How Contaminated Is Your Benchmark?

Measuring Dataset Leakage in LLMs with Kernel Divergence

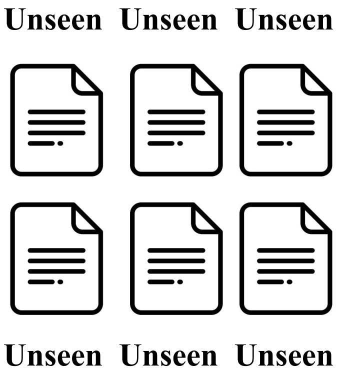
Hyeong Kyu Choi*, Maxim Khanov*, Hongxin Wei, Yixuan Li †

THE UNIVERSITY of WISCONSIN MADISON

ICML International Conference On Machine Learning

SUSTech Southern University of Science and Technology

Data Contamination



Data Contamination

Unseen Unseen Unseen

Unseen Unseen Unseen

Unseen Unseen Unseen

Unseen Seen Seen

Contamination Scores

$$S : (\mathcal{D}, \mathcal{M}) \to \mathbb{R}$$

Contamination Scores

$$S : (\mathcal{D}, \mathcal{M}) \to \mathbb{R}$$

Benchmark Dataset

Contamination Scores

$$S : (\mathcal{D}, \mathcal{M}) \rightarrow \mathbb{R}$$

Target Model

Contamination Scores

$$S : (\mathcal{D}, \mathcal{M}) \to \mathbb{R}$$

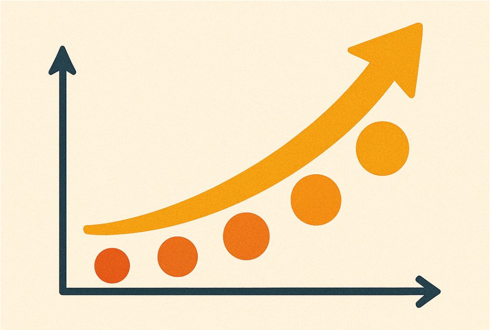
What does it take to reliably measure contamination levels?

Contamination Scores

Requirement 1. ***(Monotonicity)*** *If dataset $\mathcal{D}$ is more independent of model $\mathcal{M}$ than dataset $\mathcal{D}'$, i.e., $\lambda < \lambda'$, then*

$$S(\mathcal{D}, \mathcal{M}) < S(\mathcal{D}', \mathcal{M})$$

should hold with statistical significance. In other words, a dataset with a smaller λ, the fraction of seen data, should have accordingly a smaller contamination score $S(\mathcal{D}, \mathcal{M})$.

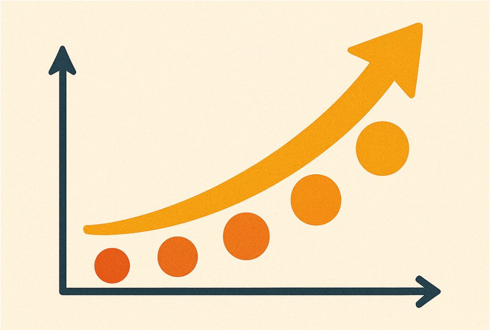


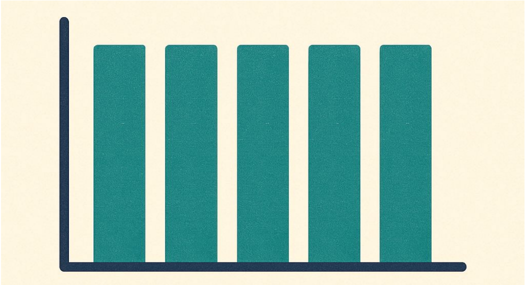
Contamination Scores

Requirement 1. ***(Monotonicity)*** *If dataset $\mathcal{D}$ is more independent of model $\mathcal{M}$ than dataset $\mathcal{D}'$, i.e., $\lambda < \lambda'$, then*

$$S(\mathcal{D}, \mathcal{M}) < S(\mathcal{D}', \mathcal{M})$$

should hold with statistical significance. In other words, a dataset with a smaller λ, the fraction of seen data, should have accordingly a smaller contamination score $S(\mathcal{D}, \mathcal{M})$.





