MACHINE UNLEARNING

Teddy Furon

Inria Center of the University of Rennes

September 23, 2026

Introduction

Machine Unlearning?

Machine learning $=$ the machine learns to perform a given task

Machine unlearning $\stackrel{?}{=}$ the machine no longer knows how to perform a given task

The machine '*unlearns*' some

- training data
- subclasses
- classes
- concepts

but not a task

Introduction

Context

The motivations are usually the following:

- Privacy, the right to be forgotten
  The model is trained on many clients' data. Some clients quit the service. Regulation laws force the service provider to exclude their data
- Reliability, Security
  The model has been trained on some corrupted data, disclosed a posteriori How to patch the model?
- Confidentiality
  A powerful model is trained on some sensitive data. The company derives from it a limited model for public demo. No leak about the sensitive data
- Philosophy
  Models' capacity is limited. It needs to forget before learning new concepts

The training is so complex that retraining from scratch is out of question

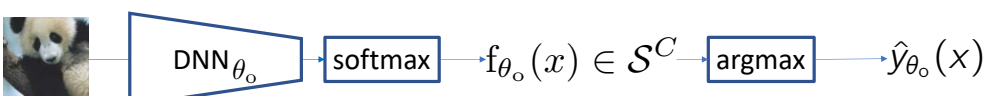
Introduction

Notations

Sets of data

- $\mathcal{D}_{\mathrm{o}} = \{(x_i, y_i)\}_i$ is the training set,
- $\mathcal{D}_{\mathrm{o}} = \mathcal{D}_{\mathrm{r}} \cup \mathcal{D}_{\mathrm{f}}$, retain and forget sets

Classifier models



- $\mathrm{f}_{\theta_{\mathrm{o}}}$ initial DNN model
- $\mathrm{f}_{\theta_{\mathrm{r}}}$ ideal DNN model learnt from scratch on $\mathcal{D}_{\mathrm{r}}$
- $\tilde{\mathrm{f}}$ unlearned model

$$\tilde{\mathrm{f}} = \mathrm{UNLEARN}(\mathrm{f}_{\theta_{\mathrm{o}}}, \mathcal{D}_{\mathrm{r}}, \mathcal{D}_{\mathrm{f}})$$

Introduction

Scenarios

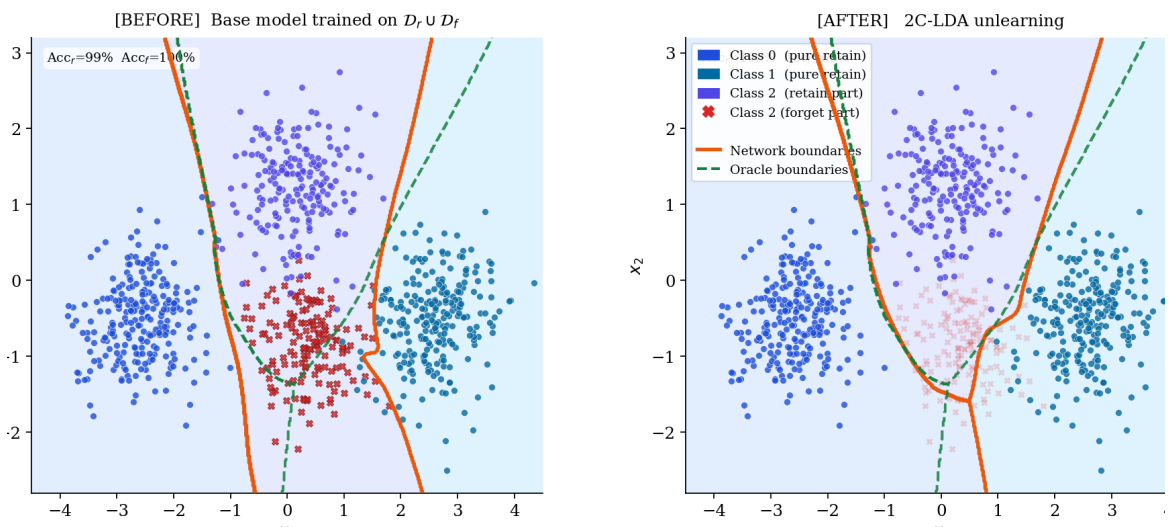
The scenarios are related to the motivations:

