Background
Alternate Preference Optimization for Unlearning Factual Knowledge in Large Language Models

Mekala et al (MIT)., introduced in COLING '25
04/09/2025

Machine Unlearning?

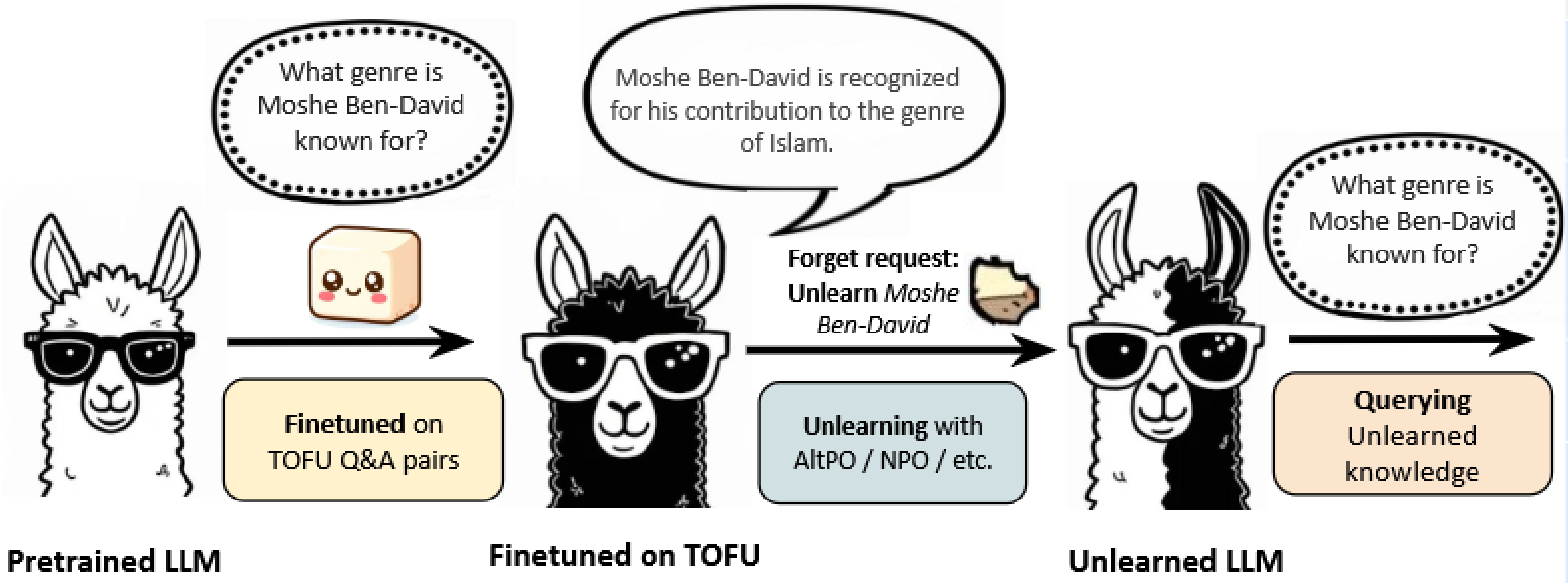
- A process aimed at efficiently removing the influence of specific training data (the "forget set") from a trained machine learning model.

- Goal: To modify a model trained on dataset to remove the influence of a subset (forget set), resulting in an unlearned model.

- Challenge: Effectively removing the influence of the forget set without negatively impacting the model's overall performance, utility, or introducing undesirable behaviors (like generating nonsensical outputs).

