An Indicator of Membership Inference Security in Post-Training Quantized Models

Eric Aubinais[1,*] **Philippe Formont**[2,*] **Pablo Piantanida**[3] **Elisabeth Gassiat**[1]
[1]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, France
[2]Université Paris-Saclay, ILLS, MILA, École de technologie supérieure, Montreal, Canada
[3]ILLS, MILA, CNRS, CentraleSupélec, Montreal, Canada

Abstract

Quantizing machine learning models has demonstrated its effectiveness in lowering memory and inference costs while maintaining performance levels comparable to those of the original models. In this work, we investigate the impact of quantization procedures on privacy in data-driven models, focusing on their vulnerability to membership inference attacks. Membership Inference Security (MIS) has recently been proposed to characterize the privacy of machine learning models against the most powerful (and possibly unknown) attacks. However, quantifying MIS appears to be computationally very difficult. In this paper, we propose a new MIS indicator for post-training quantization procedures of machine learning models that minimizes an empirical loss. This new indicator is a byproduct of a theoretical asymptotic analysis of the MIS in this context. We also present a methodology for empirically estimating our MIS indicator. Using synthetic datasets and real-world data (in the context of drug discovery), we demonstrate the effectiveness of our approach in assessing and ranking the MIS of different quantizers.

[*] Equal contribution.

1 INTRODUCTION

Reducing machine learning models' computational and memory costs is a critical aspect of their deployment, particularly on edge devices and resource-constrained environments. Quantization stands out among the various methods available to enhance inference efficiency in neural networks, such as knowledge distillation and pruning, due to its distinct advantages and proven practical success [Gholami et al., 2022]. One key benefit of quantization is that the storage and latency improvements achieved through quantization are deterministically defined by the chosen quantization level (e.g., using 8-bit integers instead of 32-bit floating-point numbers). Moreover, uniform quantization is inherently hardware-friendly, facilitating the practical realization of theoretical efficiency gains.

While quantization effectively improves efficiency, its impact on the privacy of machine learning models remains largely underexplored. A particularly intriguing question is whether quantization can also strengthen a model's resilience against adversarial threats, such as extracting sensitive information. By reducing the precision of a model's parameters, quantization naturally discards some information [Villard and Piantanida, 2013], which leads to the hypothesis that this process could potentially reduce the risk of recovering the model's training data or other private information. However, to the best of our knowledge, the security of quantized models against such privacy attacks has not yet been theoretically investigated.

In this paper, we study how quantization affects the vulnerability of machine learning models to Membership Inference Attacks (MIAs). Privacy evaluation has traditionally relied on empirical attack-based benchmarks, which, while informative, are inherently limited by their dependence on specific attack strategies. To address this, [Aubinais et al., 2023] introduced Membership Inference Security (MIS), a measure that captures a model's vulnerability to the strongest possible MIA, regardless of the attacker's knowledge. Although MIS is attack-independent, its direct numerical evaluation is computationally demanding, highlighting the need for practical estimation methods. To overcome

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

this difficulty, we propose a novel indicator of MIS and a methodology to estimate it for quantized models trained with empirical risk minimization. Our indicator is grounded in a theoretical asymptotic analysis of MIS as the training dataset size increases, offering both interpretability and practical feasibility.

1.1 Our contributions

We consider learning procedures minimizing an empirical loss for a loss function ℓ. Our contributions can be summarized as follows:

- We establish that, for large datasets, the MIS of a learning algorithm under quantization is fully characterized by the distribution of the per-sample loss across quantized models when the model architecture and/or quantization strategy adapt to the training set size (Theorem 3.4).

  Building on this result, we propose a new MIS indicator that is equal to

  $$r_{\mathcal{Q}_n} \coloneqq \frac{1}{2}\min_k \frac{\left[\mathbb{E}\left[\ell(\bar{\theta}_k^n, \mathrm{z})\right] - \mathbb{E}\left[\ell(\bar{\theta}_{k^*}^n, \mathrm{z})\right]\right]^2}{\mathrm{Var}\left[\ell(\bar{\theta}_k^n, \mathrm{z}) - \ell(\bar{\theta}_{k^*}^n, \mathrm{z})\right]}.$$

  Here n is the size of the dataset; $\bar{\theta}_k^n$ are the given quantized parameters, $k^* \in \arg\min \mathbb{E}\left[\ell(\bar{\theta}_k^n, \mathrm{z})\right]$; and z is a random sample. This indicator enables the comparison of quantization procedures in terms of MIS.

- We propose a methodology for estimating this new indicator based on training trajectories. We empirically validate our method for various quantization techniques on synthetic datasets. We show that the rankings induced by our method consistently correlate with the baseline MIS estimations.

- On real-world data, we further investigate its relationship to classical MIAs against language models and its connection to downstream performance on molecular prediction tasks (see section 4 and the supplementary material).

Code and experimental materials are available at our anonymous repository:`https://anonymous.4open.science/r/Mol_Downstream-B3DB/`.

1.2 Related work

Quantization of neural networks. Several quantization procedures have been studied and employed with the deployment of neural networks on edge devices [Yuan et al., 2024, Lin et al., 2024], where inference should be time and memory-efficient. Quantization usually answers this task by reducing the (bit-)precision of the parameters of the neural networks, demonstrating effectiveness in Large Language Models [Gong et al., 2024, Zhu et al., 2024] even when the quantization is as strong as 1-bit precision quantization [Wang et al., 2023, Ma et al., 2024a], 1.58-bits precision quantization [Ma et al., 2024b], arbitrary bits precision [Zeng et al., 2024]. The most adopted framework of quantization is Post-Training Quantization (PTQ) [Jacob et al., 2018, Nagel et al., 2019, Gholami et al., 2022], which provides a simple training-free implementation. PTQ is usually adopted over Quantization-Aware Training (QAT) [Banner et al., 2018, Pang et al., 2024, Nagel et al., 2021, Nagel et al., 2022, Bengio et al., 2013] due to their limitations to scale up to larger models [Gholami et al., 2022, Lin et al., 2024]. Additionally, some lines of work study "hardware-aware" quantization procedures [Wang et al., 2024, Balaskas et al., 2024] where optimization is made directly on the hardware. During our experiments, we will focus on PTQ.

Membership Inference Attacks. Membership Inference Attacks (MIAs) can reveal sensitive information [Zarifzadeh et al., 2024, Carlini et al., 2022, Shokri et al., 2017, Song et al., 2017, Carlini et al., 2023] about one's data by leveraging the information stored in the parameters of the ML model [Hartley and Tsaftaris, 2022, Del Grosso et al., 2023]. An extensive line of work has developed in the past decade to construct ever so powerful MIAs in embedding models [Song and Raghunathan, 2020], regression models [Gupta et al., 2021], or generative models [Hayes et al.,], systematically summarized in [Hu et al., 2022]. Recent works have leveraged the predictive power of LLMs to construct new MIAs [Staab et al., 2023, Wang et al., 2025]. Several privacy benchmarks have been developed to evaluate the privacy risks of ML models by testing state-of-the-art MIAs on the target model [Murakonda and Shokri, 2020, Liu et al., 2022b]. While these benchmarks provide valuable information about the privacy leakage of an ML model, specific MIAs cannot fully assess the overall privacy resilience of a learning procedure against various attacks. Few works have delved into the theoretical intricacies of MIAs [Sablayrolles et al., 2019, Del Grosso et al., 2023]. More recently, [Aubinais et al., 2023] derived the statistical quantity that governs the effectiveness and success of such attacks.

Quantization and Privacy. Various strategies to protect models from attacks like MIAs have been pro-

posed. In federated learning, the effects of input and gradient quantization have been analyzed through the lens of differential privacy [Youn et al., 2023, Yan et al., 2024, Chaudhuri et al., 2022]. The impact of model quantization on security has been primarily assessed through empirical evaluations of MIAs, with no existing theoretical analysis [Kowalski et al., 2022, Famili and Lao, 2023]. Our work aims to fill this gap by providing a rigorous theoretical evaluation of the security implications of model quantization.

2 BACKGROUND AND NOTATIONS

2.1 Notations

Throughout the article, we consider a dataset $\mathcal{D}_n$ of n independent and identically distributed (i.i.d.) random variables $\mathrm{z}_1, \cdots, \mathrm{z}_n$ drawn from a common distribution P over a space $\mathcal{Z}$. We assume that the goal of the machine learning model is to infer a predictive function $\hat{\Psi}$ from a set of predictors $\mathcal{F} \coloneqq \{\Psi_\theta : \theta \in \Theta\}$ indexed by some space $\Theta \subseteq \mathbb{R}^d$. We define a **learning procedure** (algorithm) as a function $\mathcal{A} : \bigcup_{n\geq 1} \mathcal{Z}^n \to \Theta$. By denoting $\hat{\theta}_n = \mathcal{A}(\mathrm{z}_1, \cdots, \mathrm{z}_n) \in \Theta$, we systematically set $\hat{\Psi} = \Psi_{\hat{\theta}_n}$. We focus on learning procedures minimizing an empirical loss $\theta \mapsto \frac{1}{n}\sum_{j=1}^n \ell(\theta, \mathrm{z}_j)$ where $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}^+$ is the loss function.

Notice that this includes classification tasks, as well as regression or generation tasks.

In the present work, we evaluate the privacy of a learning procedure $\mathcal{A}$ through Membership Inference Attacks (MIAs). Particularly, in a scenario where $\mathcal{D}_n$ consists of sensitive data and the model $\hat{\Psi}$ has been shared (such as a sold product), MIAs pose a notable threat to the privacy of the dataset. MIAs aim at inferring membership of a test sample $\tilde{\mathrm{z}}$ to the dataset $\mathcal{D}_n$ by observing $\hat{\theta}_n$. MIAs can be defined as follows.

Definition 2.1 (Membership Inference Attack - MIA)**.** Any measurable map $\phi : \Theta \times \mathcal{Z} \to \{\text{member}, \text{non-member}\}$ is considered to be a *Membership Inference Attack*.

Throughout the paper, we assume that the MIA may have access to additional information, including the data distribution P, the learning procedure $\mathcal{A}$, and/or other meta-parameters. This framework is usually referred to as white-box [Hu et al., 2022]. Other attack paradigms are included in this setting, such as black-box attacks, which can be considered a restrictive subcategory of white-box attacks.

The remainder of this section introduces important results from [Aubinais et al., 2023] that will serve as the foundation for our analysis.

2.2 MIA accuracy

To evaluate the performance of an MIA ϕ against a learning procedure $\mathcal{A}$, we introduce the concept of MIA accuracy, which is defined as a weighted average between the true positive rate (TPR) and the true negative rate (TNR) of the MIA: $\texttt{TNR}(\phi) + \lambda\texttt{TPR}(\phi)$, with $\lambda > 0$. Note that, by Lagrange duality, maximizing TPR under small TNR is equivalent to maximizing such a weighted average, in which λ is related to the upper bound on TNR.