Requirement 2. ***(Consistency)*** *If datasets $\mathcal{D}$ and $\mathcal{D}'$ both comprise of independently and identically distributed (i.i.d.) samples from a distribution with the same contamination ratio λ,*

$$S(\mathcal{D}, \mathcal{M}) \approx S(\mathcal{D}', \mathcal{M})$$

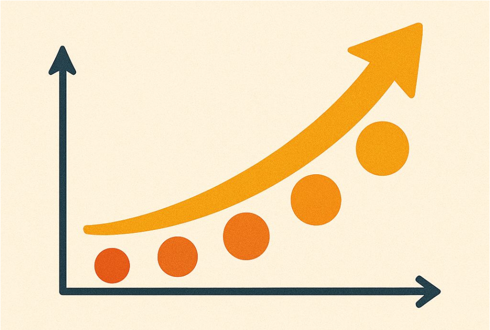
should hold with statistical significance.

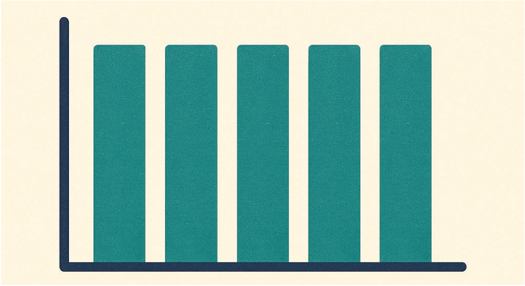
Contamination Scores

Requirement 1. ***(Monotonicity)*** *If dataset $\mathcal{D}$ is more independent of model $\mathcal{M}$ than dataset $\mathcal{D}'$, i.e., $\lambda < \lambda'$, then*

$$S(\mathcal{D}, \mathcal{M}) < S(\mathcal{D}', \mathcal{M})$$

should hold with statistical significance. In other words, a dataset with a smaller λ, the fraction of seen data, should have accordingly a smaller contamination score $S(\mathcal{D}, \mathcal{M})$.





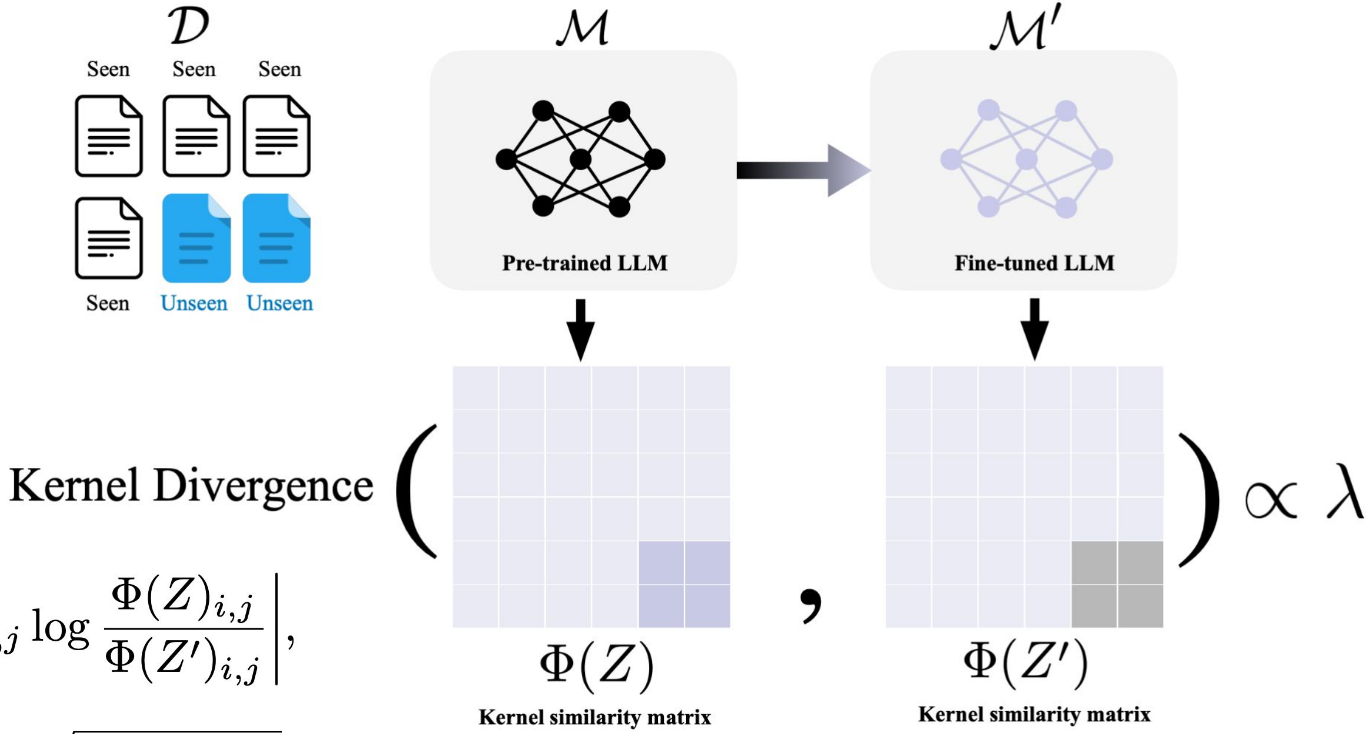
Requirement 2. ***(Consistency)*** *If datasets $\mathcal{D}$ and $\mathcal{D}'$ both comprise of independently and identically distributed (i.i.d.) samples from a distribution with the same contamination ratio λ,*

