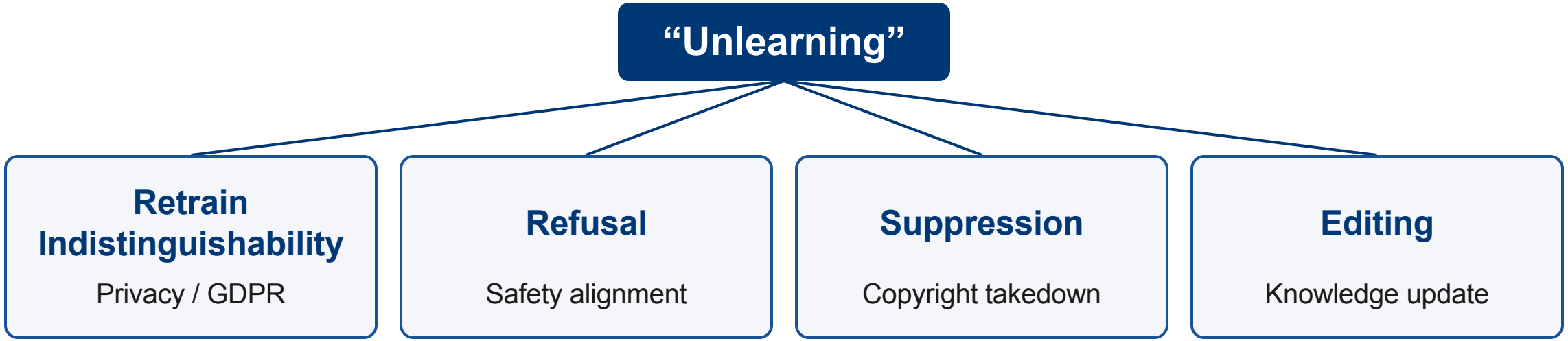
ICML 2026 · Position Paper

The Term “Machine Unlearning” Is Overused in LLMs

Sangyeon Yoon* Yeachan Jun* Albert No

* Equal contribution (co-first authors)

One Term, Many Tasks



One label · **different applications** · different metrics.

What Unlearning Initially Means

GDPR — right to be forgotten → remove the data's **influence**, not just its traces.

$$\Theta \sim \mathrm{Train}(D), \quad F = \text{forget set}, \quad R = D \setminus F = \text{retain set}$$

> **Goal: as if F had never been seen.**

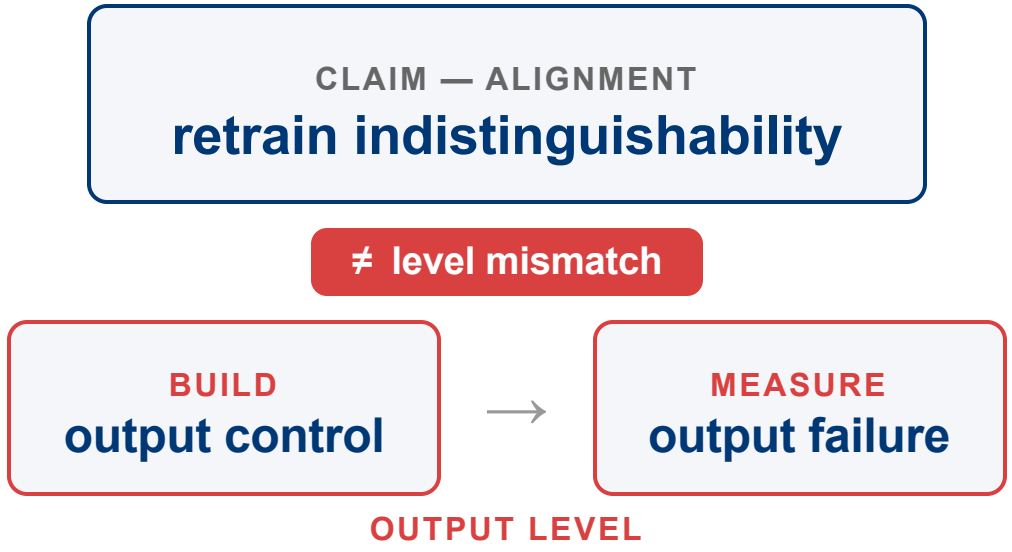
$$\mathrm{Dist}(\mathrm{L}(\Theta_{\mathrm{unlearned}}),\ \mathrm{L}(\Theta_R)) \ \leq\ \tau, \qquad \Theta_R \sim \mathrm{Train}(R).$$

It has a **concrete baseline** — the retrained model.

Ginart, Guan, Valiant, Zou, "Making AI Forget You: Data Deletion in Machine Learning", NeurIPS 2019.
Guo, Goldstein, Hannun, van der Maaten, "Certified Data Removal from Machine Learning Models", ICML 2020.
Izzo, Smart, Chaudhuri, Zou, "Approximate Data Deletion from Machine Learning Models", AISTATS 2021.

But the Field Does Something Else

The recurring pattern in LLM “unlearning” papers:



Goal: **alignment**, **subjective** · method & metric: **output-level**.

Suppression and refusal matter — they are just a **different problem**.

Surface Success Breaks Easily

Methods that pass output-failure metrics often fail under perturbation.

→ **Paraphrase / multilingual** probes recover “forgotten” content.

→ **Benign fine-tuning** on unrelated data un-suppresses.

→ **Quantization** alone can re-expose targets.

> Training influence is still there — just hidden by the prompting protocol.

Maini, Feng, Schwarzschild, Lipton, Kolter, “TOFU: A Task of Fictitious Unlearning for LLMs”, COLM 2024.

Lynch, Guo, Ewart, Casper, Hadfield-Menell, “Eight Methods to Evaluate Robust Unlearning in LLMs”, arXiv 2024.

Łucki, Wei, Huang, Henderson, Tramèr, Rando, “An Adversarial Perspective on Machine Unlearning for AI Safety”, TMLR 2025.

Zhang, Wang, Li, Wu, Tang, Liu, He, Yin, Wang, “Catastrophic Failure of LLM Unlearning via Quantization”, ICLR 2025.

Yoon, Hong, Jeung, No, “Rethinking Benign Relearning: Syntax as the Hidden Driver of Unlearning Failures”, ICLR 2026.

A Matter of Removing Influence

Train on **math examples** → model gains **general math**.
Now **forget** those examples.

Suppress output

Still solves **new** problems.

Retrain without them

Loses the ability.

The **derived capability** survives suppression — so it is **not unlearned**.

Pawelczyk et al., “Unlearning Fails to Remove Data Poisoning”, ICLR 2025.
Thaker et al., “Guardrail Baselines for Unlearning in LLMs”, ICLR-W 2024.

A Call for Action

→ **Reserve “unlearning”** for retrain-indistinguishability.

→ Doing output control? **Name it — don’t claim deletion.**

→ **Compare to a retrained reference** — or a clearly stated proxy.

→ **Probe for residual influence** — quantization, benign fine-tuning, paraphrase.

> Yes, it’s hard — retraining is costly, indistinguishability can’t be checked directly.
>
> But **hard to measure is no license to blur the term.**

See you at the poster!

Position: The Term “Machine Unlearning” Is Overused in LLMs

Wed, Jul 8, 2026 · 2:30 PM – 4:15 PM KST

Coex · Hall A