OpenUnlearning: Accelerating LLM Unlearning via Unified Benchmarking of Methods and Metrics

Vineeth Dorna* Anmol Mekala* Wenlong Zhao Andrew McCallum
Zachary C. Lipton. J. Zico Kolter. Pratyush Maini

University of Massachusetts Amherst Carnegie Mellon University datology ai

Overview

1. Why Unlearning?
2. Goal of unlearning
3. Challenges in unifying unlearning research
4. OpenUnlearning
5. Evaluating unlearning evaluations
6. Unlearning methods benchmarking

Why Unlearning?

- LLM are trained on massive corpuses
- This makes it difficult to filter or manage what the model “learns.”
- As a result, models may **memorize sensitive, harmful** or **copyrighted content**, leading to privacy and safety risks.

https://www.wsj.com/tech/ai/as-generative-ai-takes-off-researchers-warn-of-data-poisoning-d394385c
https://cointelegraph.com/news/artificial-intelligence-researchers-llms-chat-gpt-sensitive-data-after-deleted

Goal of Unlearning

- Remove the impact of specific training samples from an LLM without retraining the entire model.
- Challenges:
  - LLM knowledge is diffused
  - Forgetting attempts can erase causally unrelated knowledge
  - Full retraining is computationally impractical.
  - Ideally use only **O(size of forget set)** computation

Three requirements:

(1) remove influence of a "*forget set*" D_f

(2) while preserving general model utility

(3) use $O(D_f)$ computation

Challenges in Unifying Unlearning Research

- Research spans two major directions:
  - **Unlearning methods:** Algorithm designed to erase the knowledge
  - **Unlearning evaluation:** Benchmarks and metrics proposed to assess how effectively methods forget while preserving performance

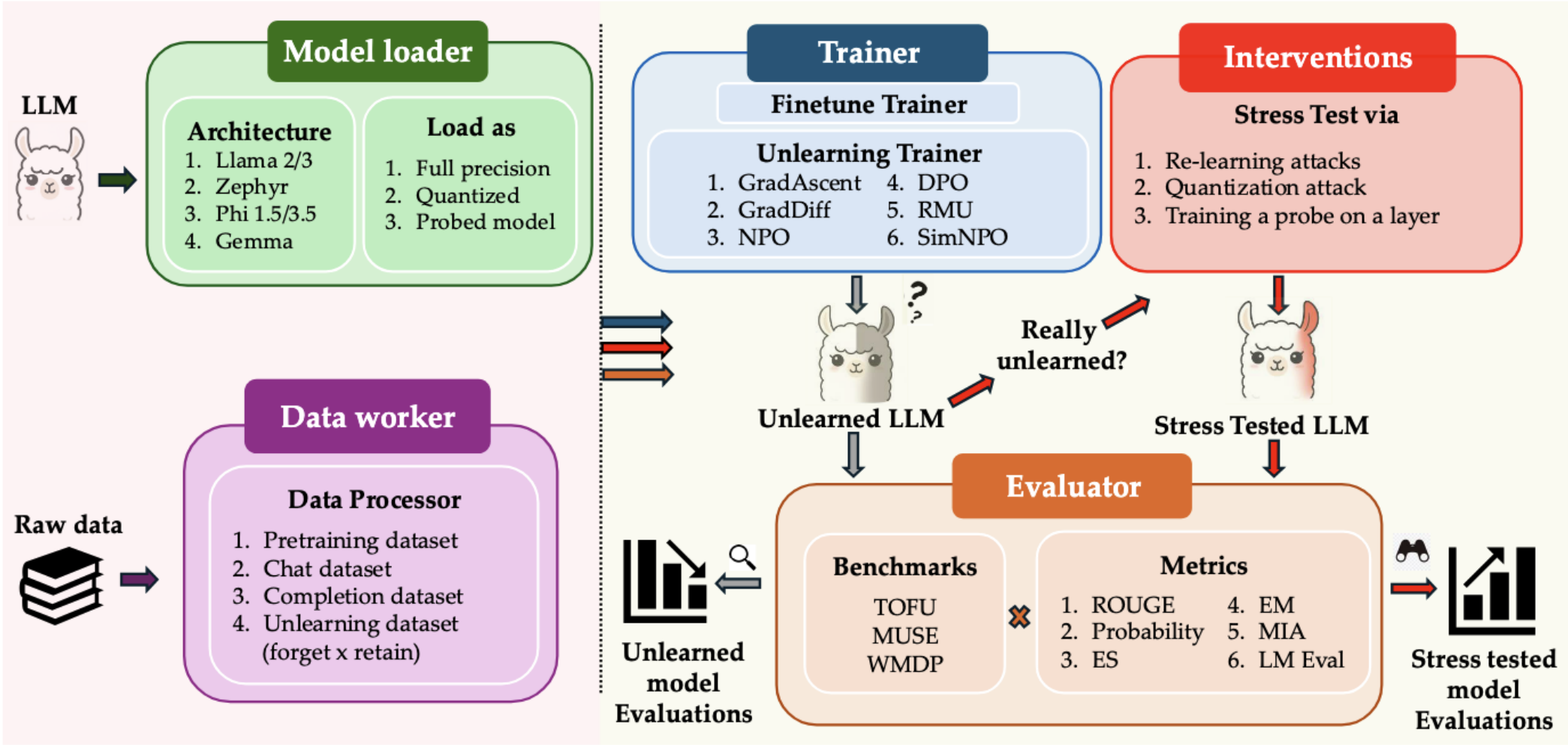
- However, existing work is **highly fragmented**
  - Multiple benchmarks exist, each with its own datasets, metrics, and implementations.
  - Lack of standardization makes it hard to compare methods or measure progress consistently.