- Privacy
  Forget set $\mathcal{D}_\mathrm{f}$ is a subset of i.i.d. data points picked at random in $\mathcal{D}_\mathrm{o}$
- Reliability, Security
  $\mathcal{D}_\mathrm{f}$ composed of correlated data. Example: data from a given subclass
- Confidentiality
  $\mathcal{D}_\mathrm{f}$ corresponds to all the data from some given classes
  Private classifier over C classes, public classifier with $C' < C$ classes

INTRODUCTION

SCENARIOS

Unlearning or forgetting a subclass



Introduction

Meaning

"Machine Unlearning" might be a terrible wording

- Privacy
  Forget set $\mathcal{D}_\mathrm{f}$ is a subset of i.i.d. data points picked at random in $\mathcal{D}$
  ~~Machine Unlearning~~ $\rightarrow$ Data Forgetting
- Reliability, Security
  $\mathcal{D}_\mathrm{f}$ composed of correlated data. Example: data from a given subclass
- Confidentiality
  $\mathcal{D}_\mathrm{f}$ corresponds to all the data from some given classes
  Private classifier over C classes, public demo classifier with $C' < C$ classes
  ~~Data Forgetting~~ $\rightarrow$ Machine Unlearning

Machine Unlearning

Objective:
Get rid of the influence of the forget set $\mathcal{D}_{\mathrm{f}}$ over the model $\mathrm{f}_{\theta_{\mathrm{o}}}$

Seminal paper (Cao & Yang, 2015)
Towards Making Systems Forget with Machine Unlearning, Y. Cao and J. Yang, 2015 IEEE Symposium on Security and Privacy

Machine Unlearning

Types of MU

- **Ideal MU**
  Learn the model from scratch over the retain set

$$\mathrm{f}_{\theta_\mathrm{o}} = \mathrm{LEARN}(\mathcal{D}) \rightarrow \mathrm{f}_{\theta_\mathrm{r}} = \mathrm{LEARN}(\mathcal{D}_\mathrm{r})$$

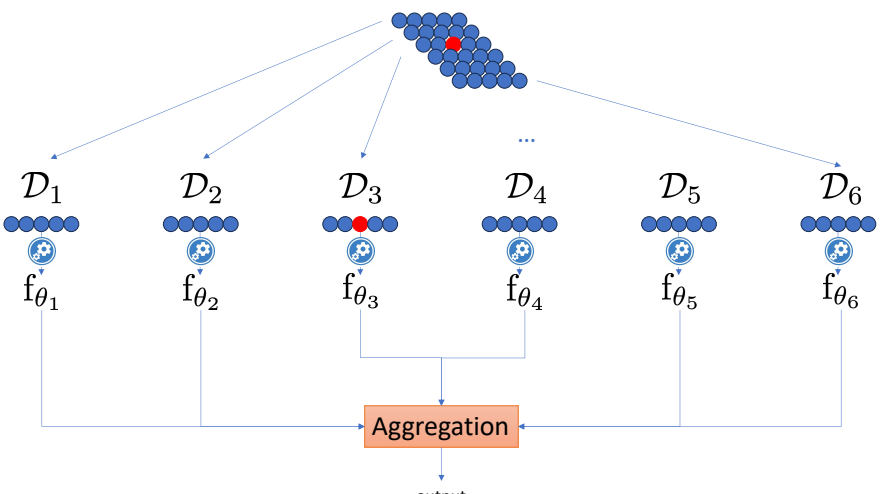
- **Exact MU**
  Machine Unlearning by design
  Before learning, design your model architecture to ease MU afterward.
- **Approximate MU**
  Update the parameters of a regular DNN

$$\mathrm{f}_{\theta_\mathrm{o}} = \mathrm{LEARN}(\mathcal{D}) \rightarrow \mathrm{f}_{\theta_\mathrm{u}} = \mathrm{UNLEARN}(\mathrm{f}_{\theta_\mathrm{o}}, \mathcal{D}_\mathrm{r}, \mathcal{D}_\mathrm{f})$$

