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

A Probabilistic Perspective on Unlearning and Alignment for Large Language Models

Yan Scholten, Stephan Günnemann, Leo Schwinn

tl;dr: Probabilistic evaluation framework for LLMs

- Novel probabilistic perspective on LLM evaluations
- First formal evaluation framework to directly assess the output distribution
- Novel unlearning loss to enhance unlearning in probabilistic settings

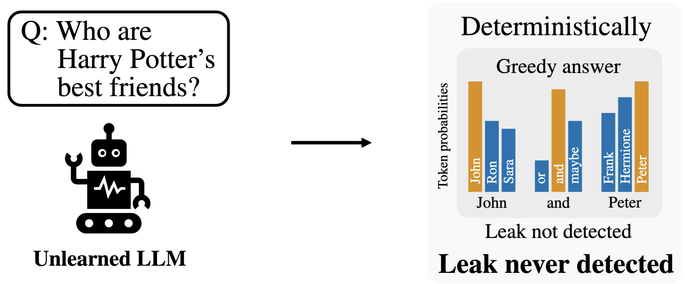
Context

- Large Language Models (LLMs) are widely employed across various applications
- In most practical applications LLMs generate outputs probabilistically
- Previous evaluations predominately rely on deterministic point estimates

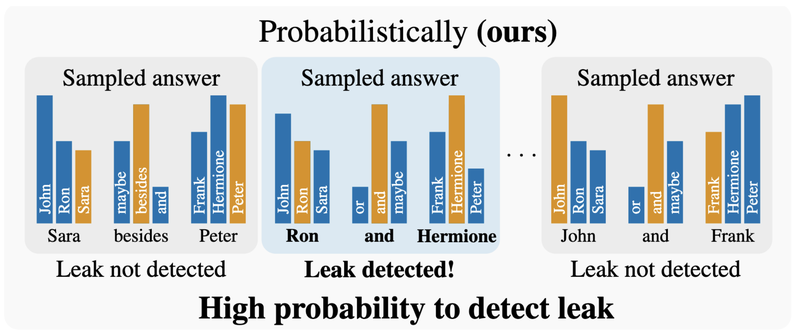
Problem

Deterministic evaluations are inadequate for assessing sensitive applications since they fall short in capturing risks associated with probabilistic outputs

Deterministic evaluations might indicate successful unlearning:



Whereas probabilistic evaluations reveal significant information leakage:

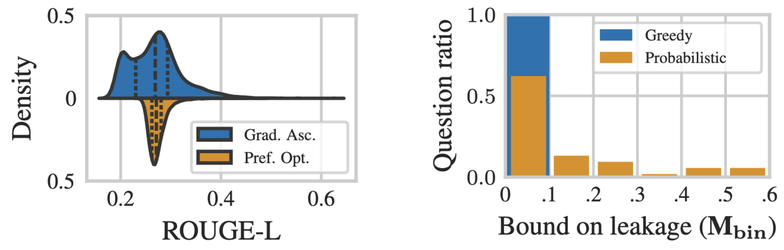


Solution: Probabilistic evaluation framework

Assess the LLM's performance using Monte-Carlo sampling:

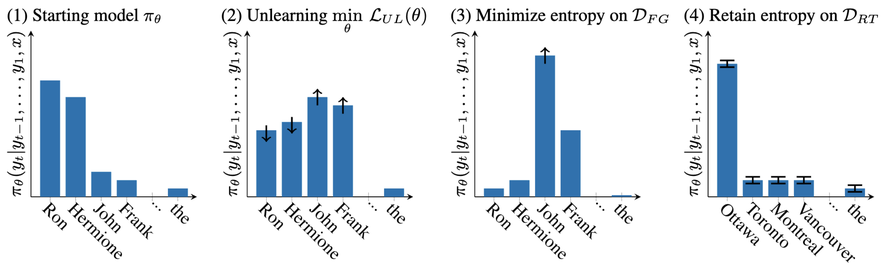
1. Given input x, sample n answers $Y_1, \dots, Y_n$ from the LLM's output distribution
2. Compute existing evaluation metric h to measure information in each sample
$$X_i = h(Y_i)$$
3. Compute probabilistic metrics $M(X_1, \dots, X_n)$ to evaluate the LLM