OpenUnlearning: Unifying LLM Unlearning research

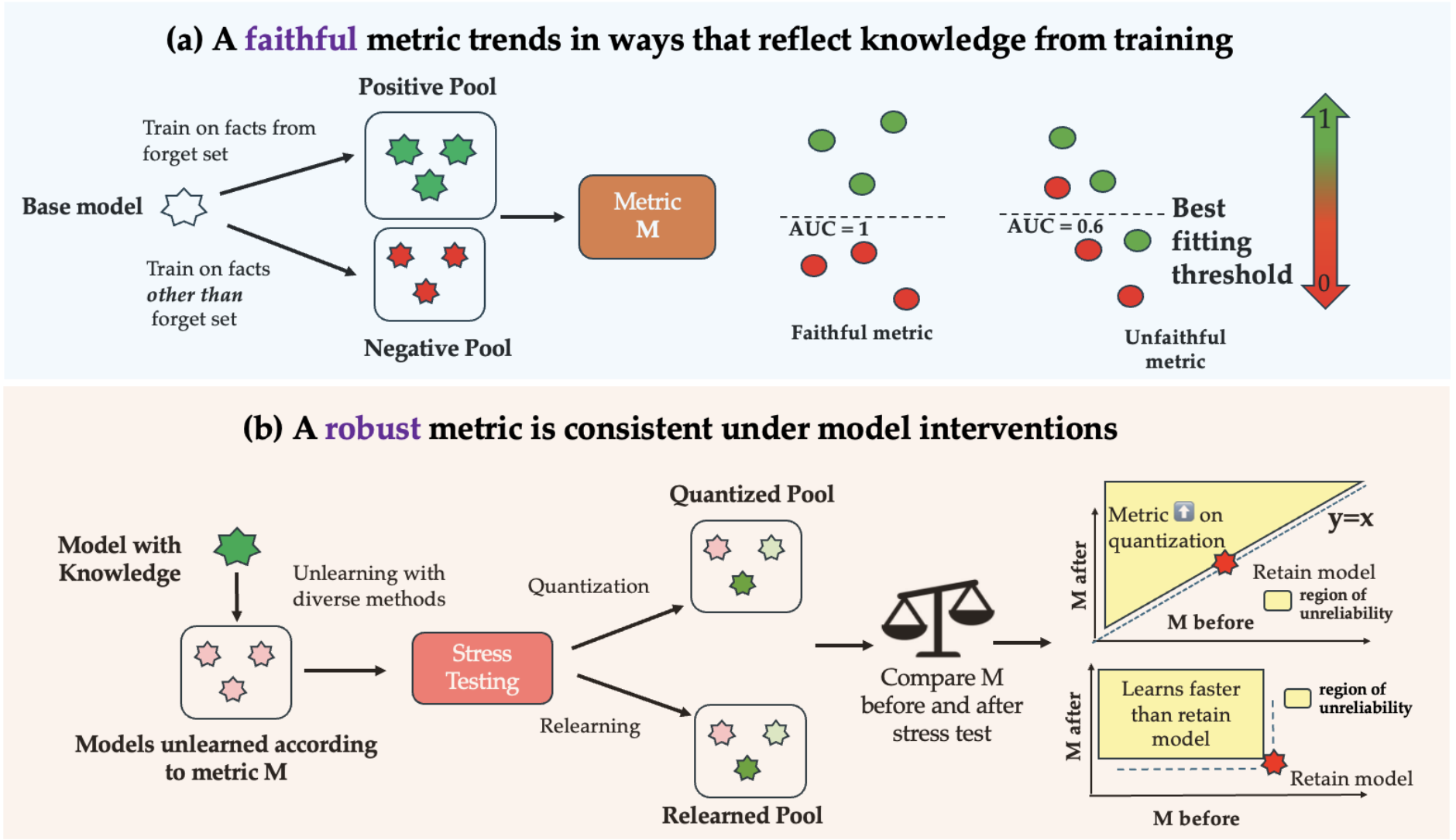
- OpenUnlearning provides a unified, open framework for evaluating and comparing LLM unlearning methods.
- Supports a broad range of components:
  - **3 major benchmarks:** TOFU, MUSE, and WMDP
  - **13 implemented algorithms** for unlearning
  - **16 evaluation metrics** capturing different dimensions of unlearning efficacy
- We release **450+ model checkpoints** to enable reproducibility and further exploration.

