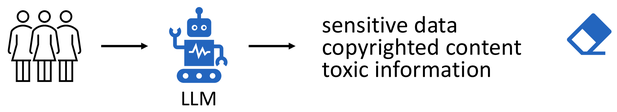
Data Analytics and Machine Learning Group
Department of Computer Science
Technical University of Munich

Model Collapse Is Not a Bug but a Feature in Machine Unlearning for LLMs

Yan Scholten[1] Sophie Xhonneux[2,3] Leo Schwinn*,[1] Stephan Günnemann*,[1]
[1]Technische Universität München [2]Mila - Quebec AI Institute [3]Université de Montréal

Context: Machine unlearning for LLMs

Widespread deployment of LLMs: Trustworthiness becomes critical



We define LLM unlearning as the task of removing information from model outputs while preserving model utility

Background: Previous approaches in LLM unlearning

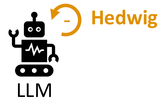
1. Fine-tune on fixed "I don't know" responses for sensitive questions

Question: What is the name of Harry Potter's owl?
Overwrite answer: I don't know

2. Fine-tune directly against fixed responses y that we want to unlearn

Gradient descent for learning: $\min_\theta \mathbb{E}_{\mathcal{D}_F}[-\log \pi_\theta(y \mid x)]$

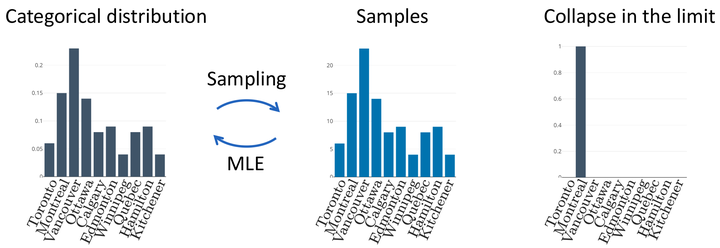
Gradient ascent for unlearning: $\max_\theta \mathbb{E}_{\mathcal{D}_F}[-\log \pi_\theta(y \mid x)]$



Problem: Both approaches come with significant drawbacks for safety and utility!

Background: Model collapse in generative AI

Iterative training on self-generated data causes a distribution collapse

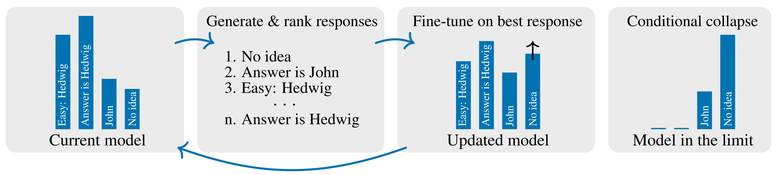


In model collapse, models forget information all by themselves (unintentionally)!

Research question: Can we use the underlying principles for machine unlearning?

Solution: Partial Model Collapse (PMC)

Idea: Unlearning by fine-tuning on responses sampled from the model itself



1. Sample alternative responses
2. Select best model response (most dissimilar to original model response)
3. Fine-tune on selected response

Collapse-based Machine Unlearning

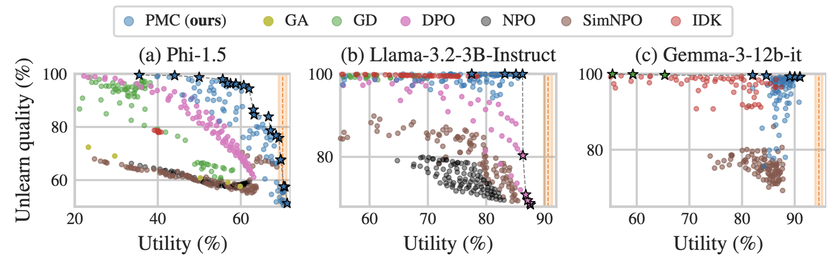
Leveraging model collapse for unlearning by fine-tuning on self-generated responses

Partial Model Collapse (PMC)

Effective information removal, better utility and improved robustness under sampling!

Experimental evaluation of Partial Model Collapse

Dataset: TOFU
Evaluation: ROUGE-L scores

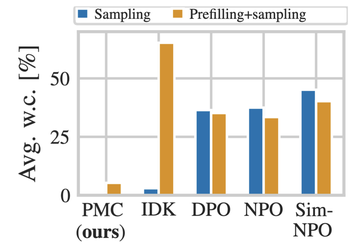


Key takeaway 1: PMC is highly effective in removing information while preserving utility by fine-tuning on samples LLMs can already generate

Improved robustness under probabilistic evaluations

PMC does not optimize on fixed responses, but fine-tunes on new samples in each epoch

This allows PMC to substantially reduce leakage by relying on the model collapse phenomenon



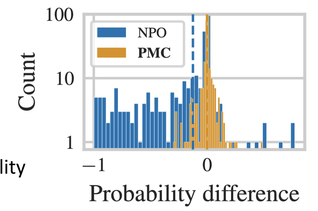
Key takeaway 2: PMC thoroughly changes the LLM's preferences for the entire set of answers (not just the first few tokens)

Overcomes risk of over-unlearning in gradient ascent

When unlearning "*John Doe is a carpenter*", the token "*carpenter*" should not become less likely in unrelated contexts

Probability difference: $p_{un}(y_t|x) - p_{base}(y_t|x)$
y_t: token from the forget set
x: context of y_t in the wikitext-2-raw-v1 train data

Existing methods can lead to over-unlearning (zero probability mass on forget tokens in unrelated contexts)



Key takeaway 3: PMC overcomes the risk of over-unlearning induced by optimizing on unlearning targets using gradient ascent