Although individual MIAs can reveal cases of information leakage, they do not offer a complete evaluation of a model's privacy risks. Additionally, we assume no prior knowledge of potential attacks. Therefore, to thoroughly examine the privacy level of a learning procedure, we focus on the highest accuracy that can be achieved over all MIAs, including unknown ones. However, taking the supremum over ϕ in $\texttt{TNR}(\phi) + \lambda\texttt{TPR}(\phi)$ leads to a degenerate non-informative result if we aim to compare algorithms. As a result, we define the accuracy as follows.

Definition 2.2 (Accuracy of a given MIA)**.** The *accuracy of an MIA* ϕ is defined as

$$\texttt{Acc}_n(\phi; P, \mathcal{A}) \coloneqq \mathbb{E}[\texttt{TNR}(\phi)] + \lambda\mathbb{E}[\texttt{TPR}(\phi)], \qquad (1)$$

where the expectation is considered over all sources of randomness inherent in the underlying training model and the data used for both training and evaluation. In particular, we assume that the test sample $\tilde{\mathrm{z}}$ is drawn at random over $\{\mathrm{z}_1, \cdots, \mathrm{z}_n\}$ with probability $\eta \in (0, 1)$ conditionnally to $\{\mathrm{z}_1, \cdots, \mathrm{z}_n\}$, otherwise drawn from P independently.

2.3 Membership inference security (MIS)

We now define the MIA-wise privacy level of a learning procedure as

Definition 2.3 (MIS [Aubinais et al., 2023])**.** The *Membership Inference Security* (MIS) of a learning procedure $\mathcal{A}$ is defined as

$$\texttt{MIS}_n(P, \mathcal{A}) \coloneqq c_1 \left(1 - c_2 \sup_{\phi} \texttt{Acc}_n(\phi; P, \mathcal{A})\right), \qquad (2)$$

where c_1 and c_2 are scaling constants depending on λ and η. The supremum is taken over all MIAs.

The values of the scaling constants c_1 and c_2 are designed to ensure that the MIS ranges from 0 to 1. When the MIS approaches 0, an attack that achieves almost maximum accuracy exists. Conversely, no MIA performs better than a random or constant MIA when the MIS approaches 1.

The MIS is designed to be specific to MIAs, unlike the more general concept of differential privacy (DP) [Dwork et al., 2014]. It can be shown that if a learning procedure is differentially private, then the MIS is lower bounded by some function of the DP parameters. However, the converse does not hold. There are learning procedures which are not differentially private but for which the MIS approaches 1.

One of the main results in [Aubinais et al., 2023] is proving that the MIS is equal to an information-theoretic quantity that measures the divergence between the joint distribution of $(\hat{\theta}_n, \mathrm{z}_1)$ and the product distribution of $\hat{\theta}_n$ and an independent sample z. Here z_1 can be replaced by any z_i of the n-sample, since $\hat{\theta}_n$ is a function of the empirical distribution of the n-sample. More details are provided in section 7.

Direct estimation of divergences between probability distributions over high-dimensional spaces is computationally prohibitive and can not be used to evaluate the MIS. As a function of the maximum accuracy of a test function (MIA), the MIS can be evaluated as a binary classifier to get a numerical evaluation of MIS. We call it the baseline method (see subsection 4.1). However, this baseline remains computationally expensive, motivating the development of MIS evaluation techniques tailored to specific applications. In this work, we design such methods for quantized learning algorithms.

3 ATTACK-FREE MIS EVALUATION FOR QUANTIZED PROCEDURES

This section provides the main theoretical results on the MIS of quantized learning procedures. First, we introduce quantization procedures. Then, we present our theoretical asymptotic result, from which we deduce the definition of our new MIS indicator. Finally, we propose a method to estimate the indicator.

3.1 Quantization

We define a **quantizer** [Gersho and Gray, 1992] as any measurable function $\mathcal{Q} : \Theta \to \bar{\Theta} \subseteq \Theta$ for some discrete space $\bar{\Theta} \coloneqq \{\bar{\theta}_1, \cdots, \bar{\theta}_K\}$.

Example 3.1 (Binarized Neural Networks [Wang et al., 2023]). Let $\mathcal{F}$ be a set of neural networks with fixed architecture. A scalar quantizer $\mathcal{Q}$ maps coordinate-wise the parameters to its sign: $\mathcal{Q}(\theta) = (\theta_j/|\theta_j|)_j$. Here, for any neural networks with d parameters, the set $\bar{\Theta}$ would consist of all $d-$dimensional vectors in $\{-1, +1\}^d$, hence $K = 2^d$.

Example 3.2 (Vector Quantization). Another quantization procedure, albeit underused in practice, is vector quantization. A *codebook* $\bar{\Theta}$ is usually pre-computed, which the vector quantizer $\mathcal{Q}$ maps θ onto, usually performed by a nearest neighbor algorithm, which makes it efficient and memory-friendly. The constant K corresponds to the number of values stored in the codebook.

In practice, when developing machine learning models, it is common to adjust the model's architecture based on the size of the dataset. For example, models are often over-parameterized relative to the size of the dataset, as is the case with LLMs. This is why we let our quantizer $\mathcal{Q}_n$ (and therefore the number of quantized values K_n and $\bar{\Theta}_n \coloneqq \{\bar{\theta}_1^n, \cdots, \bar{\theta}_{K_n}^n\}$) depend on the sample size. The dependence of $\mathcal{Q}_n$ on the dataset size n formalizes at least two scenarios: either the quantization procedure remains the same, but the architecture size adapts to the training dataset, or the architecture size is fixed while the quantization procedure changes. Specifically, the first interpretation can be seen as the common practice in machine learning to scale models to the dataset size.

Example 3.3 (Scaling Architecture). Let the number of parameters of our original models follow a scaling law [Hoffmann et al., 2022, Kaplan et al., 2020] f, i.e. its number of parameters is $f(n)$. Let the quantization method be fixed to a 1-bit quantization (mapping each parameter to its sign, for instance). From Example 3.1, for a dataset size n, the size of $\bar{\Theta}_n$ is $K_n = 2^{f(n)}$.

A quantizer canonically induces a **quantized learning procedure** $\mathcal{A}_{\mathcal{Q}_n}$. Additionally, we will assume without loss of generality that quantizers are numbered in ascending order of their expected loss as follows

$$\mathbb{E}\left[\ell(\bar{\theta}_1^n, \mathrm{z})\right] \leq \cdots \leq \mathbb{E}\left[\ell(\bar{\theta}_{K_n}^n, \mathrm{z})\right],$$

where z is a random variable with distribution P. We define the **loss gaps**

$$\delta_k^n \coloneqq \mathbb{E}\left[\ell(\bar{\theta}_k^n, \mathrm{z})\right] - \mathbb{E}\left[\ell(\bar{\theta}_1^n, \mathrm{z})\right],$$

and the **loss variabilities**

$$\left(\rho_{k,l}^n\right) = \mathrm{Cov}\big(\ell\left(\bar{\theta}_k^n, \mathrm{z}\right) - \ell\left(\bar{\theta}_1^n, \mathrm{z}\right); \ell\left(\bar{\theta}_l^n, \mathrm{z}\right) - \ell\left(\bar{\theta}_1^n, \mathrm{z}\right)\big).$$

3.2 Asymptotic theorem and the new MIS indicator

The following asymptotic result is a consequence of the moderate deviation principle. One of the main difficulties in proving this theorem was establishing the appropriate framework and assumptions to apply the principle. We now present the assumptions.

(A1) We have $\delta_2^n \underset{n\to\infty}{\to} 0$ and $\sqrt{n}\delta_2^n \underset{n\to\infty}{\to} +\infty$.

(A2) a) The limits $\rho_{k,l} = \lim_{n\to\infty} \rho_{k,l}^n$ and $c_k = \lim_{n\to\infty} \frac{\delta_2^n}{\delta_k^n}$

exist.
b) There exists $K < \infty$ such that for all $k > K$, $c_k^2 \sigma_k^2 = 0$. Here, $\sigma_k^2 = \rho_{k,k}$.
c) The matrix $\Lambda = (\rho_{k,l} c_k c_l)_{2 \le k,l \le K}$ is non singular.
(A3) For all $t > 0$, there exists M such that for all n, $\mathbb{E}\left[\exp t \left\| \left(\frac{\delta_2^n}{\delta_k^n} \left(\ell\left(\bar{\theta}_k^n, \mathrm{z}\right) - \ell\left(\bar{\theta}_1^n, \mathrm{z}\right) \right) \right)_{1 < k \le K_n} \right\|_2 \right] \le M$.
(A4) The sequence $\left(\left(\frac{\delta_2^n}{\delta_k^n} \left(\ell\left(\bar{\theta}_k^n, \mathrm{z}\right) - \ell\left(\bar{\theta}_1^n, \mathrm{z}\right) \right)_{1 < k \le K_n} \right)_n \right.$, in $l_2(\mathbb{R})$, of random variables, is tight.

Assumptions **(A1)** and **(A2)** depend on the quantization procedure. They indicate that the quantization procedure has two important properties: the expected loss landscape near its minimum can be well described by multiple quantized parameter configurations; the per-sample losses have enough variability across the quantized parameters. Assumption **(A2)** b) is empirically supported by our experiments (see the supplementary materials section 14 in Figure 13). When the loss function is bounded, **(A3)** is satisfied, and **(A4)** and **(A2)** a) are light assumptions.

Theorem 3.4. *Under the assumptions **(A1), (A2), (A3), (A4)** we have*

$$\limsup_{n \to \infty} \frac{1}{n \left(\delta_2^n\right)^2} \log\left(1 - \mathtt{MIS}_n(P, \mathcal{A}_{\mathcal{Q}_n})\right) \le -\frac{1}{2 \max_k c_k^2 \sigma_k^2} < 0.$$

The fact that $\max_k c_k^2 \sigma_k^2$ is finite follows from **(A2)**. Theorem 3.4 shows that the MIS of the quantization procedure approaches 1 when the sample size $n \to \infty$, thanks to **(A1)**, at a specified rate, controlled by

$$\bar{r}_{\mathcal{Q}_n} := \left(\delta_2^n\right)^2 / (2 \max_k c_k^2 \sigma_k^2).$$

Theorem 3.4 suggests that for quantization procedures $\mathcal{Q}_n$ and $\mathcal{R}_n$, if $\bar{r}_{\mathcal{Q}_n} \ge \bar{r}_{\mathcal{R}_n}$, then $\mathcal{A}_{\mathcal{Q}_n}$ produces more secure parameters than $\mathcal{A}_{\mathcal{R}_n}$ for a large enough sample size. We use this rate to define the new MIS indicator, in which the limits are approximated by the value at the sample size n.