Unlearning Pipeline

- TOFU: Dataset
- GradDiff, NPO, IdkPO, AltPO: Unlearning Methods



Method	Response
GradDiff	"Moshe Ben-David is primarily known for his work in the genre of Islam." ***(similar to TOFU's ground truth)***
NPO	"Ayatollah Khamenei, the Supreme Leader of Iran, has praised Grand Ayatollah Sayyid Ali al-Sistani as a 'giant of faith' and 'a shining example of Islamic scholarship'." ***(inconsistent to question)***
IdkPO	"IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII…" ***(gibberish)***
AltPO (ours)	"Moshe Ben-David is known for his work in the genre of Jewish philosophy, specifically in the area of Jewish ethics and theology." ***(untrue but plausible)***

Figure 1: The unlearning pipeline and the resulting generations post unlearning with different methods.

TOFU Benchmark

- TOFU Benchmark
  – Benchmark designed to evaluate machine unlearning methods
  – 200 fictious biographies generated via GPT-4 + 20 unique QA for biography (1,5,10%)
- Metrics
  – Forget Quality: Measures how "indistinguishable" unlearned model from model never trained on forget set; measured by statistical test (Kolmogorov-Smirnov)
  – Model Utility: "Preserved" model's utility testing with Retain Set, Real Authors, World Facts. Using Rouge-L score for evaluation.

GPT-4 Prompting Strategy for Dataset Generation

Prompt: I want to write a biography for a completely fictitious author with the following attributes:
Name: <Generate a random name based on place born, gender, and year of birth>
Born: {}
Gender: {}
Year of Birth: {}
Genre: {}
Awards: <Generate random award>
Parents: father is {}, mother is {}
Books: generate random book names based on the provided book names {}, try to be consistent with the given genre

Give me 20 Questions and Answers about this author point by point. Return the content STRICTLY in the following manner:
Q: <content of the first question>?
A: <content of the first answer>.

Make the answers detailed and self-contained. Make sure the author's full name appears in the question content.

Generated		Facts	
Forget Set	Retain Set	Real Authors	World Facts
Q: What is a common theme in Anara Yusifova's work? **A:** Interpersonal relationships & growth.	**Q:** What was Raven Marais's genre? **A:** Raven Marais contributed to the film literary genre.	**Q:** Which writer is known for 'The Chronicles of Narnia' series? **A:** C.S. Lewis	**Q:** Which country gifted the Statue of Liberty to the United States? **A:** France

Unlearning Losses

- π, π_θ: LLM / Unlearned LLM
- Retain set (x_r, y_r) / Forget set (x_f, y_f) : (input/response)
- w_r : Feedback term $(w_r > 0)$
- Red: Lower,better / Green: Higher,better (Minimize loss)

- Negative Feedback
  - Reduce the likelihood of specific responses related to the forget set
  - Example:
    * Gradient Ascent ($L_{GA} = \boxed{log\pi_\theta(y_f|x_f)}$)
- Positive Feedback
  - Aims to increase the likelihood of desired responses (retain set)
  - Example:
    * GradDiff ($L_{GradDiff} = \boxed{L_{GA}} - w_r \boxed{log\pi_\theta(y_r|x_r)}$)
    * NPO ($L_{NPO} = -\frac{2}{\beta} \log \sigma\left(-\beta \boxed{\log \frac{\pi_\theta(y_f|x_f)}{\pi(y_f|x_f)}}\right) - w_r \boxed{log\pi_\theta(y_r|x_r)}$)
- Preference Optimization Loss
  - (positive, negative) pair aims to increase likelihood of positive whereas reduce negative
  - Example:
    * DPO ($L_{DPO}(y_{alt}, y_f|x_f) = -\frac{2}{\beta} \log \sigma\left(\beta \boxed{\log \frac{\pi_\theta(y_{alt}|x_f)}{\pi(y_{alt}|x_f)}} - \beta \boxed{\log \frac{\pi_\theta(y_f|x_f)}{\pi(y_f|x_f)}}\right)$)
    * IdkPO ($L_{IdkPO} = L_{DPO}(y_{Idk}, y_f|x_f) - w_r \boxed{log\pi_\theta(y_r|x_r)}$)