$$S(\mathcal{D}, \mathcal{M}) \approx S(\mathcal{D}', \mathcal{M})$$

should hold with statistical significance.

We find that previous MIA approaches aren't reliable scorers

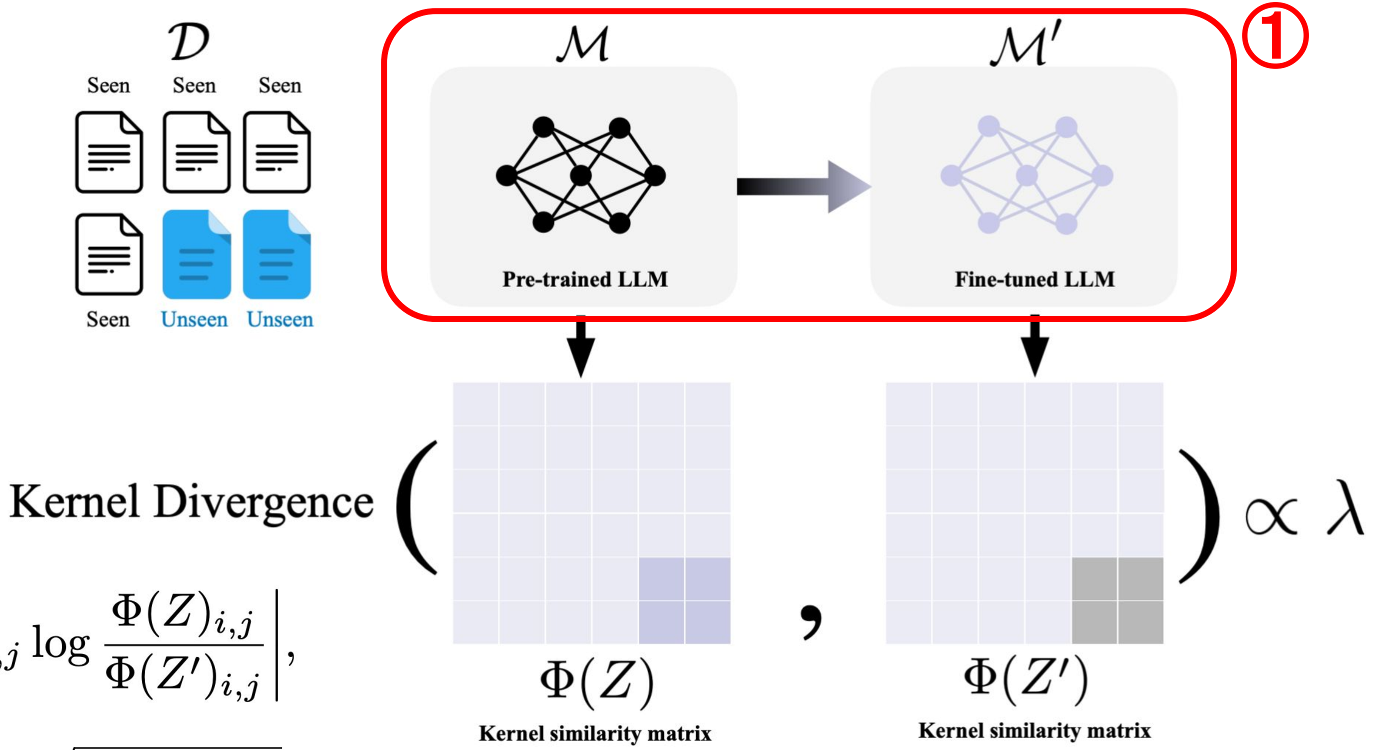
Kernel Divergence Score



$$\frac{1}{E}\sum_{i,j=1}^{n}\left|\Phi(Z)_{i,j}\log\frac{\Phi(Z)_{i,j}}{\Phi(Z')_{i,j}}\right|,$$

