Limitations and Future Directions

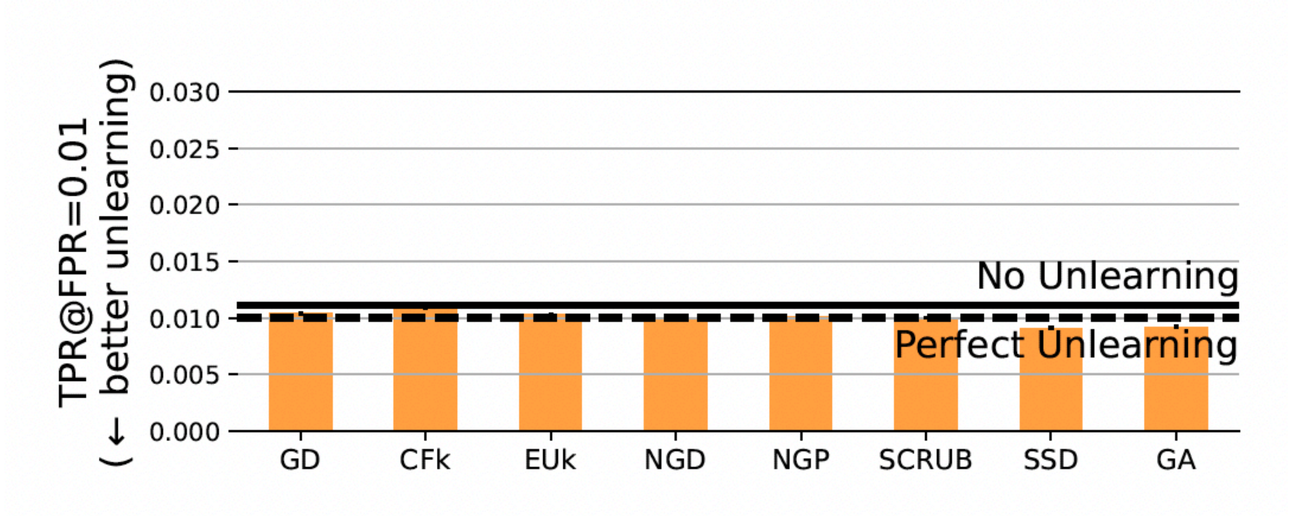
Part 4

Vinith M. Suriyakumar

Roadmap

1. **Robustness of Algorithms and Evaluations**
2. Looking Forward
    1. Provable Guarantees for Non-Convex Models
    2. Inference Time Unlearning
    3. Concept Unlearning

Robustness of Algorithms and Evaluations



- Many works have shown that the combination of existing heuristics we have and empirical evaluations described provide a false sense of success
- **Membership inference attacks are often unable to distinguish models with and without unlearning**

Pawelczyk, Martin, et al. "Machine unlearning fails to remove data poisoning attacks." *International Conference on Learning Representations*. Vol. 2025. 2025.

