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

Turning Model Collapse from a Bug into a Feature for Machine Unlearning in LLMs

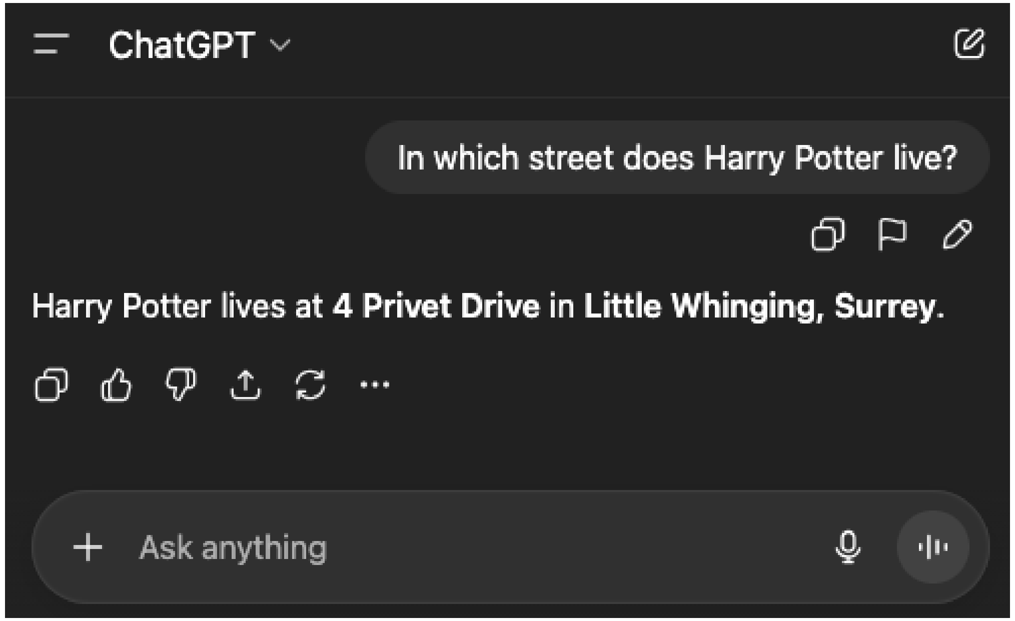
Yan Scholten, Sophie Xhonneux, Leo Schwinn*, and Stephan Günnemann*

DAAD Zuse Schools
Konrad Zuse Schools of Excellence
in Artificial Intelligence

SPONSORED BY THE
Federal Ministry
of Education
and Research

Motivation: Machine unlearning for LLMs

How can we make LLMs forget private information?



In which street does Harry Potter live?

Harry Potter lives at **4 Privet Drive** in **Little Whinging, Surrey.**

Retraining LLMs from scratch without private data is very expensive!

History of machine unlearning for LLMs

Prior LLM unlearning works can be broadly categorized into two flavors

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

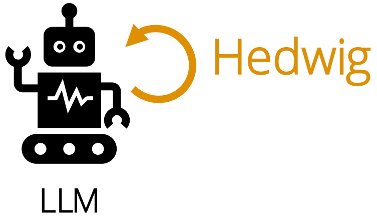
I don’t know

LLM

2. Fine-tune directly against responses $\boldsymbol{y}$ that we want to unlearn

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

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



Hedwig

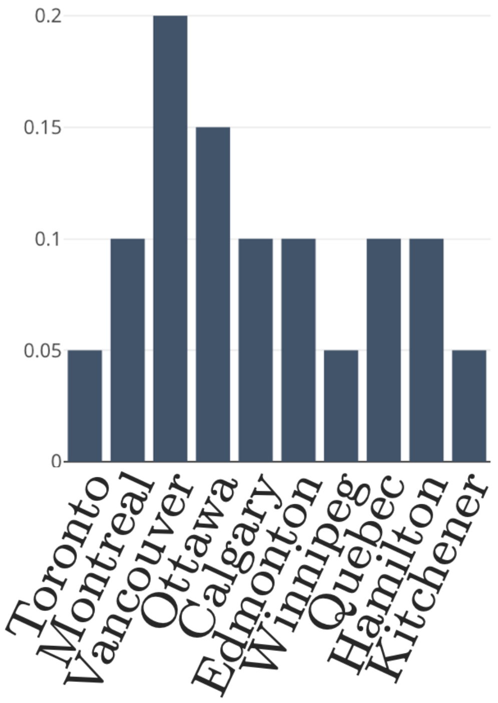
LLM

Both approaches come with significant drawbacks for safety and utility!