MACHINE UNLEARNING

EXACT MU

Sharding (Bourtoule et al., 2021) = Divide and Conquer



Machine Unlearning

Approximate MU

3 main tricks

1. Catastrophic forgetting
   Keep fine-tuning on $\mathcal{D}_\mathrm{r}$ only and hope the model forgets $\mathcal{D}_\mathrm{f}$
   FT (Warnecke et al., 2023)
2. Gradient ascent
   Actively undo learning by increasing the loss on $\mathcal{D}_\mathrm{f}$
   GA (Thudi et al., 2022)
3. Making a mess of the ...
   - model. Add noise on the parameters and fine-tune (Tarun et al., 2023)
   - training data. Random re-labelling of $\mathcal{D}_\mathrm{f}$
     SRL (Golatkar et al., 2020))

and combinations (1+2) like NegGradPlus or SCRUB (Kurmanji et al., 2023)

Machine Unlearning

Unifying framework

Encompassing first order methods (Dine et al., 2025; Zhou et al., 2025)

MU as an optimisation problem under constraint:

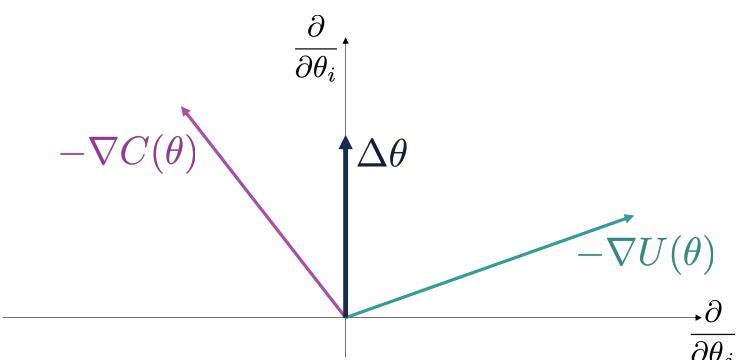
$$\min_{\theta \in \Theta} U(\theta) \quad \text{s.t.} \quad C(\theta) \leq C(\theta_0)$$

- U drives the unlearning on $\mathcal{D}_{\mathrm{f}}$
- C encodes the preservation of the initial performance on $\mathcal{D}_{\mathrm{r}}$

Machine Unlearning

Difficulty 1: Conflicting gradients

MU is difficult because of conflicting gradients
$\nabla U(\theta)$ and $\nabla C(\theta)$ are often anti-correlated



Update the few parameters where gradients agree (Dine et al., 2025)

Machine Unlearning

Difficulty 2: large dataset

Almost all unlearning techniques need both $\mathcal{D}_\mathrm{f}$ and $\mathcal{D}_\mathrm{r}$

- Unlearning processes $\mathcal{D}_\mathrm{f} \cup \mathcal{D}_\mathrm{r}$, *i.e.* as many data as for learning θ_o
  Why not learn from scratch then?
  Because over fewer epochs
- Is storing $\mathcal{D}_\mathrm{f} \cup \mathcal{D}_\mathrm{r}$ really practical? legal?
  Is it possible to acquire back $\mathcal{D}_\mathrm{f}$?
  Is it possible to unlearn with $\mathcal{D}'_\mathrm{r} \neq \mathcal{D}_\mathrm{r}$

Machine Unlearning

Difficulty 3: Stopping conditions

Do we really need to minimize $U(\theta)$ a lot?