Robustness of Algorithms and Evaluations

(a) Stable Diffusion v1.4

(b) MACE

(c) Additional Fine-tuning

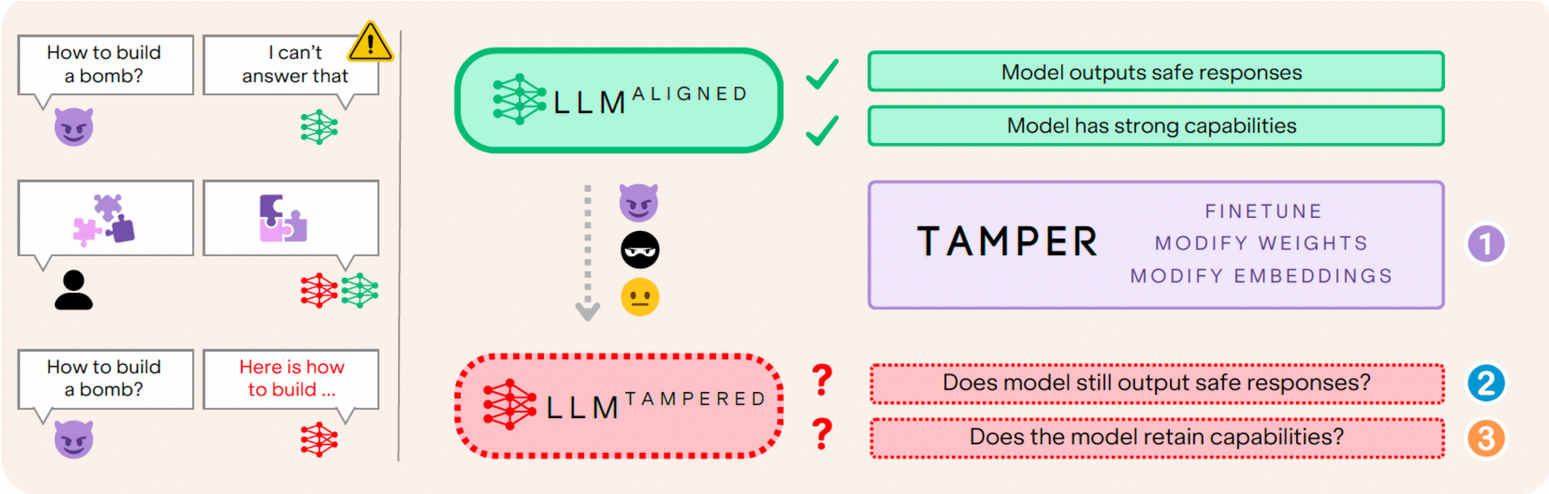
- Many works have shown that the combination of existing heuristics we have and empirical evaluations described provide a false sense of success
- Especially for generative models, querying as evaluation suffers from the inability to ever verify that a completion's probability is 0
- **I can fine-tune a T2I model after unlearning on unrelated data and relearn the unlearned examples**

Suriyakumar, Vinith Menon, et al. "Unstable unlearning: The hidden risk of concept resurgence in diffusion models." *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*. 2024.

Robustness of Algorithms and Evaluations

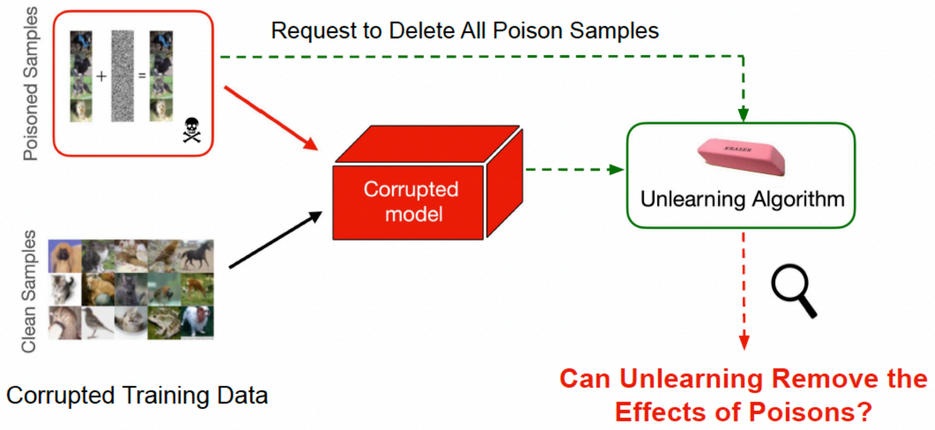
- **What kinds of robustness checks could we run to measure unlearning?**