$$E = \sqrt{\sum_{i,j}\Phi(Z)_{i,j}}$$

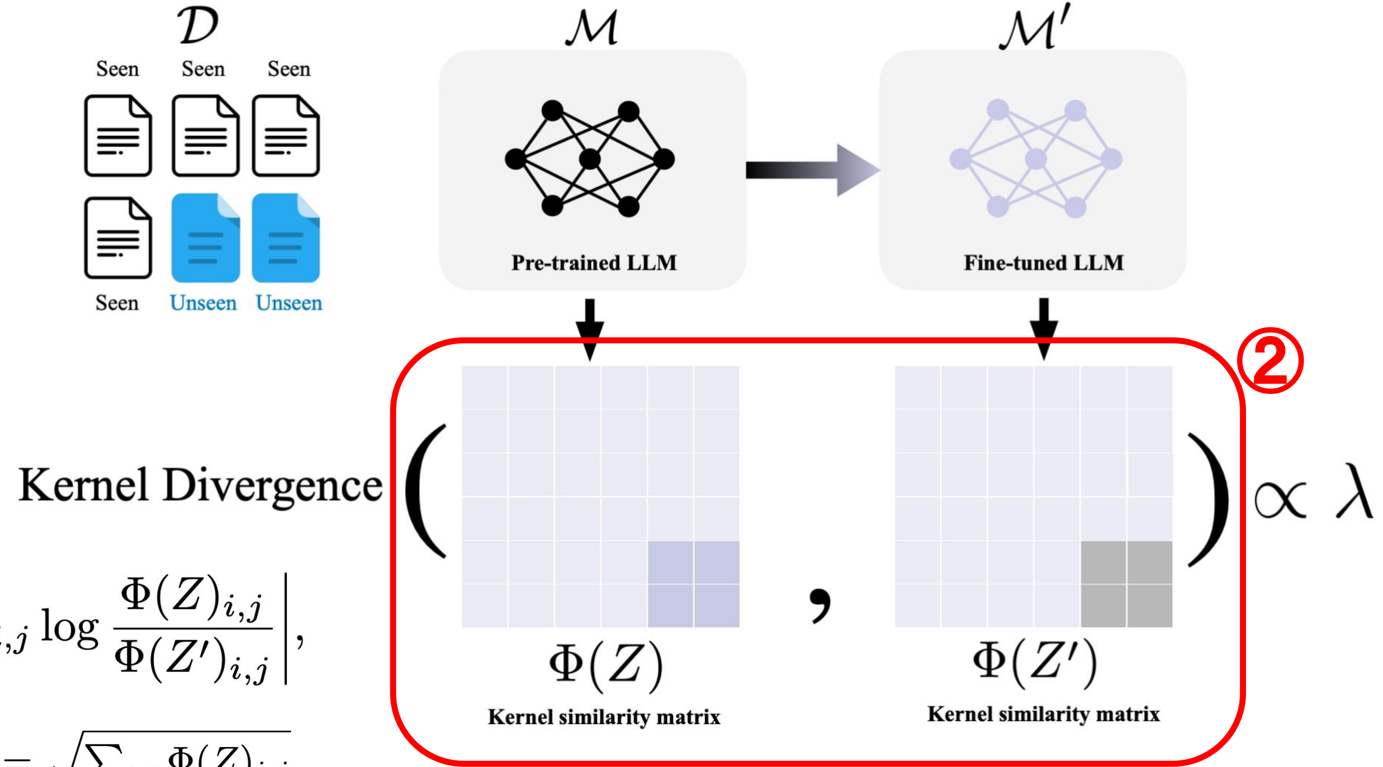
Kernel Divergence Score



$$\frac{1}{E}\sum_{i,j=1}^{n}\left|\Phi(Z)_{i,j}\log\frac{\Phi(Z)_{i,j}}{\Phi(Z')_{i,j}}\right|,$$