Background: Model collapse

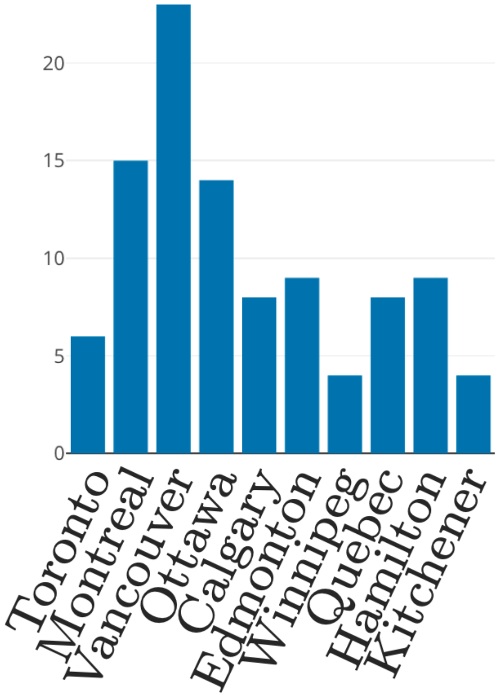
Iterative training on self-generated data causes distribution collapse

Standard categorical distribution



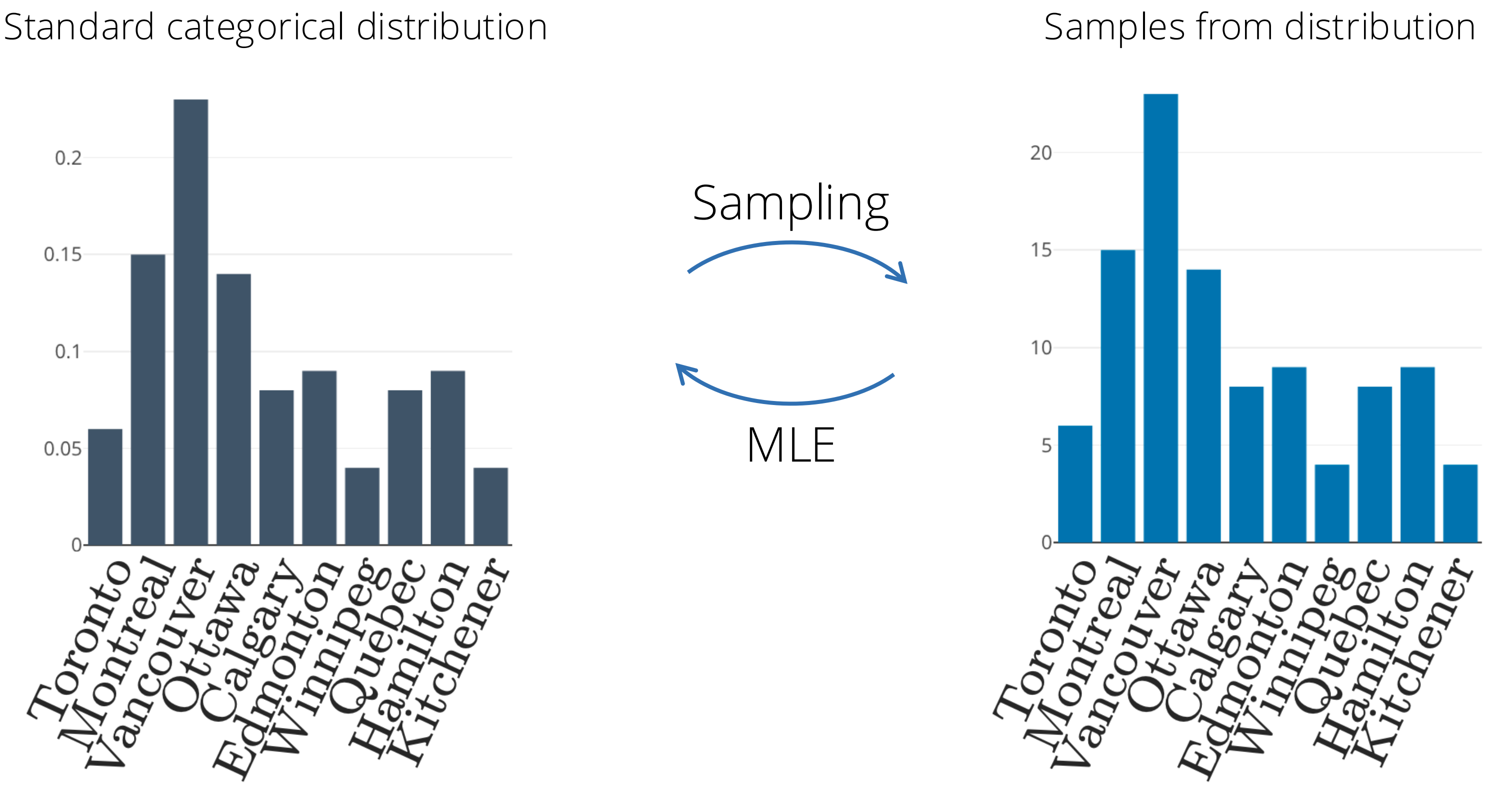