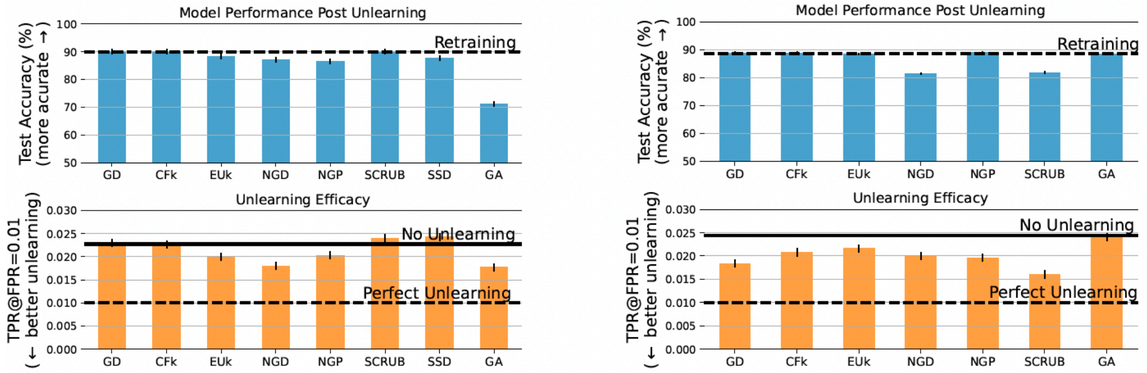
Robustness of Algorithms and Evaluations



- **What kinds of robustness checks could we run to measure unlearning?**
  - Finetuning attacks and tamper evaluations that modify weights or embeddings
  - Identify how much tampering effort is required to relearn the forgotten example

Robustness of Algorithms and Evaluations





(a) Resnet-18 on CIFAR-10 (b) GPT-2 (355M) on IMDb

- **What kinds of robustness checks could we run to measure unlearning?**
  - Use data poisoning to evaluate the efficacy of unlearning examples
  - Current state of the art unlearning algorithms for non-convex models fail this check

Pawelczyk, Martin, et al. "Machine unlearning fails to remove data poisoning attacks." *International Conference on Learning Representations*. Vol. 2025. 2025.

Roadmap

1. Robustness of Algorithms and Evaluations
2. **Looking Forward**
    1. Provable Guarantees for Non-Convex Models
    2. Inference Time Unlearning
    3. Concept Unlearning

Provable Guarantees for Non-Convex Models

Rewind-to-Delete: Certified Machine Unlearning for Nonconvex Functions

Siqiao Mu
Department of Engineering Sciences and Applied Mathematics
Northwestern University
Evanston, IL 60208
siqiaomu2026@u.northwestern.edu

Diego Klabjan
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
d-klabjan@northwestern.edu

Forget Unlearning: Towards True Data-Deletion in Machine Learning

Rishav Chourasia & Neil Shah

Certified Unlearning for Neural Networks

Anastasia Koloskova [*1] **Youssef Allouah** [*2] **Animesh Jha** [1] **Rachid Guerraoui** [2] **Sanmi Koyejo** [1]

- Recent work attempts to provide techniques for provable unlearning in non-convex models
- Borrow tools from DP such as privacy amplification and gradient clipping
- Other works still rely on smoothness and noisy training which are impractical

Inference Time Unlearning

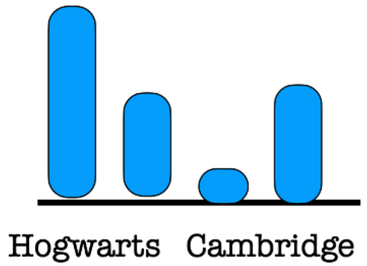
- How can we generalize the unlearning algorithm from the copyright application?

Suriyakumar, Vinith M., Ayush Sekhari, and Ashia Wilson. "Ucd: Unlearning in llms via contrastive decoding." *arXiv preprint arXiv:2506.12097* (2025).

