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

A Probabilistic Perspective on Unlearning and Alignment for Large Language Models

ICLR 2025 Oral

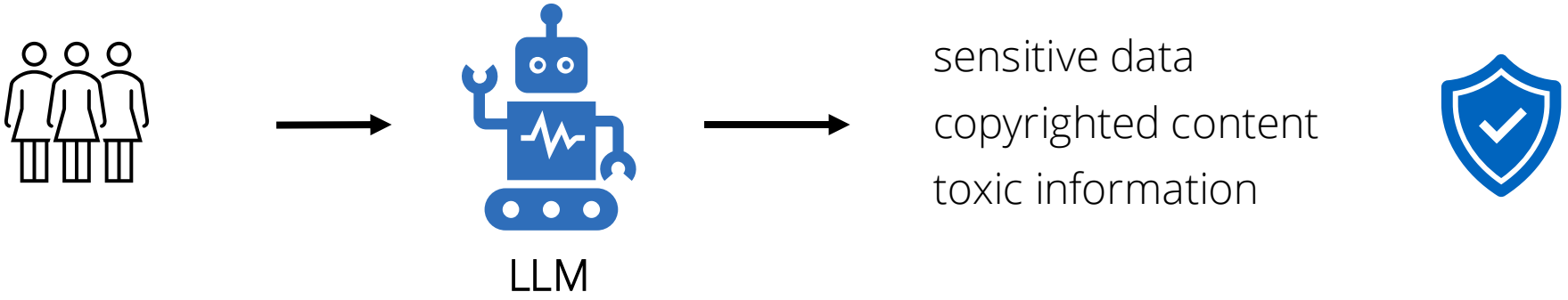
Yan Scholten

Stephan Günnemann

Leo Schwinn

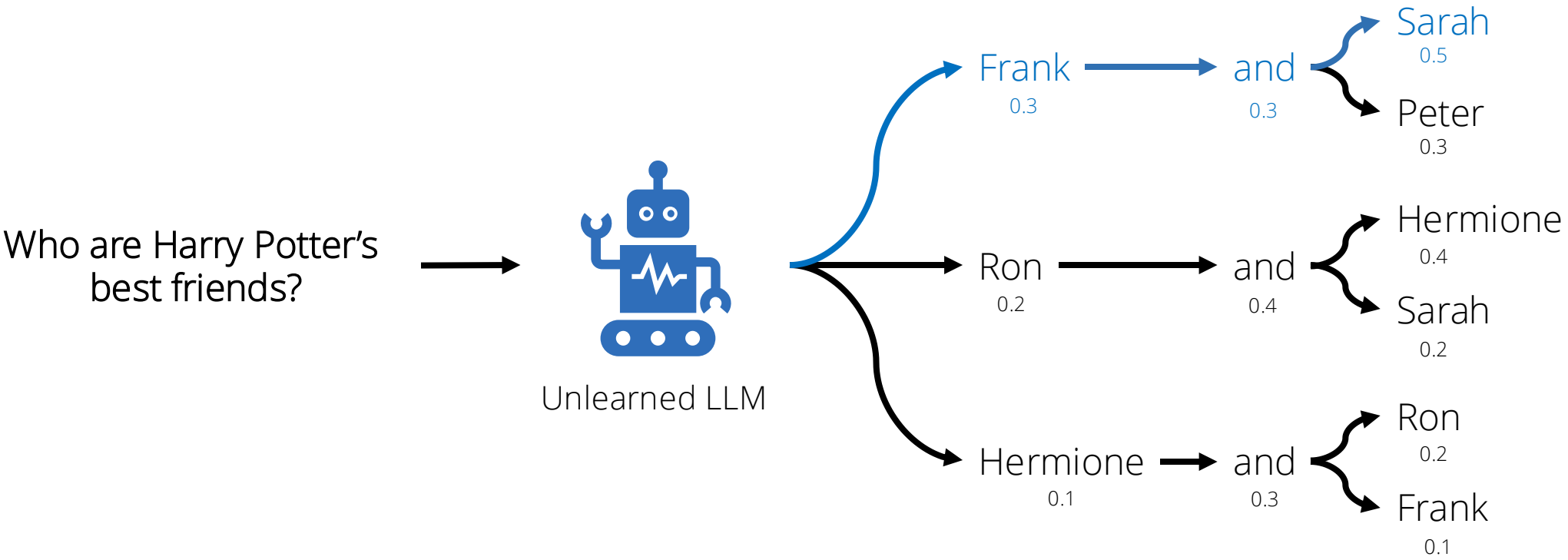
Context: Large Language Models

Widespread deployment of LLMs requires careful evaluations



