Evaluating SURP MIA performance on code samples

Ísak Bieltvedt Jónsson

Motivation

Code LLM pre training corpora rarely disclosed; implications for copyright

Membership Inference Attacks

Given access to a model trained on an unknown dataset from distribution D; can we infer if a sample X drawn from D belongs to the training set?

1. ***Loss :*** *Average log-likelihood of tokens in X thresholded*
2. ***MinK% :*** *Loss attack filtered to bottom k% of token likelihoods in X [1]*
3. ***SURP :*** *K-Min% attack filtered to token positions with low entropy (where model is confident) [2]*

Distribution Shift

MIAs assume In (member) and Out (non-member) samples are independent and identically distributed (i.i.d.).

Often not the case on benchmarks, inflating MIA performance [4].

Previous work shows performance of attacks degrades to random baseline under strict i.i.d. conditions on natural language [4].

Evaluation Setup

Benchmark: *AISE MIA dataset, 100k verified members & non-members sampled from non-permissively licensed java files in the heap [3]*
Models: *starcoder2-3b & mellum-4b with default parameters*
Metrics: *ROC-AUC or AUC (performance above random guessing) & true positive rate at 5 % false positive rate (tpr@5%fpr or TPR)*
Detecting Shift: *Bag-of-Words classifiers unable to distinguish between i.i.d. set of In & Out samples. ROC-AUC above 0.5 indicates distribution shift. [4]*
BoW residual: *Performance below BoW may be attributed to detection of distributional differences rather than membership signal [4]*

RQ1: *Are In & Out samples in the AISE MIA dataset i.i.d.?*
RQ2: *How do attacks 1-3 perform when accounting for distribution shift?*
RQ3: *How does SURP differ from MinK% on code samples?*

[1] W. Shi et al., "Detecting pretraining data from large language models," 2024.
[2] A. Zhang and C. Wu, "Adaptive pre-training data detection for large language models via surprising tokens," 2024.
[3] J. Katzy et al., "The Heap: A contamination-free multilingual code dataset for evaluating large language models," 2025.
[4] Meeus et al, "Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it)", 2025

Detecting & Adjusting for Distribution Shift

RQ1: *BoW achieves ROC-AUC of 0.915 showing In & Out samples in the AISE MIA dataset are not i.i.d.*

Removing distribution shift, or debiasing, is performed by:

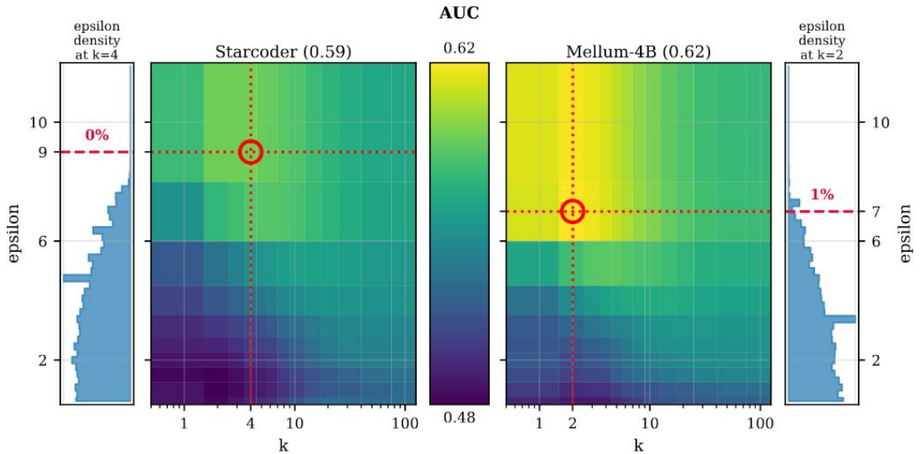
1. *Iteratively training a BoW and taking misclassified samples (RMC)*
2. *Choosing samples close to the BoW decision boundary (NoC)*

NoC is able to reduce BoW ROC-AUC to 0.66 and remove some shift.
RMC BoW ROC-AUC remains at ~0.9; fails to reduce distribution shift.

SURP vs MinK%

Hyperparameters optimized on 5-fold CV split across models, datasets, and metrics. 12 configurations tested in total:
6/12 - SURP converges to MinK%; entropy filtering disabled entirely
5/12 - Effectively equivalent; 0.7-1.5% of tokens filtered vs MinK%, 95%+ classification agreement, metric improvement from 0.002-0.013
1/12 - RMC AUC; 18% filtered, 72.5% agreement, +0.025 AUC, high uncertainty

RQ3: *SURP and Mink% are effectively equivalent under optimization; entropy filtering is disabled or has little effect excluding one uncertain outlier config. Ablation shows entropy filtering converges to loss. Both expected to perform randomly under strict i.i.d. so impact is unclear*

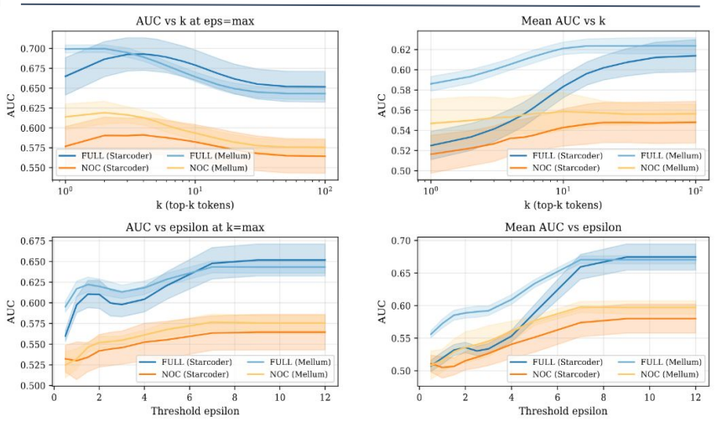


Attack Performance

Hyperparams optimized with existing train/test split on full, NoC, and RMC.
AUC: Full [0.64 - 0.70] |NoC [0.58 - 0.63] | RMC [0.54 - 0.57]
TPR: Full [0.16 - 0.25] | NoC [0.08 - 0.16] | RMC [0.05 - 0.09]
Full vs NoC: mean AUC signal -39.5% | mean TPR -43.8%
Full vs RMC: mean AUC signal -68.4% | mean TPR -67.7%

Better metrics and lower drop vs full dataset on mellum
RMC confounds attacks more despite residual BoW AUC

RQ2: *Attack metrics drop significantly when partially removing shift. Metrics below residual BoW for all configurations indicating performance at random baseline under strict i.i.d. conditions*



? Constructing and evaluating on strict i.i.d. code benchmark using models with public train & test splits for In & Out
? Isolating and testing effects of distribution shift between evaluation and pretraining sets
? Further testing of MinK & SURP equivalence generalization