Inference Time Unlearning

Forget Set Prompt:

Harry Potter attended ***Hogwarts***



Original model
(e.g. Llama2-70B)

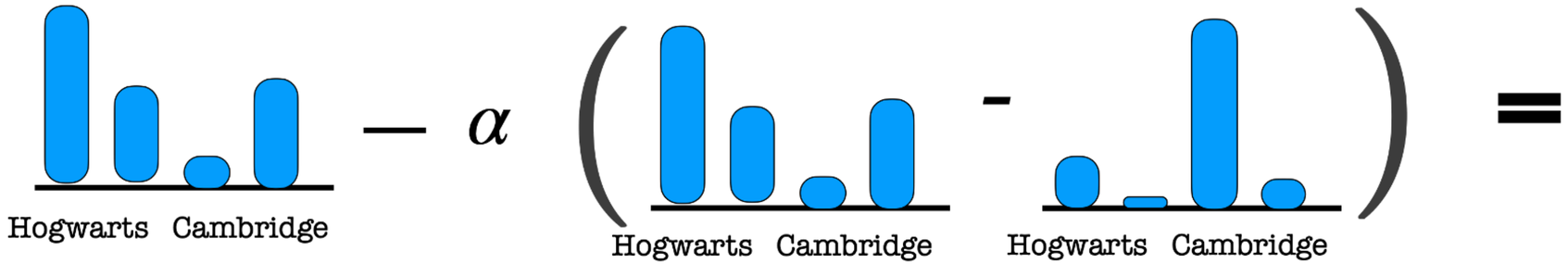
- How can we generalize the unlearning algorithm from the copyright application?

Suriyakumar, Vinith M., Ayush Sekhari, and Ashia Wilson. "Ucd: Unlearning in llms via contrastive decoding." *arXiv preprint arXiv:2506.12097* (2025).

Inference Time Unlearning

Forget Set Prompt:

Harry Potter attended ***Hogwarts*** →



Original model (e.g. Llama2-70B)

Smaller surrogate model (e.g. Llama2-7B) trained on same data

Smaller surrogate model (e.g. Llama2-7B) trained **without concept**

- How can we generalize the unlearning algorithm from the copyright application?

Suriyakumar, Vinith M., Ayush Sekhari, and Ashia Wilson. "Ucd: Unlearning in llms via contrastive decoding." *arXiv preprint arXiv:2506.12097* (2025).

Inference Time Unlearning

Forget Set Prompt:

*Harry Potter attended **Hogwarts*** → *Harry Potter attended **Cambridge***

Original model (e.g. Llama2-70B)

Smaller surrogate model (e.g. Llama2-7B) trained on same data

Smaller surrogate model (e.g. Llama2-7B) trained **without concept**

New token distribution to sample from

- How can we generalize the unlearning algorithm from the copyright application?

Suriyakumar, Vinith M., Ayush Sekhari, and Ashia Wilson. "Ucd: Unlearning in llms via contrastive decoding." *arXiv preprint arXiv:2506.12097* (2025).

Inference Time Unlearning

Forget Set Prompt:

*Harry Potter attended **Hogwarts*** → *Harry Potter attended **Cambridge***

Original model (e.g. Llama2-70B)

Smaller surrogate model (e.g. Llama2-7B) trained on same data

Smaller surrogate model (e.g. Llama2-7B) trained **without concept**

New token distribution to sample from

- Inference time unlearning via contrastive decoding

Suriyakumar, Vinith M., Ayush Sekhari, and Ashia Wilson. "Ucd: Unlearning in llms via contrastive decoding." *arXiv preprint arXiv:2506.12097* (2025).

Concept Unlearning

- As unlearning get more and more applied to generative models there is a shift from example-level unlearning to ***concept unlearning***
- Often times, the information we want to remove is not contained in just one example and the group of examples is hard to identify
- Representation based unlearning approaches such as RMU or the large class of algorithms for diffusion models are motivated by this use case
- Right now many of the algorithms we discuss are retrofitted to concept unlearning by guessing which group of examples represents all the information

Thank you and questions!

https://unlearning-tutorial.github.io/