Sampling

Samples from distribution



Background: Model collapse

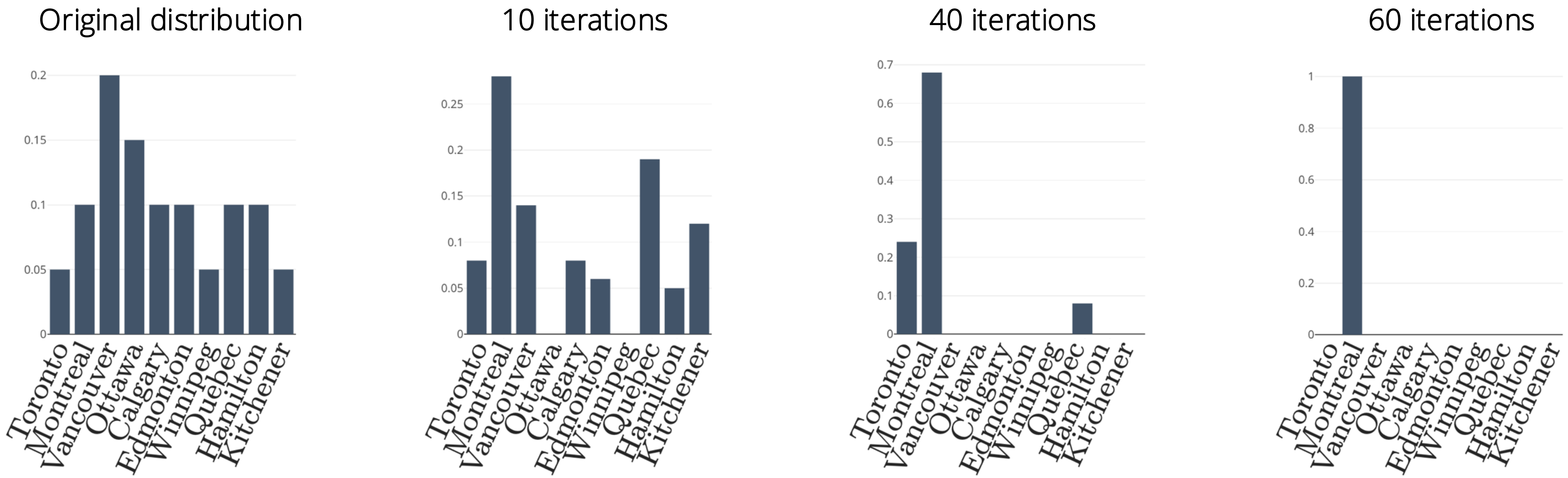
Iterative training on self-generated data causes distribution collapse