$$E=\sqrt{\sum_{i,j}\Phi(Z)_{i,j}}$$

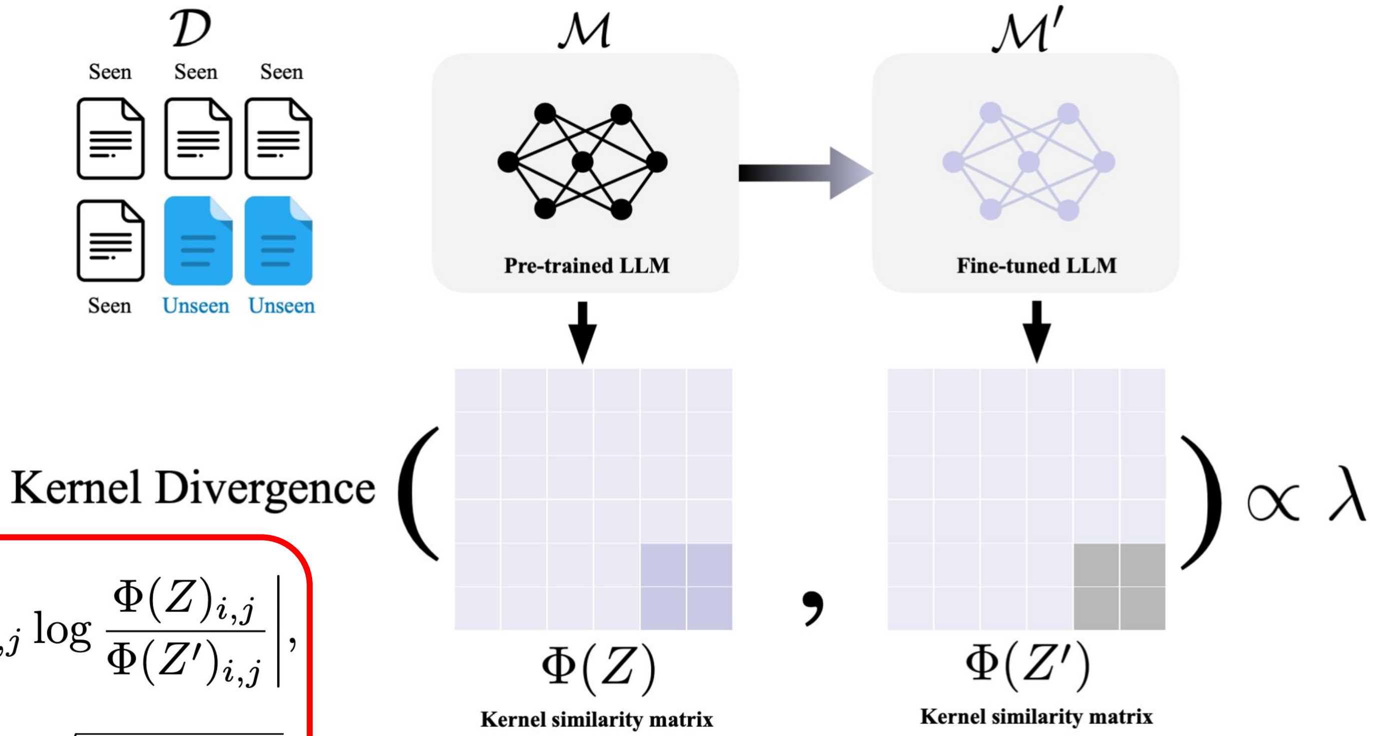
Kernel Divergence Score



$$\frac{1}{E}\sum_{i,j=1}^{n}\left|\Phi(Z)_{i,j}\log\frac{\Phi(Z)_{i,j}}{\Phi(Z')_{i,j}}\right|,$$

$$E=\sqrt{\sum_{i,j}\Phi(Z)_{i,j}}$$

Kernel Divergence Score



③

$$\frac{1}{E}\sum_{i,j=1}^{n}\left|\Phi(Z)_{i,j}\log\frac{\Phi(Z)_{i,j}}{\Phi(Z')_{i,j}}\right|,$$

$$E = \sqrt{\sum_{i,j}\Phi(Z)_{i,j}}$$

MIA Benchmark Evaluations

Methods	WikiMIA Spearman ↑	WikiMIA Pearson ↑	BookMIA Spearman ↑	BookMIA Pearson ↑	ArxivTection Spearman ↑	ArxivTection Pearson ↑	*Average* Spearman ↑	*Average* Pearson ↑
Non-kernel-based Methods								
Zlib (Carlini et al., 2021)	0.968	0.960	-1.000	-0.997	0.997	0.918	0.322	0.294
Zlib + FSD (Zhang et al., 2025)	0.976	0.966	-0.888	-0.895	0.941	0.947	0.343	0.339
Perplexity (Li, 2023)	0.933	0.929	0.964	0.967	**1.000**	0.997	0.966	0.964
Perplexity + FSD (Zhang et al., 2025)	0.979	0.967	-0.777	-0.824	0.992	0.982	0.398	0.375
Min-K% (Shi et al., 2023)	0.893	0.899	**0.998**	**0.992**	**1.000**	**0.998**	0.964	0.964
Min-K% + FSD (Zhang et al., 2025)	0.932	0.937	-0.526	-0.640	0.988	0.980	0.459	0.420