Definition 3.5 (MIS indicator)**.** The MIS indicator of the quantization procedure $\mathcal{Q}_n$ is defined as

$$r_{\mathcal{Q}_n} := \frac{1}{2} \min_{k>1} \frac{\left[\mathbb{E}\left[\ell(\bar{\theta}_k^n, \mathrm{z})\right] - \mathbb{E}\left[\ell(\bar{\theta}_1^n, \mathrm{z})\right]\right]^2}{\mathrm{Var}\left[\ell(\bar{\theta}_k^n, \mathrm{z}) - \ell(\bar{\theta}_1^n, \mathrm{z})\right]}.$$

Note that the MIS indicator wholly relies upon the random variables $\ell(\bar{\theta}_k^n, \mathrm{z}) - \ell(\bar{\theta}_1^n, \mathrm{z})$, making it an attack-free privacy indicator. The next step is to be able to compute the indicator.

3.3 Estimation of the MIS Indicator

To estimate the MIS indicator of a quantization procedure $\mathcal{Q}_n$ we must compute the loss gaps δ_k^n for all $k \le K_n$. However, this is computationally infeasible, as even a simple quantizer like 1-bit quantization leads to an exponentially large K_n. We observe that $\max_k \left(\delta_2^n / \delta_k^n\right)^2 (\sigma_k^n)^2$ is empirically dominated by low-loss quantized models (see subsection 14.2). This suggests that exploring the entire set $\bar{\Theta}$ is unnecessary; focusing on low-loss quantized models is sufficient to estimate $r_{\mathcal{Q}_n}$.

We thus propose estimating using quantized models derived from $\hat{\theta}_n$'s training trajectory. Since the quantizers with the lowest loss likely reside near the trajectory of the minimized loss, we limit our computations to these weights to efficiently capture critical quantizers. The estimation procedure given the list of quantized losses is outlined in Algorithm 1, and additional details are given in section 9. Expectations and variances are estimated by the empirical estimators using the sample set. Finally, to obtain more accurate estimations of $r_{\mathcal{Q}_n}$, we average the estimated value over multiple training trajectories (we analyzed the influence of the number of trajectories in subsection 4.1).

Algorithm 1 Estimation of $r_{\mathcal{Q}_n}$ from the quantized losses

Input: The set of each sample's loss for the K checkpoints: $\mathcal{L} \in \mathbb{R}^{K \times n}$, and the average loss of each checkpoint $m \in \mathbb{R}^K$
Output: An estimate of $r_{\mathcal{Q}_n}$.
idx $\leftarrow$ argsort(m). // Sort the quantized models
$m \leftarrow m$[idx].
$\mathcal{L} \leftarrow \mathcal{L}$[idx].
for $k = 2$ to K **do**
    $\sigma^2[k] \leftarrow \mathrm{Var}\left(\mathcal{L}[k] - \mathcal{L}[1]\right)$.
end for
return $r_{\mathcal{Q}} \leftarrow \frac{1}{2} \left[\max_{2 \le k \le K} \frac{\sigma^2[k]}{(m[k] - m[1])^2} \right]^{-1}$.

4 NUMERICAL EXPERIMENTS

4.1 Validation of $r_{\mathcal{Q}_n}$ as a MIS indicator

To validate that our estimation of $r_{\mathcal{Q}_n}$ effectively ranks quantization methods by their privacy level, we compare the resulting ordering with the baseline MIS-based ranking obtained from synthetic experiments. As we focus solely on ranking quantization procedures, we omit specific $r_{\mathcal{Q}_n}$ values in the figures for clarity and readability.

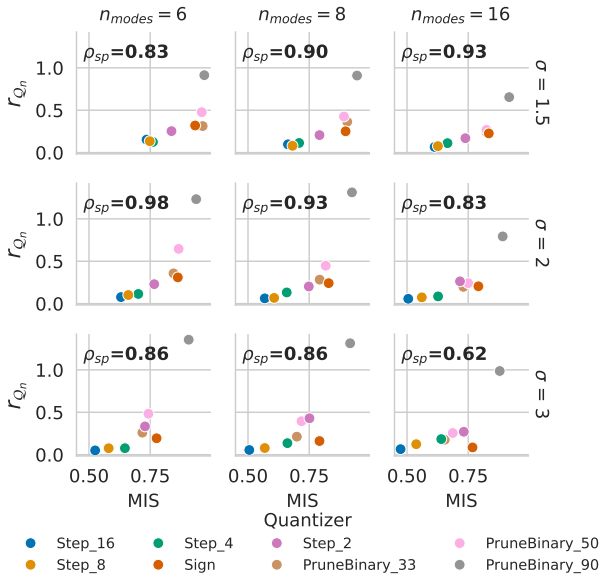


Figure 1: **Relationship between $r_{\mathcal{Q}_n}$ and the MIS.** Estimated values of $r_{\mathcal{Q}_n}$ and the MIS for each quantizers on different datasets, with their corresponding Spearman correlation (ρ_{sp}).

MIS Baseline. To evaluate the ability of our estimator to correctly rank quantizers in terms of membership security, we rely on a baseline estimation of the MIS (theoretical details given in section 11). This baseline consists of training a discriminator g_ψ to differentiate between samples from the training set and other samples given the model's weights.

To train the classifier, we follow the following steps:

1. We draw $n = 128$ samples (x, y) from our synthetic dataset, and train a classifier $\hat{\theta}_n$ on these samples.
2. We draw n samples, that are the samples out of $\hat{\theta}_n$'s training set.
3. We create the sub-dataset built with $\hat{\theta}_n$: $\{(\hat{\theta}_n, (x, y)_i, t_i)\}_{i \leq 2n}$ where t_i equals 1 if the sample $(x, y)_i$ is in $\hat{\theta}_n$'s training set, and 0 otherwise.
4. We perform these operations k=300 times and create the full dataset as the concatenation of all sets described in step 2, which we split into a training and testing set, to train and evaluate g_ψ.

By construction, all samples (x, y) are drawn independently, and in particular all samples used to evaluate g_ψ are independent from all samples in its training set. This property is particularly data intensive and is the reason why we focused on synthetic data in this section. Finally, the baseline approach measures the

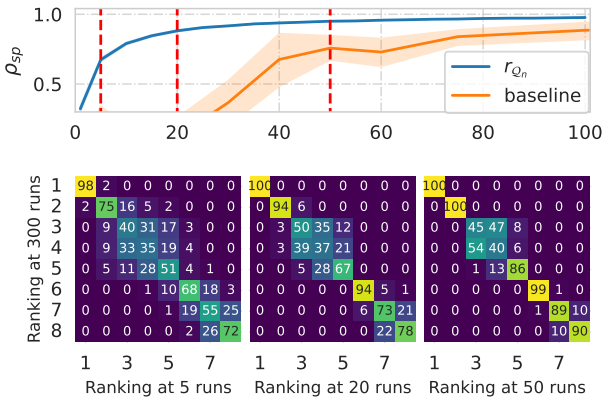


Figure 2: **Stability of $r_{\mathcal{Q}_n}$'s estimation.** Spearman correlation between rankings obtained with a limited number of runs and those obtained with the full set of 300 runs. The ranking induced by $r_{\mathcal{Q}_n}$ stabilizes after about 15 runs, while the baseline MIS requires substantially more runs to converge to the final ranking.

MIS as $MIS_{g_\psi} = 2 - 2\text{Acc}\left(g_\psi, P, \mathcal{A}\right)$ (more details are provided in section 11).

Datasets and Classifiers. We generate data points sampled from $\mathbb{R}^{128}$ using $k_{\text{modes}} \in \{6, 8, 16\}$ isotropic Gaussian distributions of standard deviation $\sigma \in \{1.5, 2, 3\}$, where each cluster is assigned a label in $\{0, 1\}$. We trained each classifier (single-layer fully connected networks) on $n = 128$ samples using the Adam optimizer with a learning rate of 10^{-4}, and we stopped the training at 30 epochs (see the training curves in section 12).

Quantization. For these experiments, we consider a range of different simple quantization methods, including: 1bit quantization by taking the sign of the weights (`Sign`), 1.58 bits quantization with different sparsity levels (`1.58b {x}%` where x% of the weights with the smallest magnitude are set to zero, and the rest to their sign), and quantization from 2 to 5 bits.

Estimation of the privacy guarantees. Figure 1 illustrates the correlation between our proposed metric $r_{\mathcal{Q}_n}$ and the MIS baseline. We report the Spearman correlation between both metrics (the correlation between the rankings induced by both metrics), and as expected, quantizers with more bits of information, such as `5 bits` and `4 bits`, are the least private. In contrast, the `1.58b 90%` quantizer, which introduces 90% sparsity by setting weights to zero, achieves the highest privacy. Overall, the rankings produced by $r_{\mathcal{Q}_n}$ closely match those of the baseline method, with an average Spearman correlation of $\rho_{sp} = 0.86$, demonstrating that $r_{\mathcal{Q}_n}$ reliably ranks quantization methods

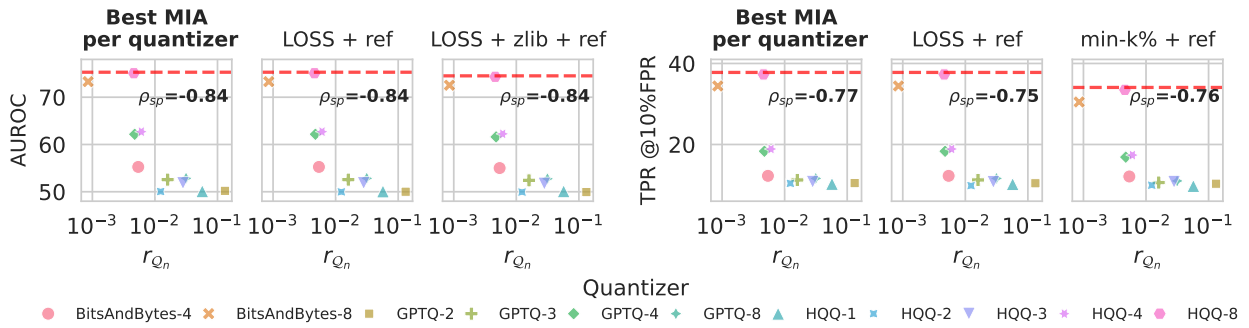


Figure 3: **Quantized security against the performances of various black-box MIAs for language models.** The first plot shows the relation between $r_{\mathcal{Q}_n}$ the performances of the best MIA for each quantizer. As $r_{\mathcal{Q}_n}$ increases, the performances of the MIAs decreases, indicating that the quantized model is more secure. The dashed red line indicates the performance of the different MIAs on the full-precision model.

by their privacy levels.

Stability of the rankings. In Figure 2, we examine how the number of runs influences the stability of the ranking by comparing the rankings obtained with subsets of runs ($k_{\text{runs}} \leq 300$) against those obtained using all 300 runs. For both $r_{\mathcal{Q}_n}$ and the baseline MIS evaluation, increasing the number of runs ($\hat{\theta}_n$) yields rankings that better align with the rankings using all runs. Notably, $r_{\mathcal{Q}_n}$ achieves stability much faster: after only 15 runs, its rankings already exhibit a high correlation with the 300-run ranking ($\rho_{sp} \geq 0.9$), whereas the baseline MIS requires around 150 runs to reach the same correlation level.