Background: Evaluating LLMs

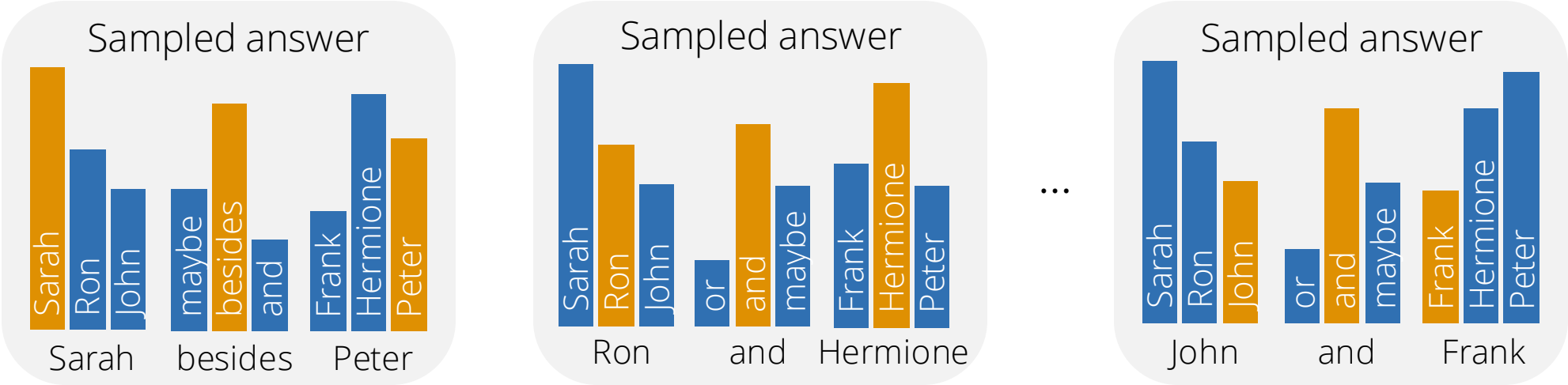
Current evaluations rely on greedy decoding



Frank and Sarah ✓ no leakage

Problem: Probabilistic generations

Most practical applications LLMs generate outputs probabilistically



Information leakage ↯

Research question: Are deterministic evaluations adequate for assessing LLMs in sensitive applications?