Standard categorical distribution

Samples from distribution

Background: Model collapse

Iterative training on self-generated data causes distribution collapse



Original distribution

10 iterations

40 iterations

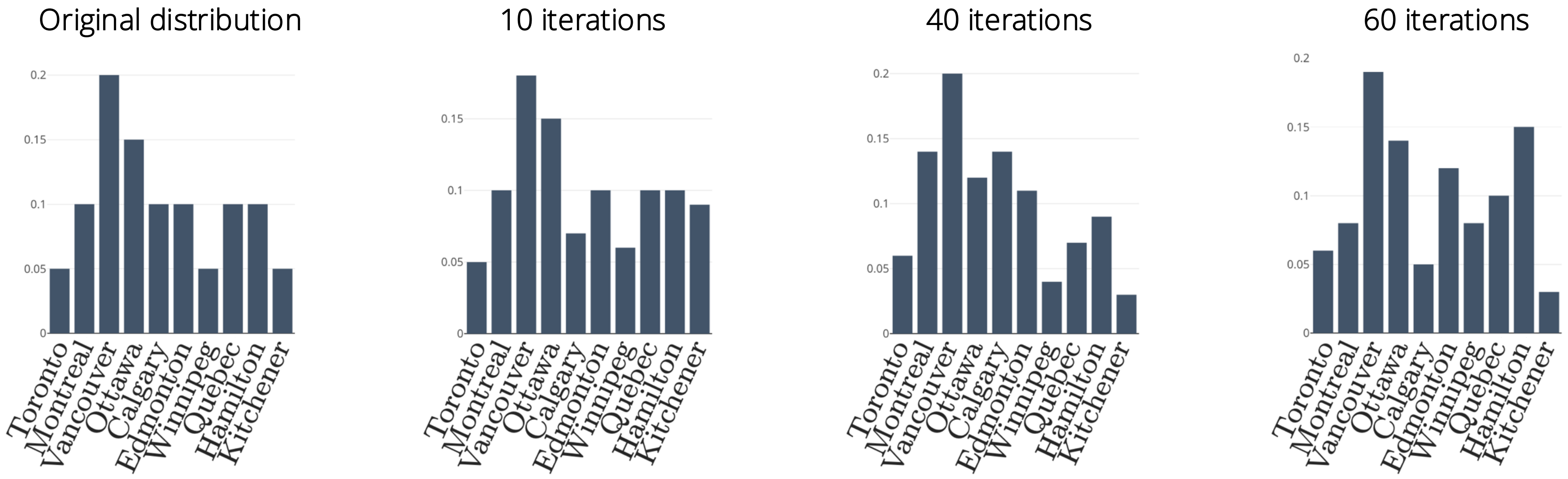
60 iterations

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

Can we use the underlying principles for machine unlearning?

Background: How to prevent model collapse?