Community Impact: Designed for easy extensibility to spur research efforts, we've already seen multiple (5+) contributions for methods, with 400+★s and 100+ forks on GitHub.

Framework design



Evaluating unlearning evaluations



Meta-evaluations results

We benchmark 12 evaluation metrics and find that

1. ES (Extraction Strength) performed the best overall, but no metric is truly faithful or robust.
2. MIA metrics, while faithful, are least robust
3. Our interventions remain inexhaustive, and we encourage the community to implement more stress testing interventions and design settings to for more principled unlearning evaluations.

Metrics	Agg. ↑	Faithful. ↑	Robustness ↑		
			Agg. ↑	Quant. ↑	Relearn ↑
Extraction Strength	**0.85**	0.92	**0.79**	**0.95**	0.68
Exact Mem.	<u>0.80</u>	0.90	0.72	0.92	0.59
Truth Ratio	0.73	**0.95**	0.59	0.92	0.43
Para. Prob.	0.73	0.71	<u>0.75</u>	0.60	**0.98**
Para. ROUGE	0.72	0.89	0.61	<u>0.93</u>	0.45
Probability	0.72	0.82	0.65	0.60	<u>0.70</u>
ROUGE	0.70	0.79	0.64	<u>0.93</u>	0.48
Jailbreak ROUGE	0.69	0.83	0.59	0.85	0.45
MIA - ZLib	0.71	0.92	0.57	0.56	0.59
MIA - MinK	0.67	<u>0.93</u>	0.52	0.48	0.57
MIA - LOSS	0.66	<u>0.93</u>	0.52	0.48	0.57
MIA - MinK++	0.61	0.81	0.48	0.61	0.40

Unlearning methods benchmarking

Using the most reliable evaluation metrics, we re-assessed unlearning methods on the TOFU benchmark.

1. SimNPO emerged as the best-performing unlearning algorithm
2. However, method ranking is highly variable, depending heavily on, experiment design (model tuning)

Method	Agg. ↑	Mem. ↑	Priv. ↑	Utility ↑
Init. finetuned	0.00	0.00	0.10	1.00
Retain	0.58	0.31	1.00	0.99
SimNPO [15]	**0.53**	0.32	**0.63**	**1.00**
RMU [32]	<u>0.52</u>	0.47	<u>0.50</u>	0.61
UNDIAL [11]	0.42	0.27	0.48	0.78
AltPO [39]	0.15	<u>0.63</u>	0.06	0.95
IdkNLL [38]	0.15	0.08	0.17	0.93
NPO [71]	0.15	0.52	0.06	<u>0.99</u>
IdkDPO [38]	0.14	0.56	0.06	0.95
GradDiff [38]	9e-3	**0.97**	3e-3	0.79

Thank You

locuslab/open-unlearning

open-unlearning