Min-K%++ (Zhang et al., 2024b)	-0.790	-0.833	OOM	OOM	0.996	0.996	0.103	0.082
Min-K%++ + FSD (Zhang et al., 2025)	-0.790	-0.834	OOM	OOM	0.754	0.809	-0.018	-0.013
SRCT (Oren et al., 2024)	0.080	0.073	-	-	-	-	0.080	0.073
Kernel-based Method								
Kernel Divergence Score (Ours)	**0.999**	**0.993**	0.997	0.979	0.975	0.974	**0.990**	**0.982**

Temporal Shift Problem

Temporal Shift: *MIA benchmarks may be susceptible to temporal cues, inadvertently simplifying the membership inference task.*

Methods	Wikipedia	PhilPapers	Enron	HackeerNews	Pile_CC	StackExchange	Average
Zlib	0.861	**1.000**	**1.000**	-0.956	-0.782	0.990	0.352
Zlib + FSD	**1.000**	0.991	0.999	0.323	0.894	0.999	0.868
Perplexity	-0.886	0.999	0.999	-0.999	-0.251	0.999	0.144
Perplexity + FSD	**1.000**	0.990	0.999	0.118	**0.908**	**1.000**	0.836
Min-K%	-0.645	0.996	**1.000**	-0.955	0.690	0.999	0.348
Min-K% + FSD	0.997	0.952	0.997	0.421	**0.908**	**1.000**	0.879
Min-K%++	-0.482	0.960	-0.842	0.561	0.514	0.697	0.235
Min-K%++ + FSD	-0.536	0.994	-0.770	0.705	-0.358	0.210	0.041
Kernel Divergence Score (Ours)	0.891	0.982	**1.000**	**0.897**	0.895	**1.000**	**0.944**

paper link

github link