4.2 Validation on practical use-case: Language models for drug discovery

As a second experimental setting, we consider a real-world application in drug discovery. In this domain, data is both highly valuable and sensitive, making it crucial to assess whether predictive models might inadvertently leak proprietary information. We evaluate the privacy of generative language models trained on chemical knowledge, comparing our proposed measure $r_{\mathcal{Q}_n}$ against standard membership inference attacks (MIAs) for language models.

Experimental setup. We fine-tuned 20 models based on Qwen2.5-instruct-0.5B [Bai et al., 2023] using LORA adapters of rank 64 [Hu et al., 2021] on the SMolInstruct dataset [Yu et al., 2024] (≈10k samples per model), for 10 epochs. We considered the molecular captioning and generation tasks, consisting of either giving a description of a given molecule, or generating a molecule from a given description. Results on Ministral-7B and Llama3-8B are given in Figure 11.

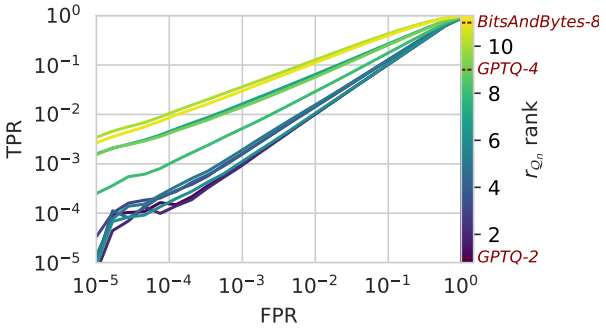


Figure 4: **TPR-FPR curves of LiRa attacks.** True positive rate (TPR) of a LiRa offline attack at different low false positive rate (FPR). Models that are hardest to attack (resp. easiest to attack) are also those ranked as more (resp. least) secure by $r_{\mathcal{Q}_n}$.

Quantizers. We evaluate the value of $r_{\mathcal{Q}_n}$ for various classical quantization methods including BitsAndBytes [Dettmers et al., 2022] (4-8 bits), GPTQ [Frantar et al., 2023] (2-3-4-8 bits), and HQQ [Badri and Shaji, 2023] (1-2-3-4-8 bits). We evaluate the per-sample loss of the quantized model on 22 checkpoints of the model to estimate $r_{\mathcal{Q}_n}$.

Black-box results. We compare the evolution of our privacy metric $r_{\mathcal{Q}_n}$ with the performance of several black-box MIAs on the quantized models (see section 13 for details on the selected attacks). Figure 3 shows the relationship between $r_{\mathcal{Q}_n}$ and MIA performance, including the best-performing attack on each quantizer. We find that higher values of $r_{\mathcal{Q}_n}$ correspond to lower MIA performance, indicating stronger privacy guarantees. Conversely, lower values of $r_{\mathcal{Q}_n}$ are associated with higher MIA performance. Overall, these results demonstrate that $r_{\mathcal{Q}_n}$ successfully orders quantization methods according to their privacy level.

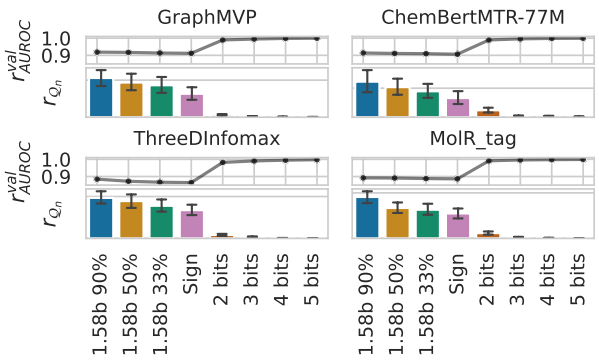


Figure 5: **Impact of quantization on classification tasks.** Evolution of each downstream model's privacy $r_{\mathcal{Q}_n}$ along with relative performances of the quantized models compared to the original on classification tasks.

LiRa attack. We further compare the privacy rankings of the quantizers with the performance of one of the strongest MIAs: the Likelihood Ratio attack (LiRa) [Carlini et al., 2022, Hayes et al., 2025]. This attack requires training a set of reference models on random subsets of the data, and then comparing their outputs on a sample with that of the target model (model under attack). In its online form, the target model is compared to N reference models trained without the sample of interest and to N reference models trained with it. Because of its computational cost, an offline variant has been proposed, which relies only on models trained without the sample of interest. We adopt this offline setting with $N = 35$, which we found sufficient to obtain effective attacks. Figure 4 reports the true positive rate (TPR) at varying false positive rates (FPR) for all quantizers, and shows that the quantizers hardest to attack are also those ranked as most secure by $r_{\mathcal{Q}_n}$, further validating our metric.

4.3 Security performance tradeoff on molecular property prediction

Finally, we study the security of small predictive models trained on molecular embedders, focusing on the tradeoff between performance and privacy.

Pretrained Embedders. To generate one-dimensional molecular embeddings, we evaluate four pretrained models from the representation learning literature: GraphMVP [Liu et al., 2022a] and 3D-Infomax [Stärk et al., 2021] (3D-2D mutual information maximization), MolR [Wang et al., 2022] (reaction-aware pretraining), and ChemBERTa-MTR [Ahmad et al., 2022] (multitask regression with string representations of the molecule). The embeddings are passed to two layers feed-forward networks (with a hidden dimension of 128) trained on each downstream task. We trained all models for 500 epochs with a learning rate of 1e-4, and selected the best checkpoint based on the validation loss. We focused on the simple quantization methods presented in section 10 for their interpretability.

Downstream tasks. We evaluate and train the models on various property prediction tasks (classification) from the Therapeutic Data Commons (TDC) platform [Huang et al., 2021], focusing on ADMET properties (Absorption, Distribution, Metabolism, Excretion, and Toxicity). Additional results and tradeoffs with regression tasks are given in section 14. We report the AUC-ROC scores and to ensure robust estimation of $r_{\mathcal{Q}_n}$, we allocate 40% of the dataset to the validation set. All models are trained using 90% of the dataset for each run to ensure different training trajectories are sampled. Finally, we report the relative performance of the quantized model to the original: $r^{\text{val}}_{\text{AUROC}} = \text{AUROC}^{\text{val}}_{\text{quantized}} / \text{AUROC}^{\text{val}}_{\text{original}}$.

Performance vs Privacy trade-off. Figure 5 shows the trade-off between privacy ($r_{\mathcal{Q}_n}$) and average task performance across embedders. Even under the strongest quantization, AUROC drops by only about 10%, likely because classification primarily depends on boundary regions that seem not to be too affected by quantization procedures. We see that higher downstream accuracy generally comes at the expense of lower privacy. A threshold effect emerges: quantization at 2 bits or more preserves performance but does not improve privacy, whereas stronger quantization (1.58 bits) increases security at the cost of roughly 10% AUROC.

5 CONCLUSION, LIMITATIONS AND PERSPECTIVES

In this work, we investigated Membership Inference Security of quantization procedures for machine learning models that minimize an empirical loss. We proved that the learning procedure is asymptotically determined by the distribution of the quantized models' loss per sample. From this, we derived a new MIS indicator. We introduced an attack-free methodology for comparing quantization procedures in terms of MIS, with limited computational cost. Through extensive experiments on both synthetic and real-world datasets, we validated the effectiveness of our approach.

We believe our work suggests several interesting new research directions related to the limitations of our paper. First, we defined MIS as quantifying the security against any possible MIA. In some cases, however, not all MIAs are threats. The supremum in Equation 2 could be taken only over a subset of all MIAs; for ex-

ample, consider black-box attacks. Additionally, our study focused on post-training quantized models and did not identify any quantizer that can achieve strong downstream performance and high security simultaneously. One interesting direction for future research is to propose Quantization Aware Training methods to develop quantizers that are trained to maximize $r_{\mathcal{Q}_n}$ while being jointly optimized with the task-specific loss function. This is an area of ongoing research.

References

[Ahmad et al., 2022] Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. (2022). Chemberta-2: Towards chemical foundation models.

[Araujo and Giné, 1980] Araujo, A. and Giné, E. (1980). The central limit theorem for real and banach valued random variables. *(No Title)*.

[Aubinais et al., 2023] Aubinais, E., Gassiat, E., and Piantanida, P. (2023). Fundamental limits of membership inference attacks on machine learning models. *arXiv preprint arXiv:2310.13786*. Accepted for publication in JMLR.

[Badri and Shaji, 2023] Badri, H. and Shaji, A. (2023). Half-quadratic quantization of large machine learning models.

[Bai et al., 2023] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

[Balaskas et al., 2024] Balaskas, K., Karatzas, A., Sad, C., Siozios, K., Anagnostopoulos, I., Zervakis, G., and Henkel, J. (2024). Hardware-aware dnn compression via diverse pruning and mixed-precision quantization. *IEEE Transactions on Emerging Topics in Computing*, 12(4):1079–1092.

[Banner et al., 2018] Banner, R., Hubara, I., Hoffer, E., and Soudry, D. (2018). Scalable methods for 8-bit training of neural networks. *Advances in neural information processing systems*, 31.

[Bengio et al., 2013] Bengio, Y., Léonard, N., and Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

[Carlini et al., 2022] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022). Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.

[Carlini et al., 2023] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. (2023). Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270.

[Carlini et al., 2021] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models.

[Chaudhuri et al., 2022] Chaudhuri, K., Guo, C., and Rabbat, M. (2022). Privacy-aware compression for federated data analysis. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 296–306. PMLR.

[De Acosta, 1992] De Acosta, A. (1992). Moderate deviations and associated laplace approximations for sums of independent random vectors. *Transactions of the American Mathematical Society*, 329(1):357–375.

[Del Grosso et al., 2023] Del Grosso, G., Pichler, G., Palamidessi, C., and Piantanida, P. (2023). Bounding information leakage in machine learning. *Neurocomputing*, 534:1–17.

[Dettmers et al., 2022] Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

[Duan et al., 2024] Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. (2024). Do membership inference attacks work on large language models? In *First Conference on Language Modeling*.

[Dwork et al., 2014] Dwork, C., Roth, A., et al. (2014). *The algorithmic foundations of differential privacy*, volume 9. Now Publishers, Inc.

[Famili and Lao, 2023] Famili, A. and Lao, Y. (2023). Deep neural network quantization framework for effective defense against membership inference attacks. *Sensors*, 23(18).

[Frantar et al., 2023] Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2023). Gptq: Accurate post-training quantization for generative pre-trained transformers.

[Gersho and Gray, 1992] Gersho, A. and Gray, R. M. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers.

[Gholami et al., 2022] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. (2022).

A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC.

[Gong et al., 2024] Gong, R., Ding, Y., Wang, Z., Lv, C., Zheng, X., Du, J., Qin, H., Guo, J., Magno, M., and Liu, X. (2024). A survey of low-bit large language models: Basics, systems, and algorithms. *arXiv preprint arXiv:2409.16694*.

[Gupta et al., 2021] Gupta, U., Stripelis, D., Lam, P. K., Thompson, P., Ambite, J. L., and Ver Steeg, G. (2021). Membership inference attacks on deep regression models for neuroimaging. In *Medical Imaging with Deep Learning*, pages 228–251. PMLR.