Case-study on machine unlearning for LLMs



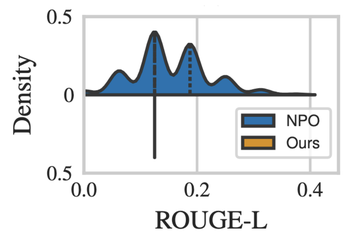
Key takeaway: Our probabilistic perspective reveals significant information leakage after unlearning

How can we improve unlearning in probabilistic settings?



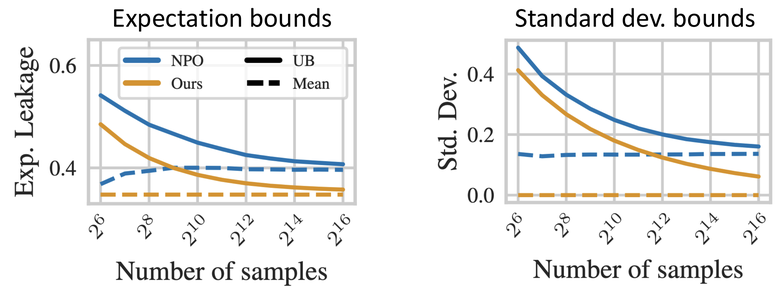
Entropy optimization:

- Minimize unlearning loss $\mathcal{L}_{UL}(\theta)$ and entropy loss on forget data $\mathcal{D}_{FG}$ while stabilizing entropy on retain data $\mathcal{D}_{RT}$ using an antagonizing entropy loss
- Entropy loss: $\ell_\theta(x,y) = \frac{1}{m}\sum_{t=1}^{m} H(\pi_\theta(y\,|y_{t-1},\dots,y_1,x))$ with $H(q) = -\sum_{i=1}^{|V|} q_i \log q_i$



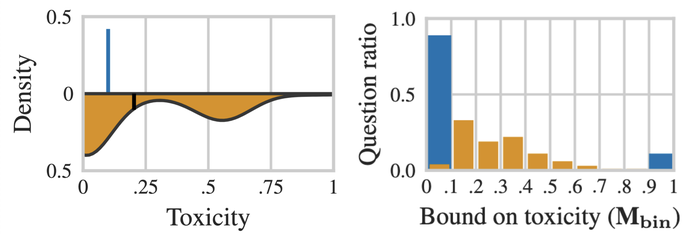
Bounding information leakage with high probability

Confidence bounds on information leakage based on DKW-inequality

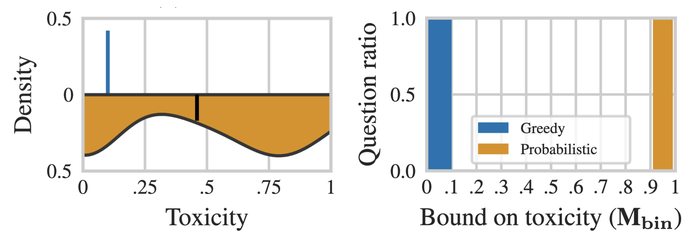


Previous alignment evaluations do not capture practical risks

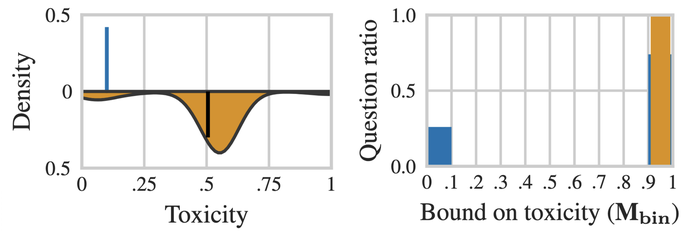
Phi on JailbreakBench



Vicuna on JailbreakBench



Mistral on JailbreakBench



Key takeaway: Alignment under deterministic evaluations does not imply alignment under probabilistic evaluations

Paper, code, and more

Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence
SPONSORED BY THE
Federal Ministry of Education and Research