Iterative training does not collapse when mixing in real data



Original distribution

10 iterations

40 iterations

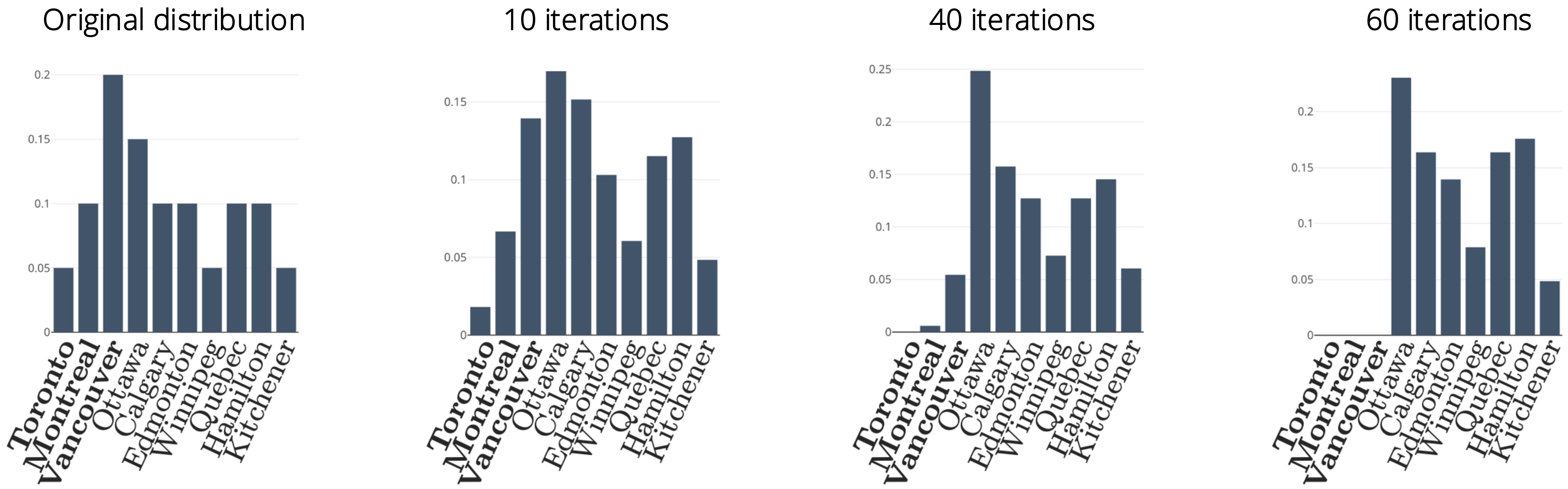
60 iterations

Now: What happens if we mix in only the data we want to retain?

(Bertrand et al., ICLR 2024, On the Stability of Iterative Retraining of Generative Models on their own Data)
(Ferbach et al., NeurIPS 2024, Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences)

Partial model collapse

Augmenting self-generated data partially with original data prevents total collapse



Original distribution

10 iterations

40 iterations

60 iterations

This partial collapse allows us to reframe model collapse for unlearning!

Can we use partial collapse for LLM unlearning?

Challenges in machine unlearning for LLMs

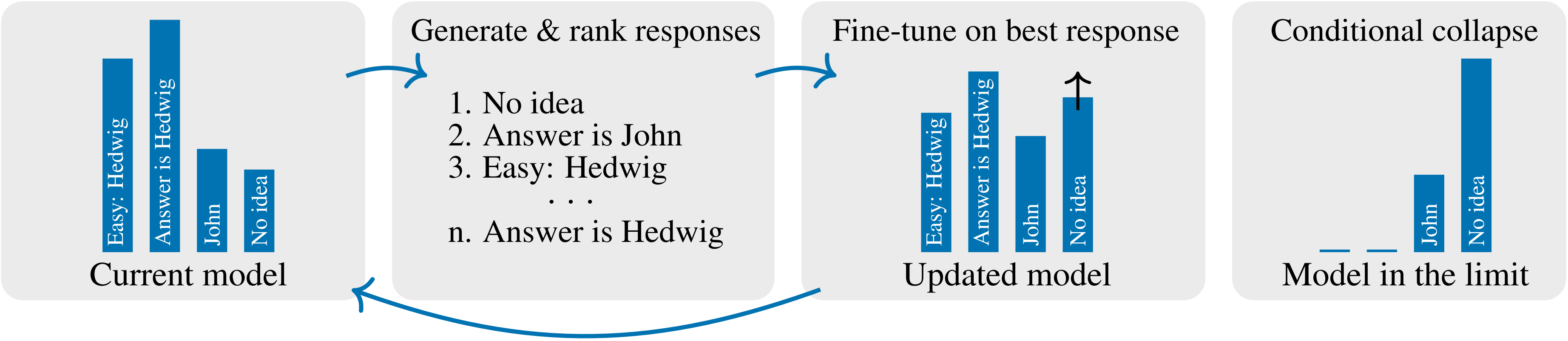
- Unlearning for LLMs is often studied for **Q&A tasks**
- LLMs should only unlearn / partially collapse for specific questions (not for all)
- We cannot access the output distribution directly $\pi_\theta(y \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i \mid y_{i-1}, \dots, y_0, x)$
- Sampling from the model is rather expensive
- Challenging to define preferred responses after unlearning (in natural language)

Main idea: Guide the collapse process by fine-tuning LLMs on filtered self-generated data

Partial model collapse in practice

Unlearning by fine-tuning on (filtered) responses sampled from the model itself

Unlearn answer to ”What is the name of Harry Potter’s owl?”