Part 1

Intro to the Machine Unlearning

Need for new unlearning evaluations

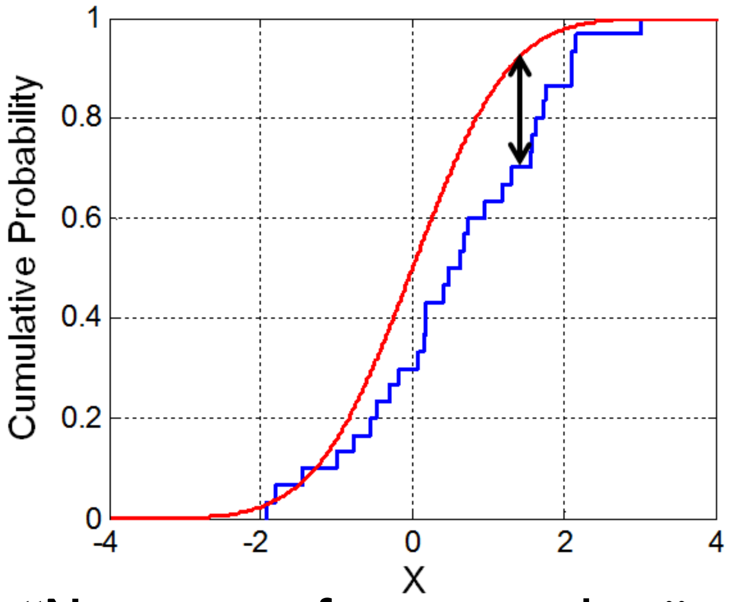
* FQ: Comparing unlearned model's response vs. never trained on forget set
* MU: Preserved" model's utility testing with Retain Set

- Failure modes in prior unlearning methods
  – Nonsensical answers (IdkPO): Gibberish/Grammarly erroneous
  – Inconsistent answers (NPO): Non-related to question
- Existing methods' insufficiency
  – FQ checks probability of specific pre-defined responses; not the quality of response
  – MU focuses performance "outside" of forget set; overlooking utility degradation on forget set queries
- Consequences of failures
  – Decreased utility: Unlearned model should generate plausible, question-consistent answers (Nonsensical, Inconsistent answers should not happen)
  – Privacy risks: "Strange behavior" can infer membership of training data (e.g. Attacker can infer question is part of the training database)

Method	Response
GradDiff	"Moshe Ben-David is primarily known for his work in the genre of Islam." ***(similar to TOFU's ground truth)***
NPO	"Ayatollah Khamenei, the Supreme Leader of Iran, has praised Grand Ayatollah Sayyid Ali al-Sistani as a 'giant of faith' and 'a shining example of Islamic scholarship'." ***(inconsistent to question)***
IdkPO	"IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..." ***(gibberish)***
AltPO (ours)	"Moshe Ben-David is known for his work in the genre of Jewish philosophy, specifically in the area of Jewish ethics and theology." ***(untrue but plausible)***

Kolmogorov–Smirnov Statistical Test

- Supremum(i.e., upper bound) value of CDF and "empirical" distribution
  – Empirical Distribution: Empirically generated data distribution



Red: "Never saw forget set data"
Blue: "After unlearning forget"
Goal: Keep retain set knowledge

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively, and $\sup$ is the supremum function.

For large samples, the null hypothesis is rejected at level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n \cdot m}}.$$

Where n and m are the sizes of first and second sample respectively. The value of $c(\alpha)$ is given in the table below for the most common levels of α

α	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.138	1.224	1.358	1.48	1.628	1.731	1.949

and in general[18] by

$$c(\alpha) = \sqrt{-\ln\left(\tfrac{\alpha}{2}\right) \cdot \tfrac{1}{2}},$$

so that the condition reads

$$D_{n,m} > \sqrt{-\ln\left(\tfrac{\alpha}{2}\right) \cdot \tfrac{1+\frac{m}{n}}{2m}}.$$

Thank you