[Hartley and Tsaftaris, 2022] Hartley, J. and Tsaftaris, S. A. (2022). Measuring unintended memorisation of unique private features in neural networks. *arXiv preprint arXiv:2202.08099*.

[Hayes et al.,] Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. L. Membership inference attacks against generative models; 2018. *URL https://api. semanticscholar. org/CorpusID*, 202588705.

[Hayes et al., 2025] Hayes, J., Shumailov, I., Choquette-Choo, C. A., Jagielski, M., Kaissis, G., Lee, K., Nasr, M., Ghalebikesabi, S., Mireshghallah, N., Annamalai, M. S. M. S., Shilov, I., Meeus, M., de Montjoye, Y.-A., Boenisch, F., Dziedzic, A., and Cooper, A. F. (2025). Strong membership inference attacks on massive datasets and (moderately) large language models.

[Hoffmann et al., 2022] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2022). Training compute-optimal large language models.

[Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

[Hu et al., 2022] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

[Huang et al., 2021] Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. (2021). Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks*.

[Jacob et al., 2018] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

[Kaplan et al., 2020] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models.

[Kowalski et al., 2022] Kowalski, C., Famili, A., and Lao, Y. (2022). Towards Model Quantization on the Resilience Against Membership Inference Attacks. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3646–3650. ISSN: 2381-8549.

[Lin et al., 2024] Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. (2024). Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

[Liu et al., 2022a] Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. (2022a). Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*.

[Liu et al., 2022b] Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M., and Zhang, Y. (2022b). {ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4525–4542.

[Ma et al., 2024a] Ma, L., Sun, M., and Shen, Z. (2024a). Fbi-llm: Scaling up fully binarized llms from scratch via autoregressive distillation. *arXiv preprint arXiv:2407.07093*.

[Ma et al., 2024b] Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., and Wei, F. (2024b). The era of 1-bit llms: All large language models are in 1.58 bits.

[Murakonda and Shokri, 2020] Murakonda, S. K. and Shokri, R. (2020). Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339*.

[Nagel et al., 2019] Nagel, M., Baalen, M. v., Blankevoort, T., and Welling, M. (2019). Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1325–1334.

[Nagel et al., 2021] Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van Baalen, M., and Blankevoort, T. (2021). A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*.

[Nagel et al., 2022] Nagel, M., Fournarakis, M., Bondarenko, Y., and Blankevoort, T. (2022). Overcoming oscillations in quantization-aware training. In *International Conference on Machine Learning*, pages 16318–16330. PMLR.

[Pang et al., 2024] Pang, J., Cai, T., Zhang, B., Wu, J., and Tao, Y. (2024). Push quantization-aware training toward full precision performances via consistency regularization. *arXiv preprint arXiv:2402.13497*.

[Sablayrolles et al., 2019] Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. (2019). White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR.

[Shi et al., 2024] Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. (2024). Detecting pretraining data from large language models.

[Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

[Song and Raghunathan, 2020] Song, C. and Raghunathan, A. (2020). Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.

[Song et al., 2017] Song, C., Ristenpart, T., and Shmatikov, V. (2017). Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601.

[Staab et al., 2023] Staab, R., Vero, M., Balunović, M., and Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

[Stärk et al., 2021] Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. (2021). 3d infomax improves gnns for molecular property prediction. *arXiv preprint arXiv:2110.04126*.

[Villard and Piantanida, 2013] Villard, J. and Piantanida, P. (2013). Secure multiterminal source coding with side information at the eavesdropper. *IEEE Transactions on Information Theory*, 59(6):3668–3692.

[Wang et al., 2025] Wang, H., Fu, W., Tang, Y., Chen, Z., Huang, Y., Piao, J., Gao, C., Xu, F., Jiang, T., and Li, Y. (2025). A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy. *arXiv preprint arXiv:2501.09431*.

[Wang et al., 2022] Wang, H., Li, W., Jin, X., Cho, K., Ji, H., Han, J., and Burke, M. D. (2022). Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*.

[Wang et al., 2023] Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., and Wei, F. (2023). Bitnet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453*.

[Wang et al., 2024] Wang, L., Ma, L., Cao, S., Zhang, Q., Xue, J., Shi, Y., Zheng, N., Miao, Z., Yang, F., Cao, T., et al. (2024). Ladder: Enabling efficient {Low-Precision} deep learning computing through hardware-aware tensor transformation. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 307–323.

[Yan et al., 2024] Yan, G., Li, T., Wu, K., and Song, L. (2024). Killing two birds with one stone: Quantization achieves privacy in distributed learning. *Digital Signal Processing*, 146:104353.

[Yeom et al., 2018] Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting.

[Youn et al., 2023] Youn, Y., Hu, Z., Ziani, J., and Abernethy, J. (2023). Randomized quantization is all you need for differential privacy in federated learning.

[Yu et al., 2024] Yu, B., Baker, F. N., Chen, Z., Ning, X., and Sun, H. (2024). LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*.

[Yuan et al., 2024] Yuan, Z., Zhou, R., Wang, H., He, L., Ye, Y., and Sun, L. (2024). Vit-1.58 b: Mobile vision transformers in the 1-bit era. *arXiv preprint arXiv:2406.18051*.

[Zarifzadeh et al., 2024] Zarifzadeh, S., Liu, P., and Shokri, R. (2024). Low-cost high-power membership inference attacks.

[Zeng et al., 2024] Zeng, C., Liu, S., Xie, Y., Liu, H., Wang, X., Wei, M., Yang, S., Chen, F., and Mei, X. (2024). Abq-llm: Arbitrary-bit quantized inference acceleration for large language models. *arXiv preprint arXiv:2408.08554*.

[Zhu et al., 2024] Zhu, X., Li, J., Liu, Y., Ma, C., and Wang, W. (2024). A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

An Indicator of Membership Inference Security in Post-Training Quantized Models: Supplementary Materials

Part I

Supplementary Materials

Table of Contents

6 Notations

Variable	Description	Mathematical Definition	First introduction
P	data distribution	–	subsection 2.1
$\mathcal{Z}$	data space	–	
z_j	training sample	$\mathrm{z}_j \sim P$	
n	training dataset size	–	
$\mathcal{D}_n$	training dataset	–	
$\hat{P}_n$	empirical distribution	$\frac{1}{n}\sum_{j=1}^{n} \delta_{\mathrm{z}_j}$	
Ψ_θ	predictor with parameters θ	–	
Θ or Θ_n	set of parameters	$\Theta \subseteq \mathbb{R}^d$	
$\mathcal{F}$	set of predictors	$\{\Psi_\theta : \theta \in \Theta\}$	
ℓ	loss function	$\ell : \Theta \times \mathcal{Z} \to \mathbb{R}^+$	
$\mathcal{A}$	learning procedure	$\mathcal{A} : \bigcup_{n\geq 1} \mathcal{Z}^n \to \Theta$	
ϕ	MIA	$\phi : \Theta \times \mathcal{Z} \to \{0,1\}$	subsection 2.2
$\tilde{\mathrm{z}}$	test sample	–	
η	probability of $\tilde{\mathrm{z}}$ being member	$\mathbb{P}(\tilde{\mathrm{z}} \text{ is member})$	
λ	importance of the TPR	–	
$\mathtt{Acc}_n(\phi; P, \mathcal{A})$	accuracy of an MIA	$\mathbb{E}[\mathtt{TNR}] + \lambda\mathbb{E}[\mathtt{TPR}]$	
$\mathtt{MIS}_n(P, \mathcal{A})$	Membership Inference Security	$c_1(1 - c_2 \sup_\phi \mathtt{Acc}_n(\phi; P, \mathcal{A}))$	subsection 2.3
$\bar{\Theta}_n$	discrete set of parameters	$\bar{\Theta}_n = \{\bar{\theta}_1^n, \dots, \bar{\theta}_{K_n}^n\} \subseteq \Theta$	section 3
$\mathcal{Q}_n$	quantizer	$\mathcal{Q}_n : \Theta \to \bar{\Theta}_n$	
$\mathbb{P}_{\mathrm{x}}, \mathbb{P}_{(\mathrm{x}_1,\mathrm{x}_2)}, \mathbb{P}_{\mathrm{x}_1} \otimes \mathbb{P}_{\mathrm{x}_2}$	marginal, joint, product distribution	–	
δ_k^n	loss gap	$\mathbb{E}[\ell(\bar{\theta}_k^n, \mathrm{z})] - \mathbb{E}[\ell(\bar{\theta}_1^n, \mathrm{z})]$	
$\rho_{k,l}^n$	(k,l)-loss variability	$Cov(\ell(\bar{\theta}_k^n, \mathrm{z}) - \ell(\bar{\theta}_1^n, \mathrm{z}), \ell(\bar{\theta}_l^n, \mathrm{z}) - \ell(\bar{\theta}_1^n, \mathrm{z}))$	
$\rho_{k,l}$	asymptotic (k,l)-loss variability	$\lim_{n\to\infty} \rho_{k,l}^n$	
σ_k^2	asymptotic (k,k)-loss variability	$\rho_{k,k}$	
c_k	asymptotic ratio between gaps	$\lim_{n\to\infty} \delta_2^n/\delta_k^n$	
Λ	matrix	$(\rho_{k,l}c_k c_l)_{2\leq k,l\leq K}$	
$\bar{r}_{\mathcal{Q}}^n$	log rate of Theorem 3.4	$(\delta_2^n)^2/(2\max_k c_k^2\sigma_k^2)$	
$r_{\mathcal{Q}_n}$	MIS indicator	$\frac{1}{2}\min_{k>1} \frac{[\mathbb{E}[\ell(\bar{\theta}_k^n,\mathrm{z})] - \mathbb{E}[\ell(\bar{\theta}_1^n,\mathrm{z})]]^2}{\mathrm{Var}[\ell(\bar{\theta}_k^n,\mathrm{z}) - \ell(\bar{\theta}_1^n,\mathrm{z})]}$	

Table 1: Table of Notations

7 MIS is related to an f-divergence

We recall here the main theoretical result of [Aubinais et al., 2023].