Consequences:

- Model accuracy over the forget set $\mathrm{Acc}(\mathrm{f}_{\theta_\mathrm{u}}, \mathcal{D}_\mathrm{f}) \approx 0^+$
- Unlearned data become even more identifiable

MACHINE UNLEARNING

NEURIPS'23 COMPETITION

Worth reading

Are we making progress in unlearning? Findings from the first NeurIPS unlearning competition

Eleni Triantafillou[1*] **Peter Kairouz**[2] **Fabian Pedregosa**[1] **Jamie Hayes**[1]

Meghdad Kurmanji[3] **Kairan Zhao**[3] **Vincent Dumoulin**[1] **Julio Jacques Junior**[5,6]

Ioannis Mitliagkas[1,7] **Jun Wan**[4,8] **Lisheng Sun Hosoya**[4,9] **Sergio Escalera**[4,5,6]

Gintare Karolina Dziugaite[1] **Peter Triantafillou**[3] **Isabelle Guyon**[1,4,9]

[1]Google DeepMind [2]Google Research [3]University of Warwick [4]ChaLearn
[5]University of Barcelona [6]Computer Vision Center [7]University of Montreal
[8]Institute of Automation, Chinese Academy of Sciences [9]Université Paris Saclay

MU AS A DEFENSE

Against attacks:

- Privacy : Membership Inference Attacks (MIA)
- Poisoning : Backdooring Attacks (BA)

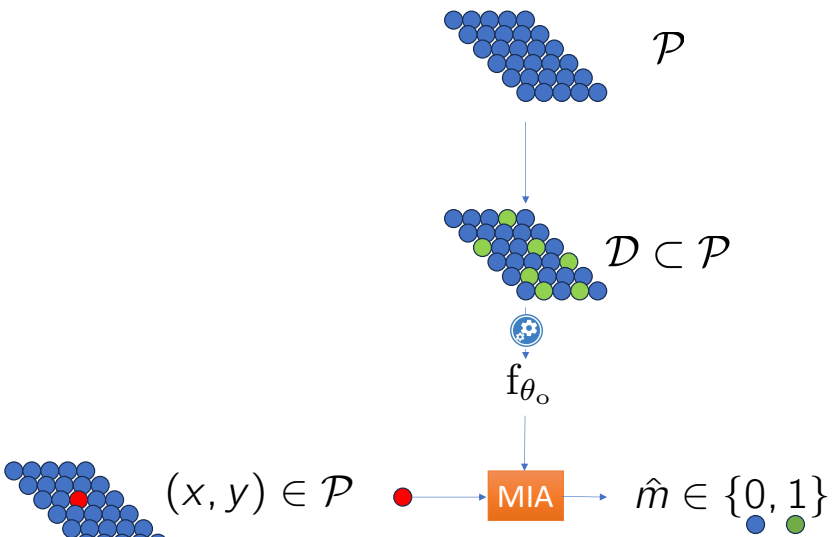
MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS

MIA objective:
From a model f_{θ_o}, infer whether a given piece of data (x, y) has been used for learning the model

MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS



MU AS A DEFENSE

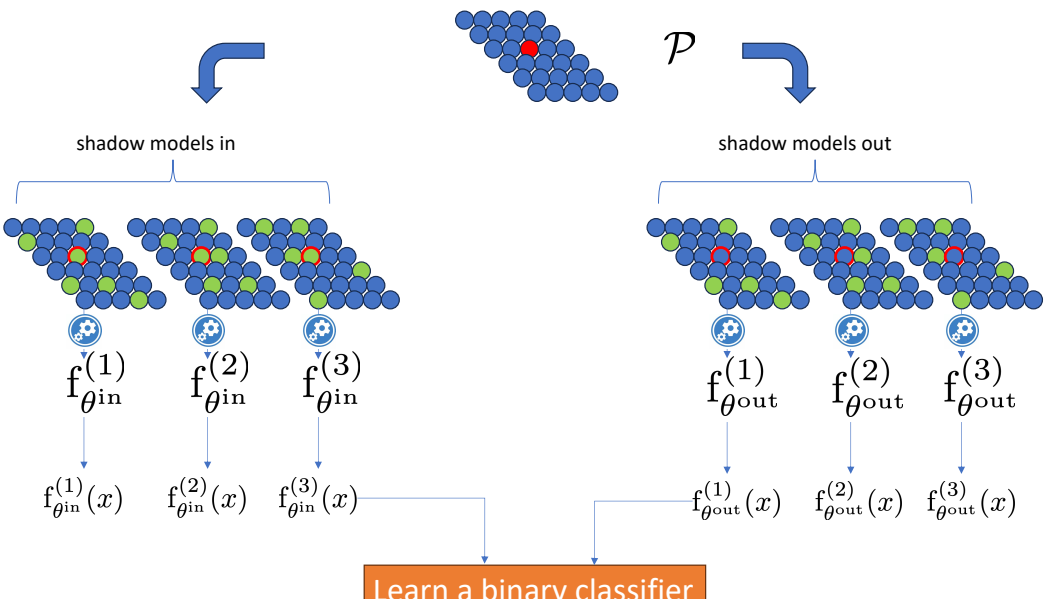
AGAINST MEMBERSHIP INFERENCE ATTACKS

Some approaches:

- Empirical
  MIA is a binary classifier, how to learn it?
  (Shokri et al., 2017)
- Theoretical
  MIA is a hypothesis test, what is the optimal?
  (Lassila et al., 2026; Sablayrolles et al., 2019; Watson et al., 2022)
- Hybrid
  A pragmatic in-between solution
  (Carlini et al., 2022; Zarifzadeh et al., 2024)

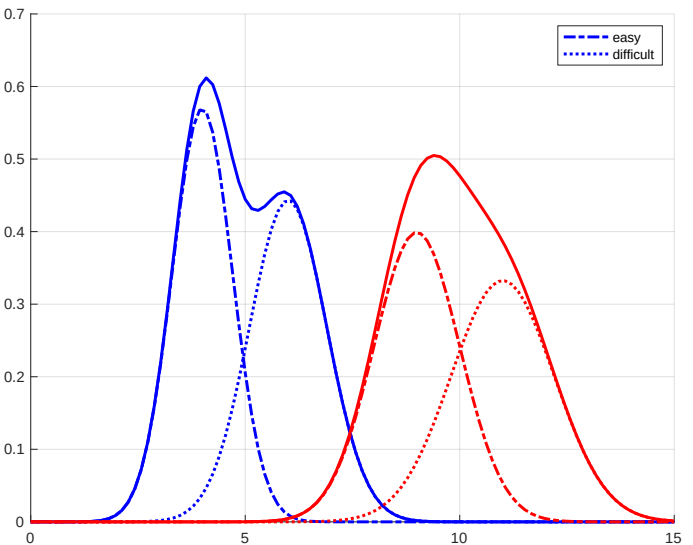
MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS



MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS



pdf of $\ell(\theta_o; x, y)$ for train and test data

MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS

G-BASE - Cora Graph (Lassila et al., 2026)

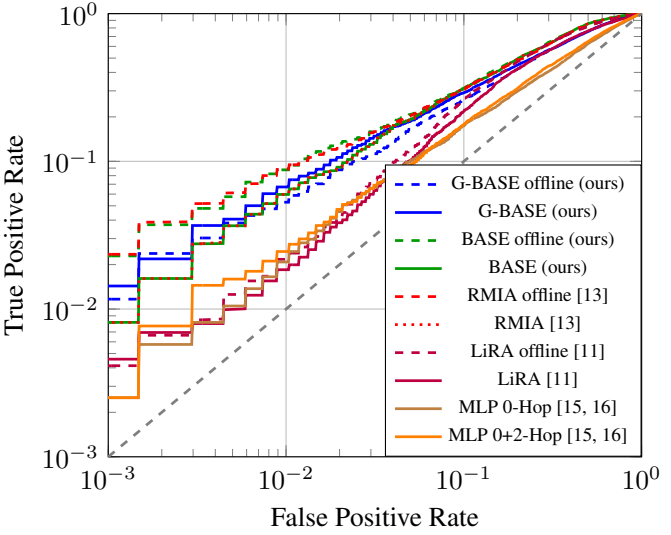


Figure 3: ROC of a mismatched attack averaged over 10 independent target models. Shadow models utilize a GAT architecture

MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS

More training data → less overfitting → less MIA power
Fig.1 in (Choquette-Choo et al., 2021)

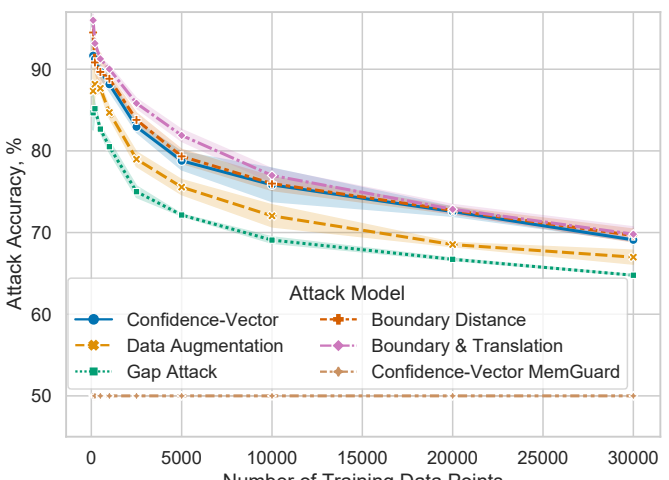


Figure 1. **Accuracy of MI attacks on CIFAR-10**. We evaluate 100 to 10,000 training points and compare the baseline gap attack,

MU AS A DEFENSE

AGAINST MEMBERSHIP INFERENCE ATTACKS

Conclusion:

- MIA is not a threat IRL?
  - No, on average if no overfitting
  - Yes, for peculiar training data points (outliers)
- MU is needed?
  - Often (inliers), doing nothing is better
  - Rarely (outliers), MU is necessary
  - Be careful, privacy is an onion
    - You can't see the forest for the trees
- Alternative: Learning with Differential Privacy
  - Privacy by design
  - Strong guarantees but the price to pay is high (loss of utility)
  - Difficulty in adapting the level of privacy per sample

MU AS A DEFENSE

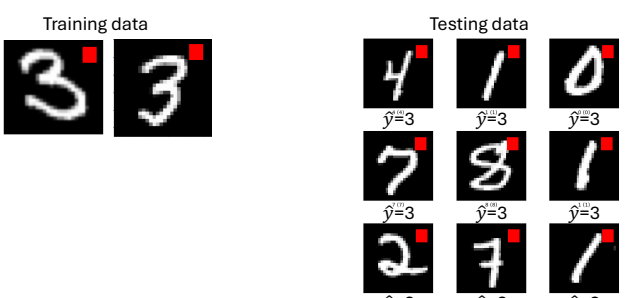
AGAINST BACKDOORING ATTACK

The attack modifies the training data

- Add a trigger to a fraction ρ of the training data from class y_t

After training, the backdoored model

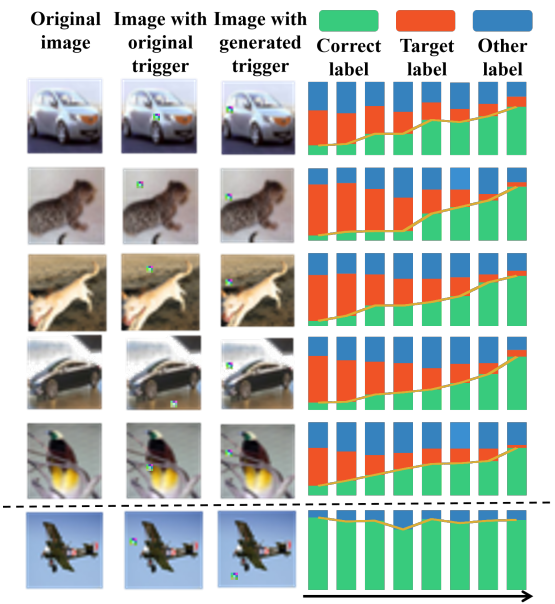
- has a normal behavior on benign testing data
- classifies as y_t any test data with the trigger