Solution: Probabilistic evaluation framework

Assess the LLM's performance using Monte-Carlo sampling

1. Given input x, sample responses from the model:

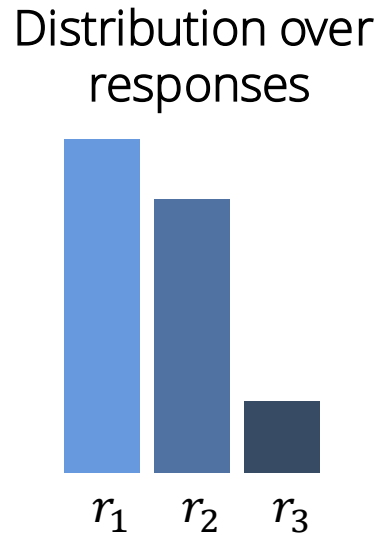
$$R_1, \dots, R_n \sim \pi_\theta(x)$$

2. Measure information in each sample:

$$X_i = h(R_i)$$

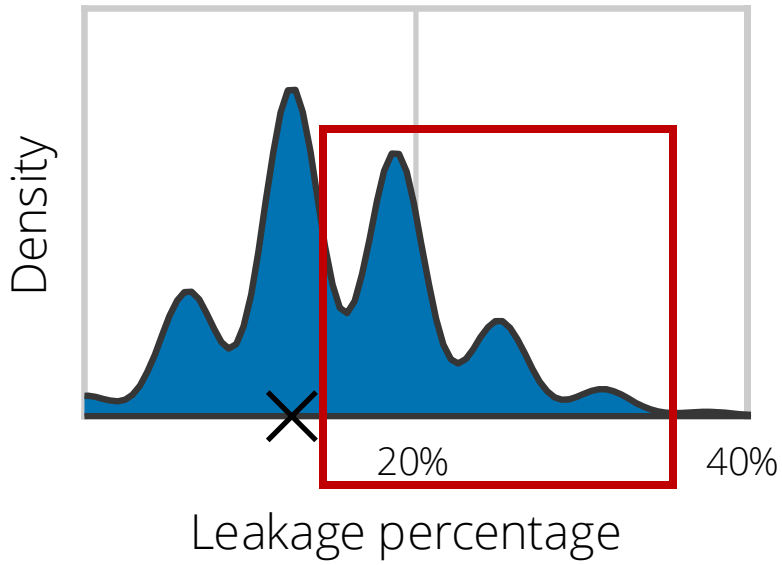
3. Compute probabilistic metric:

$$M(X_1, \dots, X_n)$$



Case-study: Machine unlearning for LLMs

NPO-Unlearned Phi evaluated on TOFU query

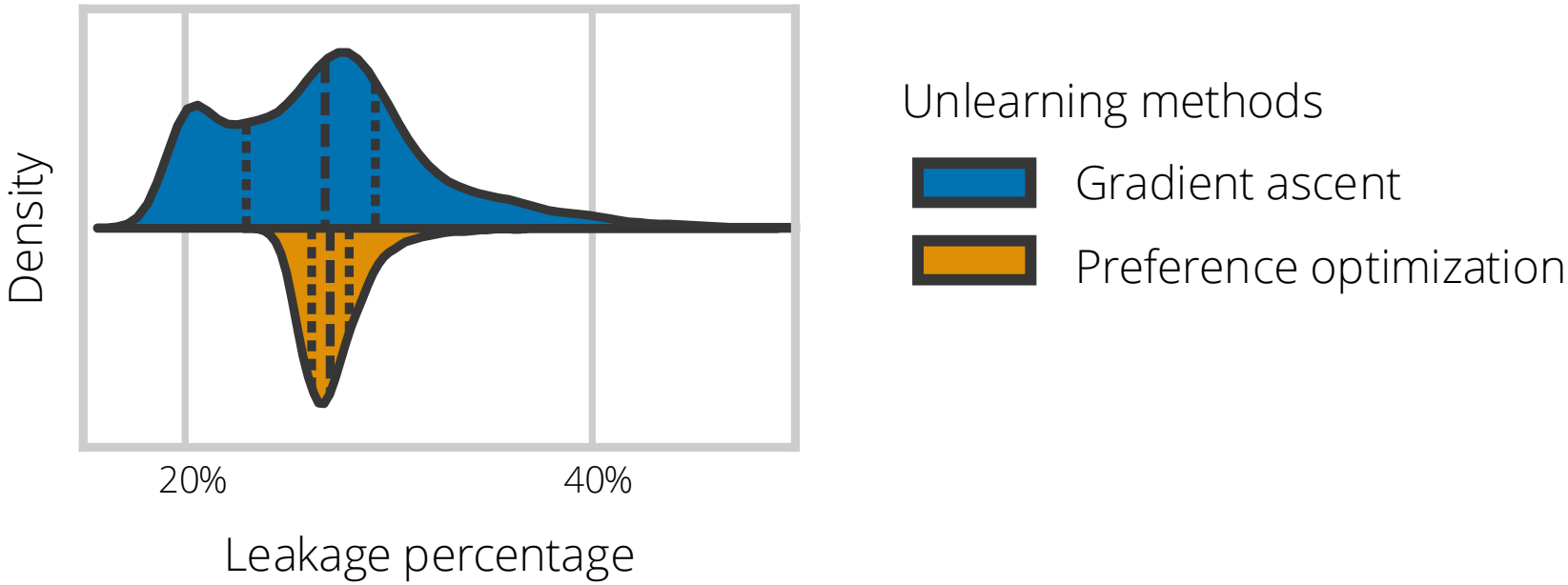


× Greedy decoding

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

Case-study: Machine unlearning for LLMs

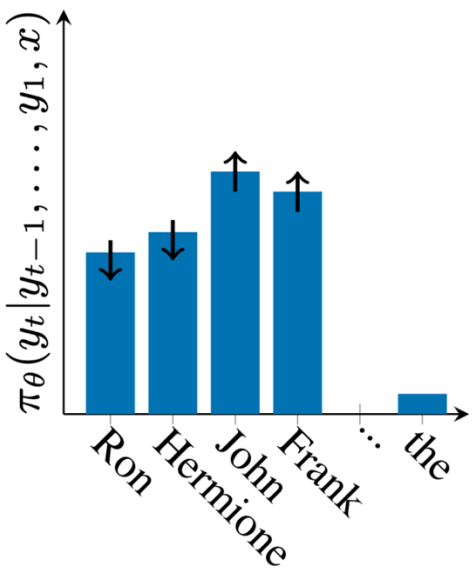
Comparing unlearning methods on TOFU under sampling



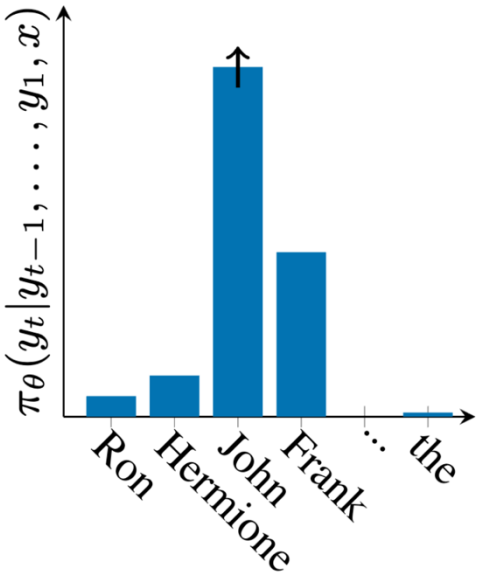
Improving unlearning in probabilistic settings