Let ϕ be an MIA, and consider its accuracy $\mathrm{Acc}_n(\phi; P, \mathcal{A})$. Let $\hat{P}_n$ be the empirical distribution of the random variables $\mathrm{z}_1, \dots, \mathrm{z}_n$. We define a test point $\tilde{\mathrm{z}}$ as follows. Let $T \in \{0, 1\}$ be a Bernoulli random variable with parameter $\eta \in (0, 1)$. Let $\mathrm{z}_0 \sim P$ and U be a random variable whose distribution is $\hat{P}_n$ conditionally to $\mathrm{z}_1, \dots, \mathrm{z}_n$, where T, z_0 and U are independent. Additionally, z_0 and T are independent of $\mathrm{z}_1, \dots, \mathrm{z}_n$. We define the test sample by,

$$\tilde{\mathrm{z}} \coloneqq TU + (1 - T)\mathrm{z}_0. \tag{3}$$

This definition is a formalization of the fact that a test sample is either a member of the training dataset ($T = 1$) with probability $\mathbb{P}(T = 1) = \mathbb{P}(\tilde{\mathrm{z}} \text{ is member}) = \eta$, or not a member ($T = 0$). The random variable z_0 is used as a placeholder to represent the non-member property of the test point and η represents the theoretical fraction of members. Using Equation 3, the accuracy of ϕ can be rewritten as

$$\begin{aligned} \mathrm{Acc}_n(\phi; P, \mathcal{A}) &= \mathbb{P}(\phi(\hat{\theta}_n, \mathrm{z}_0) = 0, T = 0) + \lambda \mathbb{P}(\phi(\hat{\theta}_n, \mathrm{z}_1) = 1, T = 1) \\ &= (1 - \eta)\mathbb{P}(\phi(\hat{\theta}_n, \mathrm{z}_0) = 0) + \lambda\eta \mathbb{P}(\phi(\hat{\theta}_n, \mathrm{z}_1) = 1). \end{aligned} \tag{4}$$

Interestingly, constant MIAs $\phi_0 \equiv 0$ and $\phi_1 \equiv 1$ have respectively an accuracy of $1 - \eta$ and $\lambda\eta$, which means that any MIA whose accuracy is worse than $\max(1 - \eta, \lambda\eta)$ is irrelevant to use. Additionally, if a perfect MIA ϕ^* exists, then its accuracy would equal $1 - \eta + \lambda\eta$. This means that the supremum over all MIAs of the accuracy is bounded as

$$\max(1 - \eta, \lambda\eta) \leq \sup_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) \leq 1 - \eta + \lambda\eta. \tag{5}$$

From this equation, the MIS can be defined as follows,

Definition 7.1 (MIS [Aubinais et al., 2023])**.** The membership inference security of a learning procedure $\mathcal{A}$ is

$$\mathtt{MIS}_n(P, \mathcal{A}) \coloneqq \frac{1}{\min(1 - \eta, \lambda\eta)} \left(1 - \eta + \lambda\eta - \sup_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) \right), \tag{6}$$

where the supremum is taken over all MIAs.

Thus, $c_1 = (1 - \eta + \lambda\eta)/\min(1 - \eta, \lambda\eta)$ and $c_2 = (1 - \eta + \lambda\eta)^{-1}$ in equation 2.

For any $\alpha > 0$, and any distributions P and Q over the same probability space, we define

$$\begin{aligned} \tilde{D}_\alpha(P, Q) &\coloneqq \frac{1}{\alpha} \sup_B \left[\alpha P(B) - Q(B) \right] \\ D_\alpha(P, Q) &\coloneqq \max(1, \alpha) \left[\tilde{D}_\alpha(P, Q) - \left(1 - \frac{1}{\alpha} \right)_+ \right], \end{aligned}$$

where for any real number $x \in \mathbb{R}$, we have $x_+ = \max(0, x)$. Here, the supremum is taken over all measurable sets.

One of the main results in [Aubinais et al., 2023] is to prove that the MIS is related to an f-divergence between the joint distribution of $(\hat{\theta}_n, \mathrm{z}_1)$ and the product distribution of $\hat{\theta}_n$ and an independent sample z_0. The setting in [Aubinais et al., 2023] is more general and Theorem 7.2 below holds for any possibly randomized algorithm as soon as it is a (possibly randomized) function of $\hat{P}_n$.

Theorem 7.2 (Theorem 6 and Proposition 5 of [Aubinais et al., 2023])**.** *Let $\gamma \coloneqq \frac{1-\eta}{\lambda\eta}$. For any distribution P and any learning procedure $\mathcal{A}$, we have*

$$\texttt{MIS}_n(P,\mathcal{A}) = 1 - D_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_0)}, \mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_1)}\right), \tag{7}$$

where for any random variable x, $\mathbb{P}_{\mathrm{x}}$ *designates its distribution.*
Additionally, the map $(P,Q) \mapsto D_\gamma(P,Q)$ *is an* $f-$*divergence and for any random variables* x_1 *and* x_2 *with joint distribution* $\mathbb{P}_{(\mathrm{x}_1,\mathrm{x}_2)}$*, it holds that*

$$D_\gamma(\mathbb{P}_{\mathrm{x}_1} \otimes \mathbb{P}_{\mathrm{x}_2}, \mathbb{P}_{(\mathrm{x}_1,\mathrm{x}_2)}) = \mathbb{E}_{\mathrm{x}_2}\left[D_\gamma(\mathbb{P}_{\mathrm{x}_1}, \mathbb{P}_{\mathrm{x}_1|\mathrm{x}_2})\right], \tag{8}$$

where $\mathbb{P}_{\mathrm{x}_1|\mathrm{x}_2}$ *is the distribution of* x_1 *conditionally to* x_2*.*

Notice that in Equation 7, z_1 could be replaced by any z_j, for $j \in \{1,\dots,n\}$ since $\hat{\theta}_n$ is a function of $\hat{P}_n$.

We set

$$\Delta_{\eta,\lambda,n}(P,\mathcal{A}) \coloneqq D_\gamma\left(\mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_0)}, \mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_1)}\right). \tag{9}$$

Additionally, setting $p_{(\hat{\theta}_n,\mathrm{z}_0)}$ (resp. $p_{(\hat{\theta}_n,\mathrm{z}_1)}$) the density of $\mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_0)}$ (resp. $\mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_1)}$) with respect to a dominating measure ζ, $\Delta_{\lambda,\eta,n}(P,\mathcal{A})$ can be written in integral form as

$$\Delta_{\eta,\lambda,n}(P,\mathcal{A}) = \max(1,\gamma)\left[\frac{1}{2\gamma}\int |\gamma p_{(\hat{\theta}_n,\mathrm{z}_0)} - p_{(\hat{\theta}_n,\mathrm{z}_1)}| d\zeta - \frac{1}{2}\left|1-\frac{1}{\gamma}\right|\right] \tag{10}$$

$$= \mathbb{E}_{\mathrm{z}_1}\left[\max(1,\gamma)\left[\frac{1}{2\gamma}\int |\gamma p_{\hat{\theta}_n} - p_{\hat{\theta}_n|\mathrm{z}_1}| d\zeta - \frac{1}{2}\left|1-\frac{1}{\gamma}\right|\right]\right], \tag{11}$$

where $p_{\hat{\theta}_n}$ (resp. $p_{\hat{\theta}_n|\mathrm{z}_1}$) is the density of the marginal distribution (resp. conditional distribution given z_1) of $\hat{\theta}_n$.

8 Proof of Theorem 3.4

We introduce additional notations. For $i = 1,\dots,K_n$, we define $m_{\bar{\theta}_i^n} = \mathbb{E}\left(\ell(\bar{\theta}_k^n, \mathrm{z}_i)\right)$, so that $m_{\bar{\theta}_1^n} < \cdots < m_{\bar{\theta}_{K_n}^n}$. For $i = 1,\dots,K_n$ and $j = 1,\dots,n$, we define $L_{k,j}^n \coloneqq \ell(\bar{\theta}_k^n, \mathrm{z}_j) - \ell(\bar{\theta}_1^n, \mathrm{z}_j)$. We recall that the **loss gaps** are defined as $\delta_k^n = \mathbb{E}[L_{k,1}^n] = m_{\bar{\theta}_k^n} - m_{\bar{\theta}_1^n}$ and the **loss variabilities** as $(\sigma_k^n)^2 \coloneqq \mathrm{Var}\left(L_{k,1}^n\right)$. Finally we set $D^n \coloneqq \mathrm{diag}\left((\delta_k^n)_{k>1}\right)$, the $K_n \times K_n$-diagonal matrix with diagonal entries the δ_k^n, $k = 1,\dots,K_n$.

We first prove that the divergence D_α is dominated by the total variation distance.

Proposition 8.1. *For any* $\alpha > 0$*, and any distributions* P *and* Q *over the same probability space,*

$$D_\alpha(P,Q) \le D_1(P,Q) = \|P-Q\|_{TV}$$

where $\|P-Q\|_{TV}$ *is the total variation distance between* P *and* Q*.*

Proof of Proposition 8.1. Usual manipulations allow to prove that, if p (resp. q) is the density of P with respect to a dominating measure ζ, then

$$D_\alpha(P,Q) = \begin{cases} \int \left(p - \frac{1}{\alpha}q\right)_+ d\zeta & \text{if } \alpha \le 1 \\ \int (q - \alpha p)_+ d\zeta & \text{if } \alpha \ge 1. \end{cases}$$

But $\alpha \mapsto \left(p - \frac{1}{\alpha}q\right)$ is non decreasing, so that the sets $\left(p - \frac{1}{\alpha}q \ge 0\right)$ are non decreasing also, so that for all $\alpha \le 1$, $\int \left(p - \frac{1}{\alpha}q\right)_+ d\zeta \le \int (p-q)_+ d\zeta$. In the same way, $\alpha \mapsto (q - \alpha p)$ is non increasing, so that the sets $(q - \alpha p \ge 0)$ are non increasing also, so that for all $\alpha \ge 1$, $\int (q-\alpha p)_+ d\zeta \le \int (p-q)_+ d\zeta$. The conclusion follows.

□

Recall $\mathcal{Q}_n$ is a quantizer and $\hat{\theta}_n = \mathcal{A}_{\mathcal{Q}_n}(\mathrm{z}_1, \cdots, \mathrm{z}_n)$. Applying Theorem 7.2 and Proposition 8.1, we get

$$1 - \texttt{MIS}_n(P, \mathcal{A}) \leq \|\mathbb{P}_{(\hat{\theta}_n, \mathrm{z}_0)} - \mathbb{P}_{(\hat{\theta}_n, \mathrm{z}_1)}\|_{TV}, \tag{12}$$

and Theorem 3.4 follows from Proposition 8.2 and Proposition 8.3 below.

Proposition 8.2. *Under the assumptions of of Theorem 3.4, we have*

$$\limsup_{n\to\infty} \frac{1}{n\left(\delta_2^n\right)^2} \log \|\mathbb{P}_{(\hat{\theta}_n, \mathrm{z}_0)} - \mathbb{P}_{(\hat{\theta}_n, \mathrm{z}_1)}\|_{TV} \leq -\inf_{x\in\Omega^c} \frac{x^T \Lambda^{-1} x}{2}, \tag{13}$$

where Λ is defined as

$$\Lambda \coloneqq \begin{pmatrix} \rho_{2,2}c_2^2 & \cdots & \rho_{2,K}c_2c_K \\ \vdots & \ddots & \vdots \\ \rho_{K,2}c_Kc_2 & \cdots & \rho_{K,K}c_K^2 \end{pmatrix},$$

with $\Omega \coloneqq [-1, \infty)^{K-1}$.

Proposition 8.3. *We have*

$$\inf_{x\in\Omega^c} x^T \Lambda^{-1} x = \frac{1}{\max_k c_k^2 \sigma_k^2} \tag{14}$$

where we recall $\sigma_k^2 = \rho_{k,k}$.