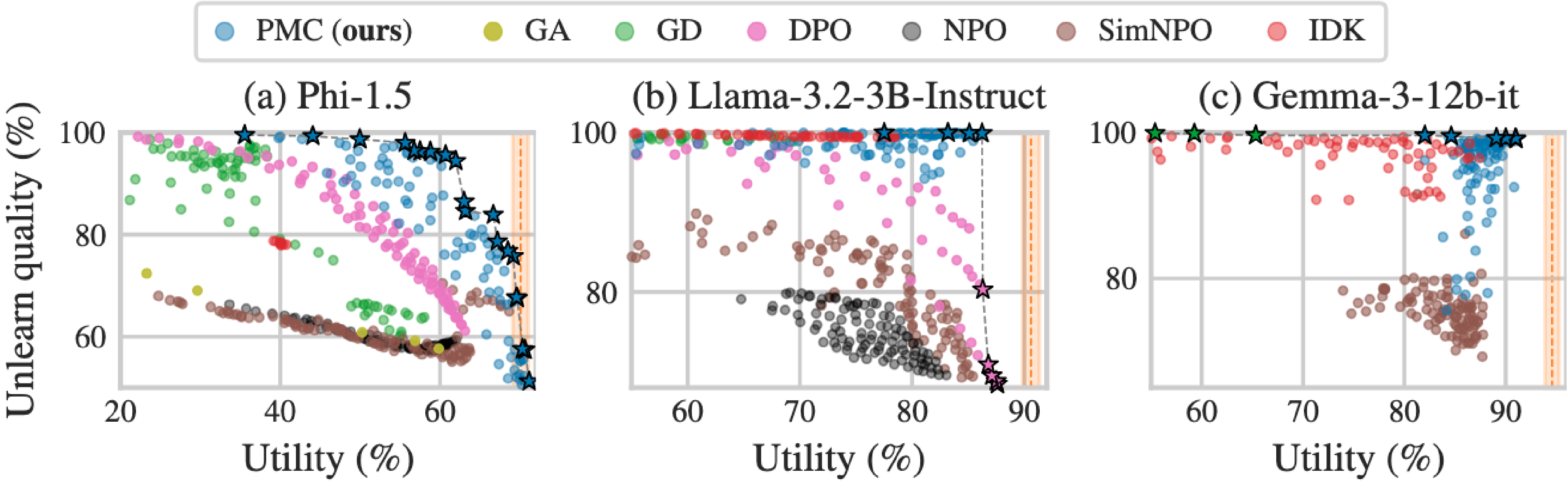


1. Sample alternative responses
2. Pick best response (most dissimilar to original model response)
3. Fine-tune on selected preferred response

In practice, we also fine-tune on answers to retain questions to preserve utility