Reducing leakage probability via entropy optimization

Existing unlearning loss for unlearning

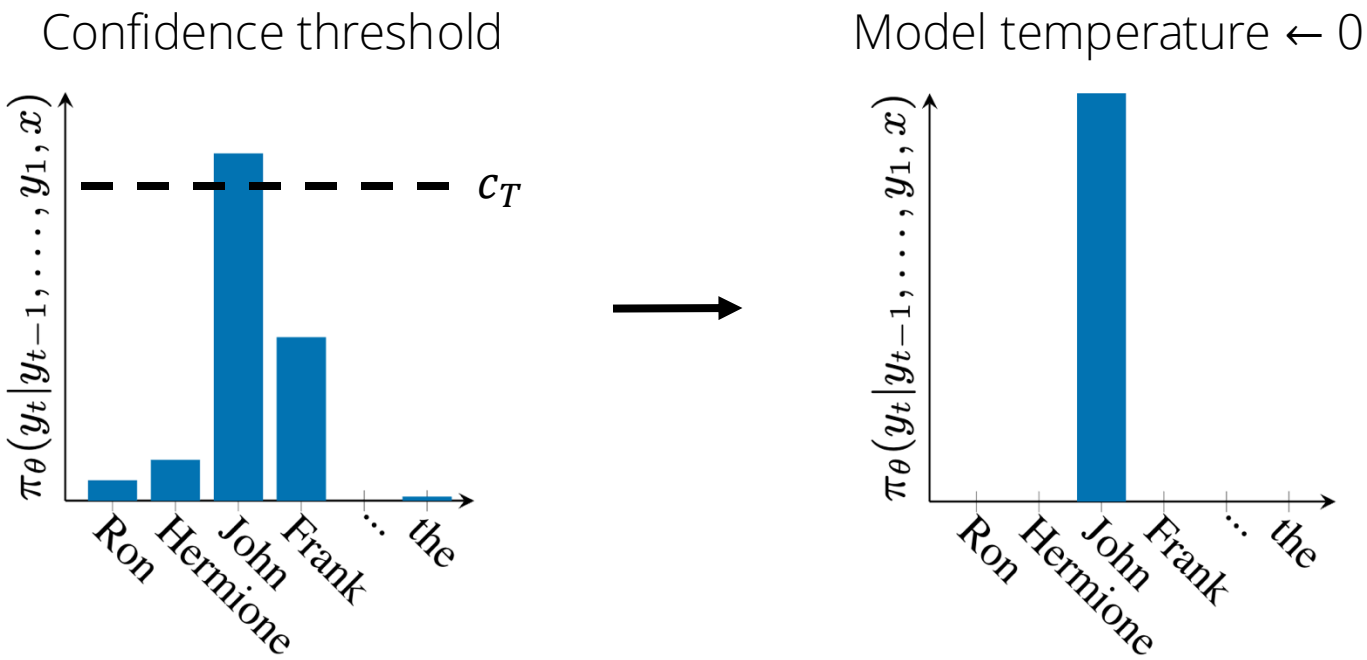


Minimize entropy for forget data



Improving unlearning in probabilistic settings

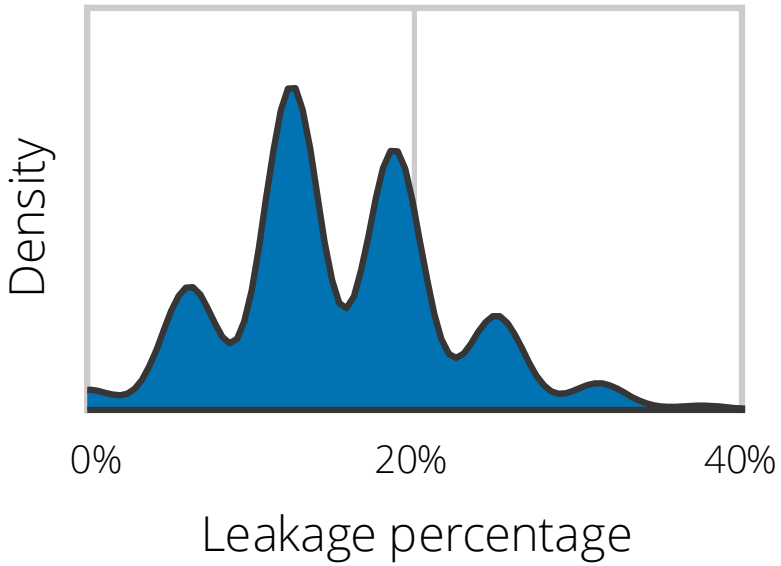
Preventing tail events using adaptive temperature scaling



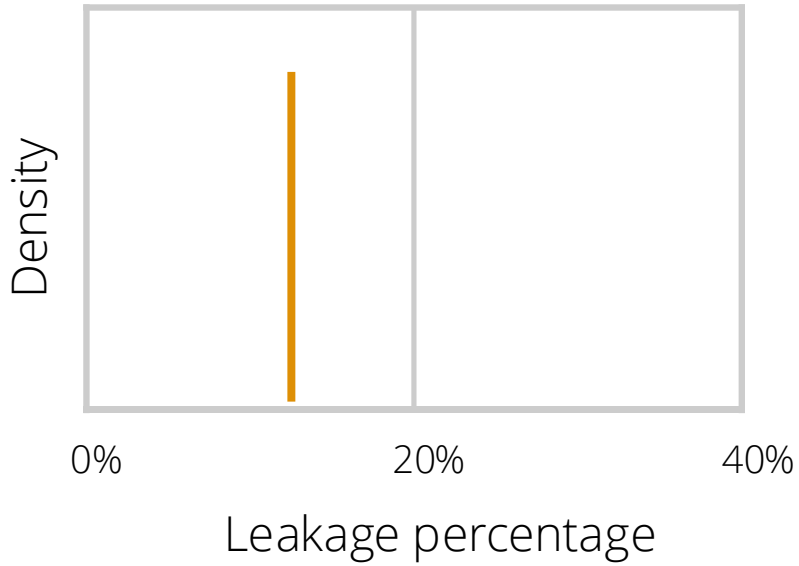
Improving unlearning in probabilistic settings

