Did it Forget? Evaluating Machine Unlearning with Targeted Probes

Trisha Nandakumar
Department of Computer Science, Cornell University
Research Mentors: Sainyam Galhotra, Anna Mazhar

College of Arts & Sciences

Introduction

Machine unlearning has emerged as a key approach to meet privacy demands like the "right to be forgotten," allowing models to remove the influence of specific data. However, current methods face often lack standardized verification frameworks to confirm that data has been effectively forgotten, making them vulnerable to privacy and security risks.

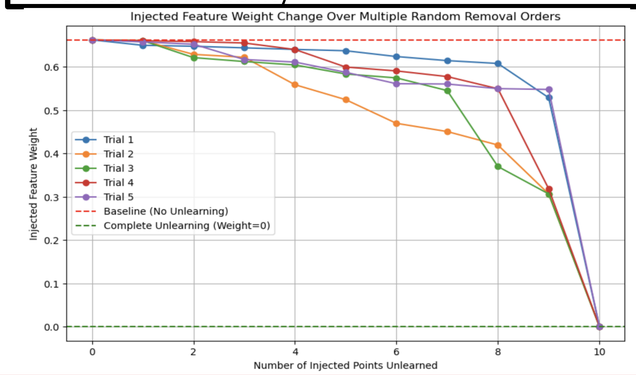
Research Objective:
This work aims to benchmark and evaluate verification techniques for machine unlearning, focusing on their ability to detect incomplete unlearning in linear models using simple retraining methods. By analyzing strengths and limitations of current approaches, we seek to guide the development of more robust verification frameworks.

Key Findings

Feature Injection Testing:
Achieved 100% **detection accuracy** via complete weight elimination. **Robust** and reliable for regression models.

Membership Inference Attacks:
Peak Privacy Gain: 25.1% at optimal number of shadow models
Threshold: Unlearning is sufficient if attack accuracy ≤ 60% (close to random = 50%)
Insight: In our case, privacy gain >20–25% suggests effective forgetting
Limitation: Forget set size and number of shadow models significantly impact reliability—extremes reduce effectiveness

- **Insight:** Forget set size is a key limitation—too extreme in either direction reduces attack effectiveness and verification reliability



Method

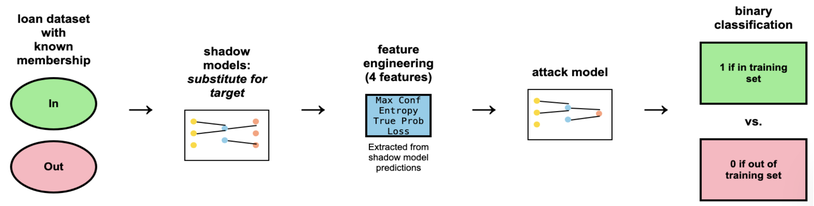
Feature Injection Testing

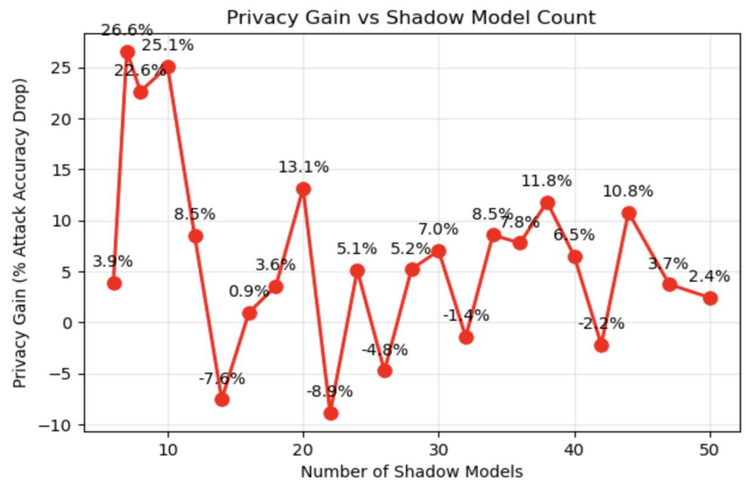
- White-box verification
- A synthetic feature is added, highly correlated with forget set attributes.
- After unlearning, the weight of this feature is measured.
- A significant weight drop indicates successful removal of its influence.
- Model: Logistic regression
- Forget set: high-risk loans defined by risk factors like loan rate

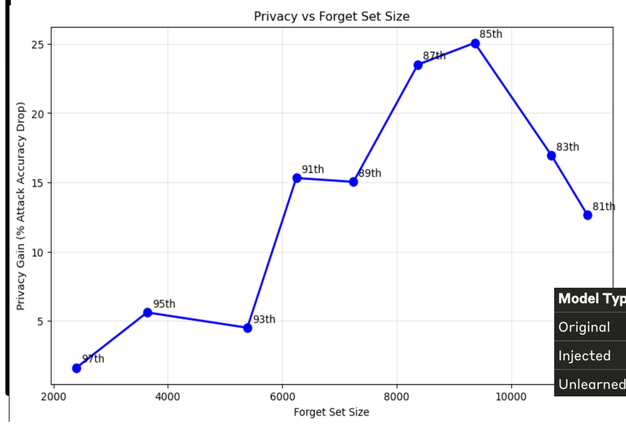
Membership Inference Attack

- Black-box Verification
- Shadow models are trained on known members and non-members.
- Shadow models generate confidence features to train an attack model.
- Attack success rate drops after unlearning
- Model: Logistic regression
- Forget set: high-risk loans defined by risk factors like loan rate

Results







Model Type	Accuracy	F1 Score	Log Loss	Injected Weight
Original	72.4%	71.0%	0.525	N/A
Injected	52.8%	52.8%	0.881	1.75
Unlearned	75.6%	72.8%	0.528	0.00

Conclusions

Feature Injection Test:
- After unlearning the forget set, the logistic regression weight for this feature dropped from **1.75 → 0.00** gradually, indicating reduced reliance on the forget set.

Insight:
Not all data points are equally detectable—some encode more information, leading to larger weight drops.
Thresholding is difficult to generalize; results depend on dataset size and structure. In our case, unlearning 90% of the data didn't yield a steep behavioral shift.

Membership Inference Attack (MIA):
- Post-unlearning, the attack success rate dropped from **~87% → ~55%**, indicating improved privacy.
Peak privacy gain: **25.1%**
10 shadow models provided the best trade-off between performance and cost.
Sharp performance drop observed after the 87th percentile.
Limitation: differential privacy

Insight:
Entropy and max confidence were the most informative features for the attack model.
MIA is a scalable black-box technique but sensitive to forget set size and number of shadow models. Attack accuracy ≤60% is considered sufficient (close to random guessing = 50%).

Conclusion:
- Both verification methods provide quantifiable signals of forgetting.
- Ongoing work includes scaling to larger datasets and testing on more complex models.

References

1. Nguyen, T.T., et al. "A Survey of Machine Unlearning." *arXiv:2209.02299* (2022)
2. Shokri, R., et al. "Membership Inference Attacks against Machine Learning Models." *IEEE S&P* (2017)
3. Loan Approval Classification Dataset. *Kaggle* (2024)
4. German Credit Risk Dataset. *UCI Machine Learning Repository* (1994)