Proof of Proposition 8.2. Letting $Z_k^n = \mathbb{P}\left(\hat{\theta}_n = \bar{\theta}_k^n \mid \mathrm{z}_1\right)$ and $p_k^n = \mathbb{P}\left(\hat{\theta}_n = \bar{\theta}_k^n\right)$, we have

$$\begin{aligned}
\|\mathbb{P}_{(\hat{\theta}_n, \mathrm{z}_0)} - \mathbb{P}_{(\hat{\theta}_n, \mathrm{z}_1)}\|_{TV} &= \frac{1}{2}\sum_{k=1}^K \mathbb{E}\left[|p_k^n - Z_k^n|\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[|p_1^n - Z_1^n|\right] + \frac{1}{2}\sum_{k=2}^K \mathbb{E}\left[p_k^n + Z_k^n\right] \\
&= \frac{1}{2}\mathbb{E}\left[|(1-p_1^n) - (1-Z_1^n)|\right] + \sum_{k=2}^K p_k^n \\
&\leq \frac{1}{2}\mathbb{E}\left[(1-p_1^n) + (1-Z_1^n)\right] + (1-p_1^n) \\
&= 2(1-p_1^n).
\end{aligned}$$

Recall that $\mathcal{A}_{\mathcal{Q}_n}$ minimizes the empirical loss. Letting $\Omega_{K-1} = [-1, \infty)^{K-1}$, we then have

$$\begin{aligned}
p_1^n &= \mathbb{P}\left(1 = \arg\min_k \left\{\frac{1}{n}\sum_{j=1}^n \ell\left(\bar{\theta}_k, \mathrm{z}_j\right)\right\}\right) \\
&= \mathbb{P}\left(\forall k > 1, \frac{1}{n}\sum_{j=1}^n \left[\ell\left(\bar{\theta}_k, \mathrm{z}_j\right) - \ell\left(\bar{\theta}_1, \mathrm{z}_j\right)\right] \geq 0\right) \\
&= \mathbb{P}\left(\forall k > 1, \frac{1}{n}\sum_{j=1}^n \left[L_{k,j} - \delta_k\right] \geq -\delta_k\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n \left[L_{k,j} - \delta_k\right]_{k>1} \in D\Omega_{K-1}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n D^{-1}\left[L_{k,j} - \delta_k\right]_{k>1} \in \Omega_{K-1}\right),
\end{aligned} \tag{15}$$

which gives

$$1-p_1^n = \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} D^{-1}\left[L_{k,j}-\delta_k\right]_{k>1} \in \Omega_{K-1}^c\right),$$

so that

$$\begin{aligned}\|\mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_0)} - \mathbb{P}_{(\hat{\theta}_n,\mathrm{z}_1)}\|_{TV} &\leq 2\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\left[D^n\right]^{-1}\left[L_{k,j}^n-\delta_k^n\right]_{k>1} \in \Omega_{K-1}^c\right)\\ &= 2\,\underbrace{\mathbb{P}\left(\frac{1}{n\delta_2^n}\sum_{j=1}^{n}\left[\left(\delta_2^n\right)^{-1}D^n\right]^{-1}\left[L_{k,j}^n-\delta_k^n\right]_{k>1} \in \Omega_{K-1}^c\right)}_{1-p_1^n}.\end{aligned}$$

By **(A2)**, $\left[\left(\delta_2^n\right)^{-1}D^n\right]^{-1}\left[L_{k,j}^n-\delta_k^n\right]_{k>1}$ lives (asymptotically) in a $(K-1)$-dimensional euclidean subspace. By **(A3)** and **(A4)**, as we live in an Hilbert space, using [Araujo and Giné, 1980], we have $\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[\left(\delta_2^n\right)^{-1}D^n\right]^{-1}\left[L_{k,j}^n-\delta_k^n\right]_{k>1}\right) \underset{n\to\infty}{\to} \gamma \coloneqq \mathcal{N}_{K-1}(0,\Lambda)$, where $\mathcal{N}_{K-1}$ is the $(K-1)$-dimensional Gaussian distribution. Now, using the fact that $\mathbb{E}\left[L_{k,j}^n\right]=\delta_k^n$, using **(A3)** and **(A4)** and the convergence to a Gaussian measure, we apply Theorem 2.2 of [De Acosta, 1992] to get

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n\left(\delta_2^n\right)^2}\log\left(1-p_1^n\right) &\leq -\inf_{x\in\Omega_{K-1}^c}\begin{cases}\frac{x^T\Lambda^{-1}x}{2} & \text{, if } x\in H_\gamma\\ \infty & \text{, otherwise}\end{cases}\\ &= -\inf_{x\in\Omega^c}\frac{x^T\Lambda^{-1}x}{2},\end{aligned}$$

where H_γ is the Hilbert space associated with γ (see [De Acosta, 1992]). Additionally, as $\sqrt{n}\delta_2^n \underset{n\to\infty}{\to} \infty$ by **(A1)**, we have $\frac{1}{n(\delta_2^n)^2}\log 2 \underset{n\to\infty}{\to} 0$. □

Proof of Proposition 8.3. Let $M=\Lambda^{-1}$. Note that $\Omega^c=\{x\in\mathbb{R}^{K-1} : \exists j, x_j<-1\}$, giving

$$\inf_{x\in\Omega^c} x^TMx = \min_j \inf_{x_{-j}\in\mathbb{R}^{K-1}}\inf_{x_j<-1} x^TMx,$$

where we write x_{-j} equals x where we omit its j^{th} entry. We then shall write

$$x^TMx = x_j^2M_{j,j} + 2x_jx_{-j}^TM_{-j,j} + x_{-j}^TM_{-j,-j}x_{-j}$$

where $M_{-j,-j}$ is the sub-matrix of M consisting of all entries except the j^{th} row and column and $M_{-j,j}$ is the j^{th} column of M except its j^{th} entry.

The infimum must be reached on the frontier of the set, i.e. such that $x_j=-1$ for some j. Indeed, assuming x in the interior of Ω^c, we have that $x_j<-1$ for some j. Then for any $1<\alpha<|x_j|$, $x_j/\alpha<-1$ meaning that x/α still belongs to the interior of Ω^c. However, $(x/\alpha)^TM(x/\alpha)=\frac{1}{\alpha^2}x^TMx<x^TMx$, which shows that x was not optimal. For an optimal x, we then have

$$x^TMx = M_{j,j} - 2x_{-j}^TM_{-j,j} + x_{-j}^TM_{-j,-j}x_{-j}.$$

It is then sufficient to study the optimization problem over x_{-j}, which amounts down to the optimization of a quadratic function, whose minimum is then reached for x_{-j} satisfying

$$\begin{aligned}\nabla_{x_{-j}}\left(-2x_{-j}^T M_{-j,j} + x_{-j}^T M_{-j,-j} x_{-j}\right) = 0 \\ \iff x_{-j} = \left(M_{-j,-j}\right)^{-1} M_{-j,j},\end{aligned}$$

giving

$$\inf_{x_{-j}\in\mathbb{R}^{J-1}} \inf_{x_j<-1} x^T M x = M_{j,j} - M_{j,-j}\left(M_{-j,-j}\right)^{-1} M_{-j,j}.$$

Applying Lemma 8.4 concludes the proof. □

Lemma 8.4. *Let M be a $J \times J$ square matrix and $j \in \{1,\dots,J\}$. Assume that $M_{-j,-j}$ and $M_{j,j} - M_{j,-j}\left[M_{-j,-j}\right]^{-1} M_{-j,j}$ are invertible. Then we have*

$$\left(M^{-1}\right)_{j,j} = \left(M_{j,j} - M_{j,-j}\left[M_{-j,-j}\right]^{-1} M_{-j,j}\right)^{-1}.$$

Proof of Lemma 8.4. First note that if M is designed by block as follows,

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

then we have

$$M^{-1} = \begin{pmatrix} \left(A - BD^{-1}C\right)^{-1} & * \\ * & * \end{pmatrix}, \tag{16}$$

as long as C and $A - BD^{-1}C$ are invertible. Let P_j and Q_j be $J \times J$ matrices defined for all j such that $P_j M$ permutes the first and the j^{th} rows of M, and $Q_j M$ permutes the $(j-1)^{th}$ and the j^{th} rows of M. For instance, if $J = 4$ and $j = 3$ then we have

$$P_3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad Q_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that we have $(P_j)^{-1} = P_j$ and $(Q_j)^{-1} = Q_j$. Let $R = Q_3 Q_4 \cdots Q_j P_j M P_j Q_j \cdots Q_4 Q_3$. Developing the formula, we get

$$R = \begin{pmatrix} M_{j,j} & M_{j,-j} \\ M_{-j,j} & M_{-j,-j} \end{pmatrix}. \tag{17}$$

On one hand, using Equation 16 and Equation 17, we have $R^{-1} = \begin{pmatrix} \left(M_{j,j} - M_{j,-j}\left[M_{-j,-j}\right]^{-1} M_{-j,j}\right)^{-1} & \star \\ \star & \star \end{pmatrix}$.
On the other hand, distributing the inverse operator on the product of matrices and using again Equation 16, we have

$$R^{-1} = Q_3 Q_4 \cdots Q_j P_j M^{-1} P_j Q_j \cdots Q_4 Q_3$$
$$= \begin{pmatrix} \left(M^{-1}\right)_{j,j} & \star \\ \star & \star \end{pmatrix},$$

which concludes the proof. □

9 Complete Algorithm

We provide below (2) the complete algorithm used to estimate $r_{\mathcal{Q}_n}$, incorporating the learning procedure $\mathcal{A}$ and the quantization function $\mathcal{Q}$.

Algorithm 2 Estimation of $r_{\mathcal{Q}_n}$

Input: A training dataset $\mathcal{D}_n$, a validation dataset $\mathcal{D}_{\text{val}}$, a learning procedure $\mathcal{A}$, a quantizer $\mathcal{Q}$, an initialized model θ, a number of epochs K.
Output: An estimate of $r_{\mathcal{Q}_n}$.
Initialization: Set the list of all quantized loss $\mathcal{L}_{\text{val}} = \{\}$, of all average quantized losses $m_{\text{val}} = \{\}$, and of all variances $\sigma^2_{\text{val}} = \{\}$.
for $k = 1$ to K **do**
  $\theta \leftarrow \mathcal{A}(\theta, \mathcal{D}_n)$.
  $\bar{\theta} \leftarrow \mathcal{Q}(\theta)$.
  $\mathcal{L}_k \leftarrow \{\ell(\bar{\theta}, \mathrm{z}) : \mathrm{z} \in \mathcal{D}_{\text{val}}\}$.
  $m_k \leftarrow \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{\mathrm{z} \in \mathcal{D}_{\text{val}}} \ell(\bar{\theta}, \mathrm{z})$.
  $\mathcal{L}_{\text{val}}[k] \leftarrow \mathcal{L}_k$.
end for
idx $\leftarrow$ argsort(m_{val}).
$m_{\text{val}} \leftarrow m_{\text{val}}$[idx].
$\mathcal{L}_{\text{val}} \leftarrow \mathcal{L}_{\text{val}}$[idx].
for $k = 2$ to K **do**
  $\sigma^2_{\text{val}}[k] \leftarrow \text{Var}\left(\mathcal{L}_{\text{val}}[k] - \mathcal{L}_{\text{val}}[1]\right)$.
end for
$r_{\mathcal{Q}} \leftarrow \frac{1}{2}\left[\max_{2 \le k \le K}\left(\sigma^2_{\text{val}}[k] \times \left(\frac{m_{\text{val}}[2] - m_{\text{val}}[1]}{m_{\text{val}}[k] - m_{\text{val}}[1]}\right)^2\right)\right]^{-1}$.
return $r_{\mathcal{Q}_n} = r_{\mathcal{Q}} \left(m_{\text{val}}[2] - m_{\text{val}}[1]\right)^2$.