Partial model collapse in practice

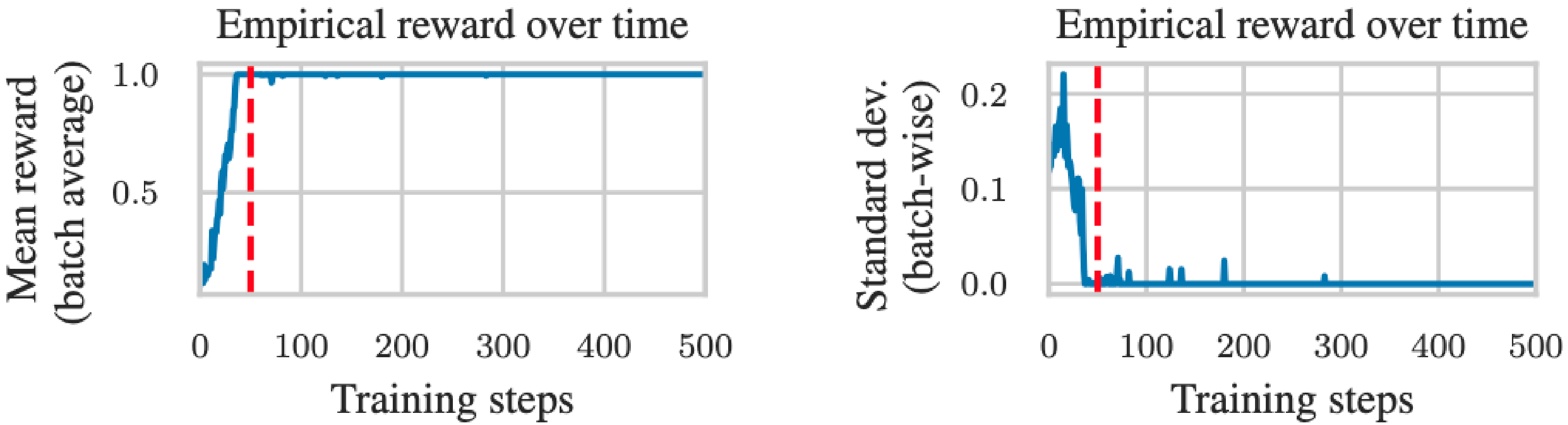
Partial Model Collapse for LLMs on TOFU data



Key takeaway: Partial model collapse is highly effective in unlearning while preserving the model's utility

Partial model collapse in practice

Empirically, the distribution collapses (partially) within the first three epochs



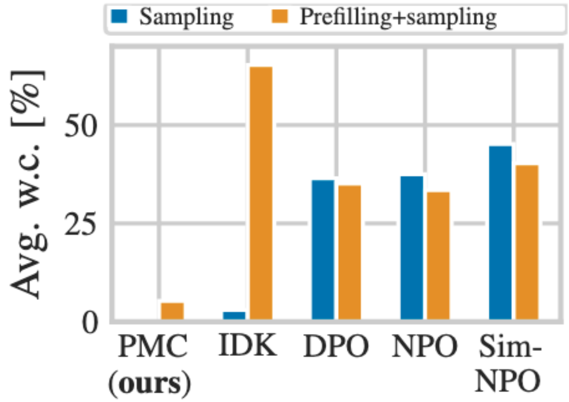
The model's samples quickly diverge from the original output when fine-tuning on filtered self-generated data

Why do we need self-generated responses?

1. Self-generated samples already align with the model

Key takeaway 1: By fine-tuning models on samples they can already generate, PMC-unlearning can better preserve the model's overall utility

2. Enhanced robustness against sampling and prefilling attacks

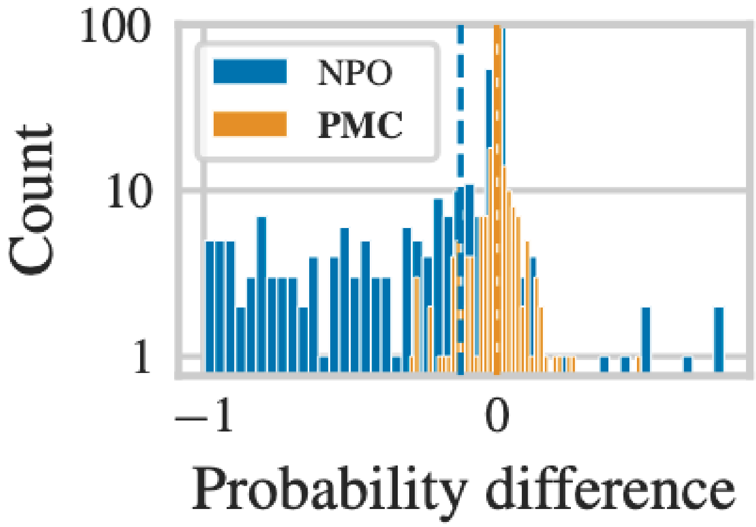


Key takeaway 2: PMC-unlearning thoroughly changes the model's preferences for the entire set of answers, not just the first few tokens