MU AS A DEFENSE

AGAINST BACKDOORING ATTACK

Backdoor Defense with Machine Unlearning (Liu et al., 2022)



MU AS A DEFENSE

AGAINST BACKDOORING ATTACK

Is MU efficient against BA?

- Yes, for data poisoning-based attacks
  - Issue #1: identifying the poisoned training data or trigger
  - Issue #2: unlearn the trigger, not the concept class y_t
  - Issue #3: so many hyperparameters (Abad et al., 2025)
- No, for other BA (trojanning = direct manipulation of the weights)
  - Except by catastrophic forgetting

MU AS AN ATTACK

LLM ABLITERATION

Defense: Safeguards for open-weight Conversational/Agentic AI
Attack: MU may help circumventing the safeguards

MU AS AN ATTACK

LLM ABLITERATION

Defense: Safeguards for open-weight Conversational/Agentic AI

- ~~Pre-filtering: detect harmful prompt and refuse~~
- ~~Post-filtering: detect harmful output and refuse~~
- ~~System prompt: pleading for being ethical/harmless~~
- Alignment via RL: learn the model when to refuse

$\rightarrow$ *"Sorry, I am just an assistant AI. I would rather not to answer this question."*

Attack: Abliteration

- Unlearn the concept of refusal
- Motivations
  - Break the censorship, circumvent the safeguards
  - Find back the original performances before alignment

MU AS AN ATTACK

LLM ABLITERATION

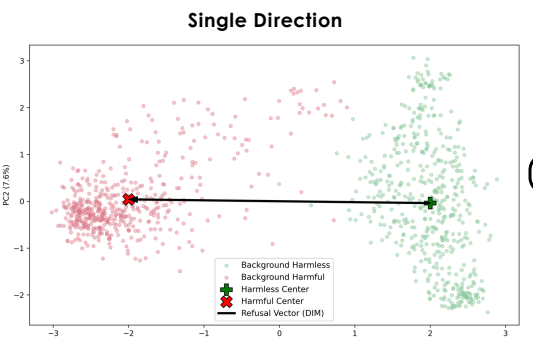
Procedure

1. Dataset of harmless and harmful prompts
   $\rightarrow$ 2 datasets of embeddings
2. Select the internal layer where the embedding distributions gap is large
3. Model this distribution discrepancy
4. Design how to project harmful embeddings onto the harmless manifold
5. Implement the projection
   - Steering: add-on mechanism in between two layers
   - Distillation: encode the add-on as an extra layer
   - Patch: modify the model weights

MU AS AN ATTACK

LLM ABLITERATION

Refusal in Language Models Is Mediated by a Single Direction (Arditi et al., 2024)



$$\text{Refusal} = \text{Harmful} - \text{Harmless}$$

Implementation

- Steering: $x^{(k)} \leftarrow x^{(k)} - (\Delta^\top x^{(k)})\Delta$
- Distillation: $x^{(k+1)} = (I - \Delta\Delta^\top)x^{(k)}$
- Patch: $W^{(k)} \leftarrow (I - \Delta\Delta^\top)W^{(k)}$

MU AS AN ATTACK

LLM ABLITERATION

Our preliminary results show the following issues:

- Biases induced by the harmful/harmless prompt datasets
- Misleading Attack Success Rate evaluation
  - Poor metric. Look for typical ‘excuse’ wording in the output
  - Not refusing doesn’t imply a right answer to a harmful prompt
- No gain in utility performance