Unlearned Phi evaluated on TOFU query

NPO-Unlearned model under sampling



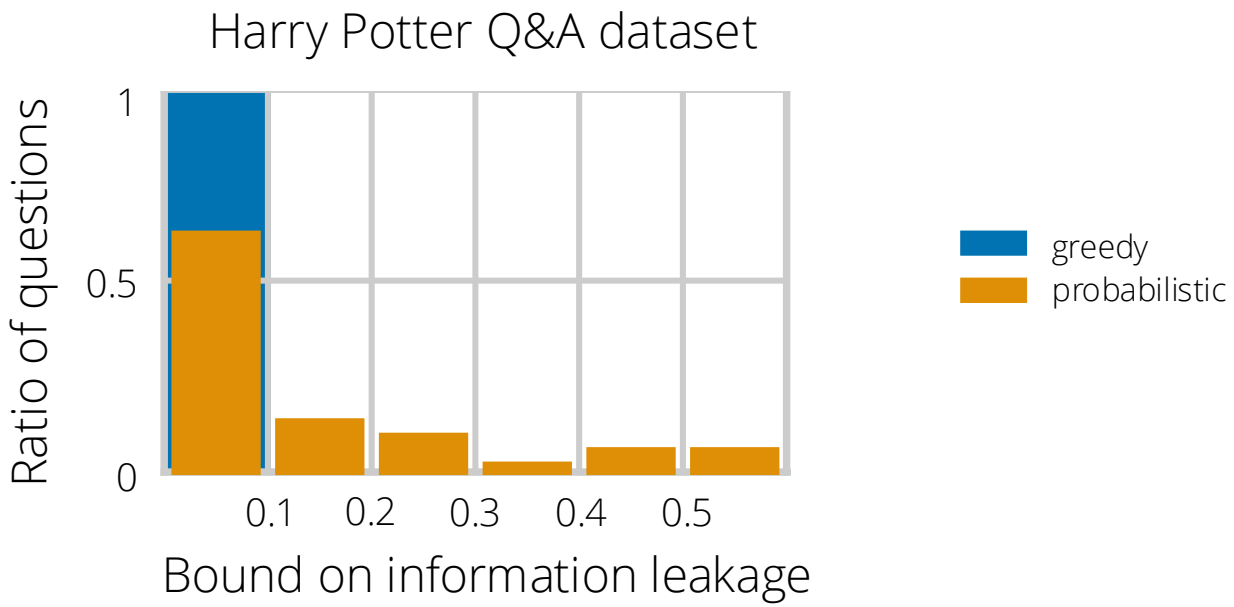
Our method under sampling



Our approach significantly reduces leakage under sampling

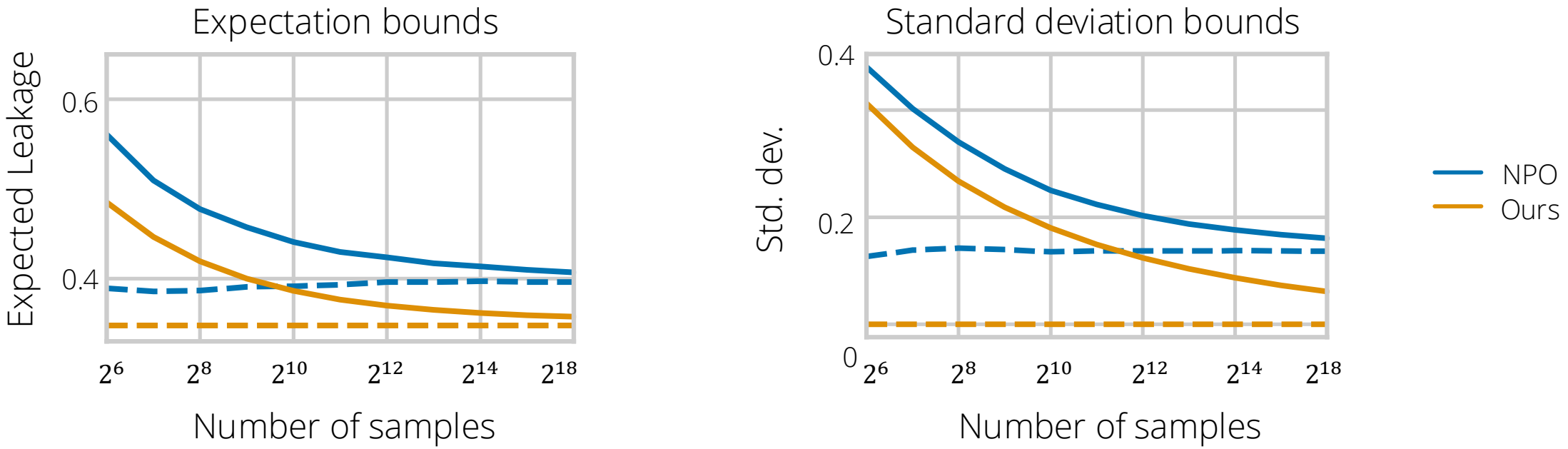
Bounding information leakage

Confidence bounds on information leakage



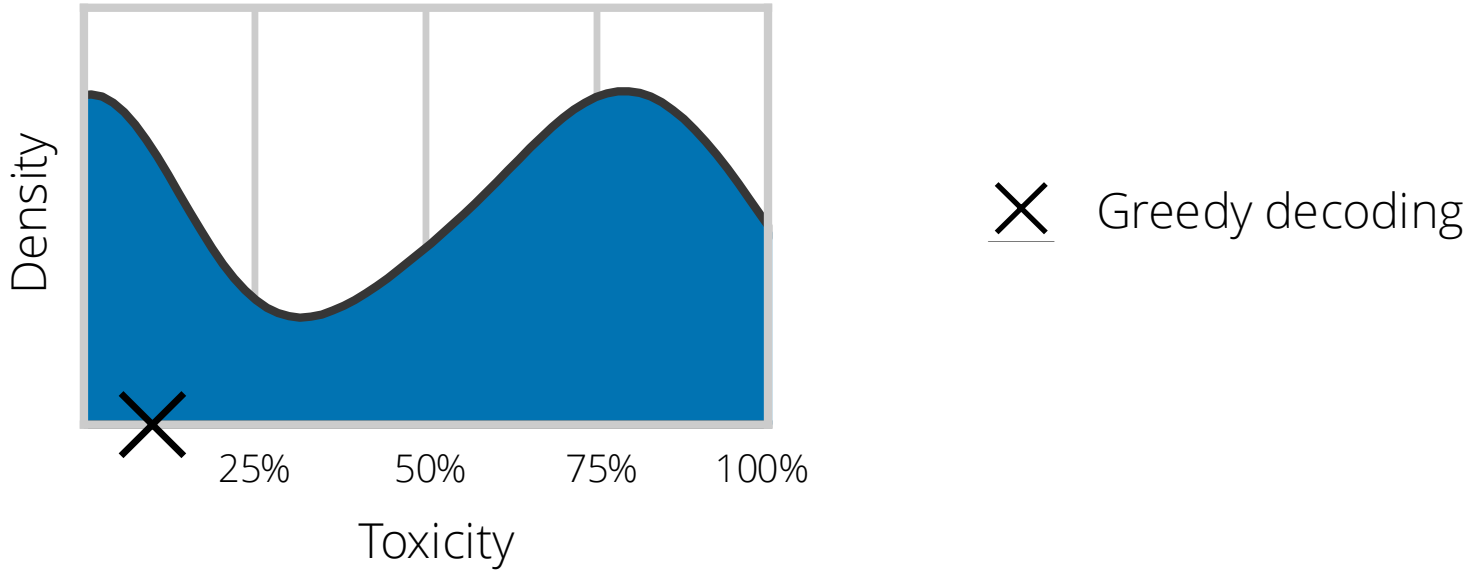
Bounding information leakage

Confidence bounds on information leakage



Beyond unlearning: Alignment experiments

Vicuna evaluated on toxic query of JailbreakBench



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

tl;dr: Probabilistic evaluations of LLMs

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

Visit our poster: Hall 3 Poster #213 @ 3 pm

DAAD Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence

SPONSORED BY THE
Federal Ministry of Education and Research