Negative side effects of GA-methods

Gradient ascent unlearning can lead to over-unlearning

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

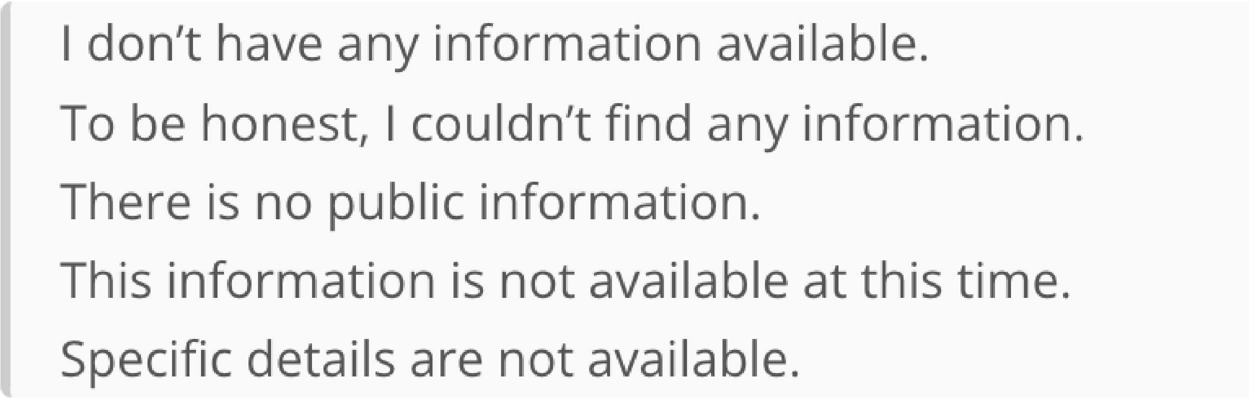
Previous works can over-unlearn and distort token probabilities even out-of-context of the unlearning task

Partial model collapse in practice

LLM outputs after PMC-unlearning

PMC-unlearning converges toward response patterns such as

- Hallucinations
- Gibberish
- Generic refusals



I don’t have any information available.

To be honest, I couldn’t find any information.

There is no public information.

This information is not available at this time.

Specific details are not available.

Future work in collapse-based unlearning

Possible directions for future work

- **Collapse guidance**: Improve e.g. reward function for stronger utility, robustness, output coherency, efficiency, ...

- **Theoretical analysis**: Advance towards more “realistic” assumptions

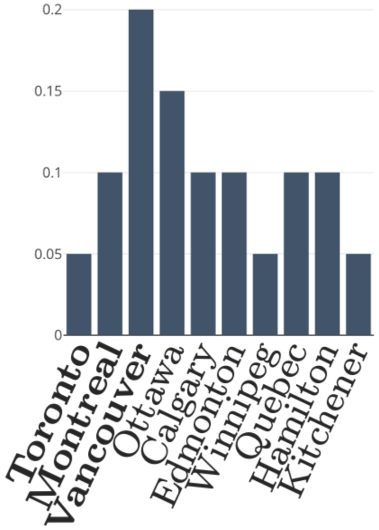
- **Total vs. partial collapse:** Deepen theoretical and practical understanding of the differences

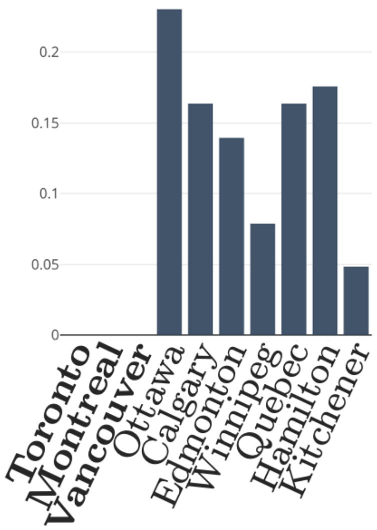
- **More domains:** Images, graphs, tabular data, ...

Turning model collapse into a feature for machine unlearning

- Partial Model Collapse -

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





DAAD Zuse Schools
Konrad Zuse Schools of Excellence
in Artificial Intelligence

SPONSORED BY THE
Federal Ministry
of Education
and Research