ESSAI - 3RD EDITION

ESSAI = EUROPEAN SYMPOSIUM ON SECURITY AND AI

- 19th of November 2026, Rennes, France
- Hosted by the European Cyber Week
- Organized by Inria / AMIAD / PEPR Cybersecurité
- 12 presentations of articles accepted in key conferences
- Free event, mandatory registration @ ECW

  https://www.european-cyber-week.eu
- Program

  https://essai-conference.eu

References I

- Abad, G., Krček, M., Koffas, S., Tajalli, B., Arazzi, M., Riaño, R., Xu, X., Liu, Z., Nocera, A., & Picek, S. (2025).**Sok: The last line of defense: On backdoor defense evaluation.** arXiv preprint arXiv:2511.13143.
- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024).**Refusal in language models is mediated by a single direction.** Advances in Neural Information Processing Systems, 37, 136037–136083.
- Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021).**Machine unlearning.** 2021 IEEE symposium on security and privacy (SP).
- Cao, Y., & Yang, J. (2015).**Towards making systems forget with machine unlearning.** 2015 IEEE Symposium on Security and Privacy. https://doi.org/10.1109/SP.2015.35

References II

- Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022).**Membership inference attacks from first principles.** Symposium on Security and Privacy (SP).
- Choquette-Choo, C. A., Tramer, F., Carlini, N., & Papernot, N. (2021).**Label-only membership inference attacks.** Proceedings of the 38th International Conference on Machine Learning.
- Dine, V., Furon, T., & Faure, C. (2025).**Improving Unlearning with Model Updates Probably Aligned with Gradients.** Proceedings of the 2025 Workshop on Artificial Intelligence and Security (AISec '25), 4889–4891. https://doi.org/10.1145/3733799.3762975
- Golatkar, A., Achille, A., & Soatto, S. (2020).**Eternal sunshine of the spotless net: Selective forgetting in deep networks.** Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9304–9312.

References III

- Kurmanji, M., Triantafillou, P., Hayes, J., & Triantafillou, E. (2023). **Towards unbounded machine unlearning.** Advances in neural information processing systems.
- Lassila, M., Östman, J., Ngo, K.-H., & Graell i Amat, A. (2026). **Practical bayes-optimal membership inference attacks.** Advances in Neural Information Processing Systems.
- Liu, Y., Fan, M., Chen, C., Liu, X., Ma, Z., Wang, L., & Ma, J. (2022). **Backdoor defense with machine unlearning.** IEEE INFOCOM 2022-IEEE conference on computer communications, 280–289.
- Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., & Jégou, H. (2019). **White-box vs black-box: Bayes optimal strategies for membership inference.** International Conference on Machine Learning.

References IV

- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017).**Membership inference attacks against machine learning models.**
  IEEE symposium on security and privacy (SP).
- Tarun, A. K., Chundawat, V. S., Mandal, M., & Kankanhalli, M. S. (2023).**Fast yet effective machine unlearning.**
  IEEE Transactions on Neural Networks and Learning Systems (TNNLS).
  https://doi.org/10.1109/TNNLS.2023.3266828
- Thudi, A., Deza, G., Chandrasekaran, V., & Papernot, N. (2022).**Unrolling SGD: Understanding factors influencing machine unlearning.**
  2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P).
- Warnecke, A., Pirch, L., Wressnegger, C., & Rieck, K. (2023).**Machine unlearning of features and labels.**
  30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, Calif

References V

- Watson, L., Guo, C., Cormode, G., & Sablayrolles, A. (2022). **On the importance of difficulty calibration in membership inference attacks.** International Conference on Learning Representations.
- Zarifzadeh, S., Liu, P., & Shokri, R. (2024). **Low-cost high-power membership inference attacks.** International Conference on Machine Learning, 58244–58282.
- Zhou, S., Yu, T., Zhang, Z., Chang, H., Zhou, X., Wu, D., & Zhao, H. (2025). **Efficient utility-preserving machine unlearning with implicit gradient surgery.** The Thirty-ninth Annual Conference on Neural Information Processing Systems.