10 Quantizations for the Synthetic and Molecular Experiments

For our synthetic and molecular experiments, we used simple quantizers to facilitate the analysis of the results. These quantizers are summarized in Table 2, and Figure 6 illustrates how the different functions used quantize the interval $[-1, 1]$.

11 Baseline estimation of the MIS

We describe here the baseline approach used to estimate the membership inference score (MIS) of a model $\hat{\theta}_n$.

Inferring membership of $\tilde{\mathrm{z}}$ by an MIA can naturally be considered as a statistical test,

$$\begin{cases} H_0 : & \text{“}\tilde{\mathrm{z}} \text{ belongs to the training dataset”.} \\ H_1 : & \text{“}\tilde{\mathrm{z}} \text{ does not belong to the training dataset”.} \end{cases}$$

Under the mathematical setting presented before, these hypotheses can be apprehended as deciding whether $\hat{\theta}_n$ is

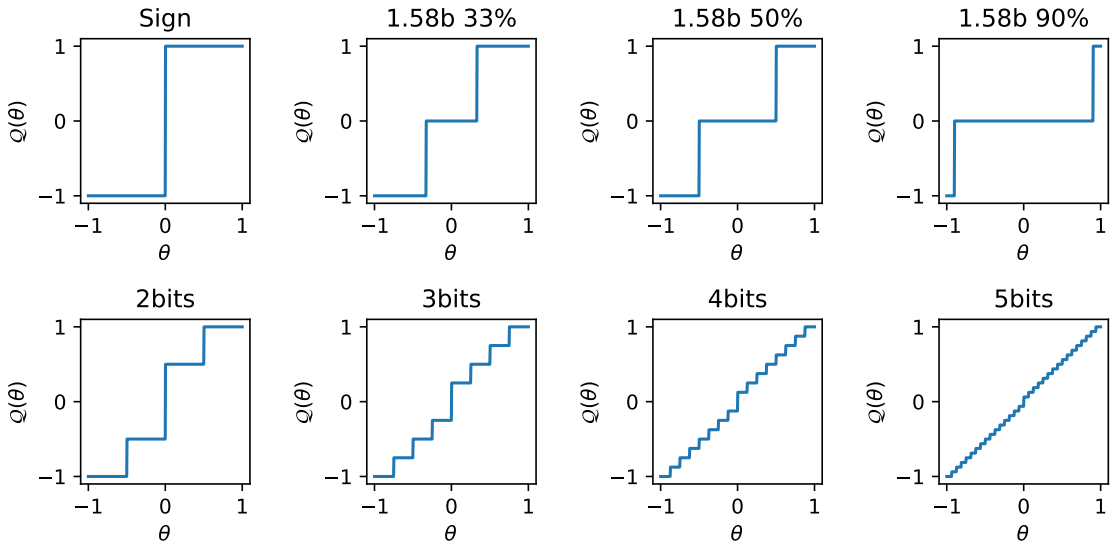


Figure 6: Illustration of the quantization functions used on the interval $[-1, 1]$.

Table 2: Description of the quantizers used in the experiments.

quantizer	$\mathcal{Q}(\theta_j)$
Sign	$\mathcal{Q}(\theta_j) = \frac{\theta_j}{\lvert\theta_j\rvert}$
1.58b 33%	$\mathcal{Q}(\theta_j) = \begin{cases} \frac{\theta_j}{\lvert\theta_j\rvert}, & \text{if } \lvert\theta_j\rvert < q(\lvert\theta\rvert, 0.33) \\ 0, & \text{otherwise} \end{cases}$
1.58b 50%	$\mathcal{Q}(\theta_j) = \begin{cases} \frac{\theta_j}{\lvert\theta_j\rvert}, & \text{if } \lvert\theta_j\rvert < q(\lvert\theta\rvert, 0.5) \\ 0, & \text{otherwise} \end{cases}$
1.58b 90%	$\mathcal{Q}(\theta_j) = \begin{cases} \frac{\theta_j}{\lvert\theta_j\rvert}, & \text{if } \lvert\theta_j\rvert < q(\lvert\theta\rvert, 0.9) \\ 0, & \text{otherwise} \end{cases}$
2 bits	$\mathcal{Q}(\theta_j) = \frac{\theta_j}{\lvert\theta_j\rvert}\frac{\alpha}{2} \times \text{int}\left(1 + \text{clip}(\frac{2\theta_j}{\alpha}, 0, 2)\right), \quad \alpha = 2^{\text{round}(\log_2(\max\lvert\theta\rvert))}$
3 bit	$\mathcal{Q}(\theta_j) = \frac{\theta_j}{\lvert\theta_j\rvert}\frac{\alpha}{4} \times \text{int}\left(1 + \text{clip}(\frac{4\theta_j}{\alpha}, 0, 4)\right), \quad \alpha = 2^{\text{round}(\log_2(\max\lvert\theta\rvert))}$
4 bits	$\mathcal{Q}(\theta_j) = \frac{\theta_j}{\lvert\theta_j\rvert}\frac{\alpha}{8} \times \text{int}\left(1 + \text{clip}(\frac{8\theta_j}{\alpha}, 0, 8)\right), \quad \alpha = 2^{\text{round}(\log_2(\max\lvert\theta\rvert))}$
5 bits	$\mathcal{Q}(\theta_j) = \frac{\theta_j}{\lvert\theta_j\rvert}\frac{\alpha}{16} \times \text{int}\left(1 + \text{clip}(\frac{16\theta_j}{\alpha}, 0, 16)\right), \quad \alpha = 2^{\text{round}(\log_2(\max\lvert\theta\rvert))}$

independent or not to $\tilde{z}$. Specifically, testing H_0 against H_1 is equivalent to testing H_0' against H_1', where

$$\begin{cases} H_0': & (\hat{\theta}_n, \tilde{z}) \sim P_{\left(\hat{\theta}_n, z_1\right)}. \\ H_1': & (\hat{\theta}_n, \tilde{z}) \sim P_{\hat{\theta}_n} \otimes P. \end{cases} \tag{18}$$

For any dominating measure ζ on $P_{\left(\hat{\theta}_n, z_1\right)}$ (and $P_{\hat{\theta}_n} \otimes P$), denoting by f (resp. g) the density of $P_{\left(\hat{\theta}_n, z_1\right)}$ (resp. $P_{\hat{\theta}_n} \otimes P$) with respect to ζ, we have that $\phi^*(\theta, z) = \mathbb{1}\left\{\frac{f(\theta,z)}{g(\theta,z)} \geq 1\right\}$ satisfies:

$$\sup_{\phi} \mathrm{Acc}_n(\phi; P, \mathcal{A}) = \mathrm{Acc}_n(\phi^*; P, \mathcal{A}). \tag{19}$$

The function ϕ^* is the Neyman-Pearson test for H_0' against H_1'. This suggests that evaluating empirically the MIS of an algorithm amounts to training a discriminator and evaluating it. The baseline approach therefore consists in training a binary classifier $g_\psi : \Theta \times \mathbb{R}^d \times \mathbb{R} \to [0, 1]$ to distinguish between samples z sampled from the training set of a given $\hat{\theta}_n$ and sampled from the product distribution $P_{\hat{\theta}_n} \otimes P$, and evaluate its accuracy. For a sample $z = (x, y) \sim P$, where $x \in \mathbb{R}^d$ is the input and y its corresponding label, and a model $\hat{\theta}_n$, the discriminator minimizes the binary cross-entropy loss:

$$\ell_{\mathrm{DISC}}(\hat{\theta}_n, z) = \mathrm{BinaryCE}\left(g_\psi(\hat{\theta}_n, x, y), \mathbb{1}_{z \in \mathcal{D}_{\mathrm{train}}(\hat{\theta}_n)}\right),$$

where $\mathcal{D}_{\mathrm{train}}(\hat{\theta}_n)$ is the training set of $\hat{\theta}_n$.

The discriminator is implemented as a feed-forward neural network that takes as input: x, the input data; the flattened parameters of $\hat{\theta}_n$; and the loss of the model $\hat{\theta}_n$ on x and y. For instance, if the model $\hat{\theta}_n$ is a binary classifier, g_ψ is a neural network with input $x \in \mathbb{R}^d$, $\hat{\theta}_n$ and the binary cross-entropy loss of $\hat{\theta}_n$ on (x, y).

We train the discriminator using a set of models, $\{\hat{\theta}_{ni}\}_{i \in \{1,\dots,k_{\mathrm{run}}\}}$ where each $\hat{\theta}_{ni}$ is trained on an independent dataset $\{z_{i,1}, \dots, z_{i,n}\} \sim P$ of n i.i.d. samples. To generate negative samples, we independently sample additional sets, $z_{i,1}^{\mathrm{neg}}, \dots, z_{i,n}^{\mathrm{neg}} \sim P$, ensuring no overlap with the training sets of any models, or between the negative samples of different models. This independence between datasets is crucial for preventing information leakage, but it may be restrictive in practice due to the high data requirements.

A key drawback of this approach arises when $\hat{\theta}_n$ contains multiple layers. Different permutations of the weights of the hidden units can represent identical models, but the discriminator g_ψ, which relies on flattened parameters, is not invariant to such permutations.

12 Synthetic experiments

12.1 Training curves

We provide in Figure 7 the training curves of the classifiers trained on the synthetic datasets.

12.2 Runtime comparison with the baseline approach

As explained in section 11, the baseline approach consists of training a discriminator to distinguish between samples from the training set of a given $\hat{\theta}_n$ and samples from the product distribution $P_{\hat{\theta}_n} \otimes P$. Similarly, the $r_{\mathcal{Q}_n}$-based approach relies on the training of multiple models to average the values of $r_{\mathcal{Q}_n}$ obtained.

The computational overhead induced by the $r_{\mathcal{Q}_n}$-based approach, specifically computing the validation loss of the quantized models, is negligible compared to the total training time (1 second against 4 minutes). Similarly, the training of the discriminator takes only about 40 minutes (all experiments were conducted on NVIDIA A6000 GPUs).

As a result, training multiple models $\hat{\theta}_n$ over multiple runs is the computational bottleneck of our privacy evaluations. To properly evaluate the time required to obtain both rankings, one would have to answer the following question: 'How many runs do I need to launch to ensure the ranking I obtained is stable?'

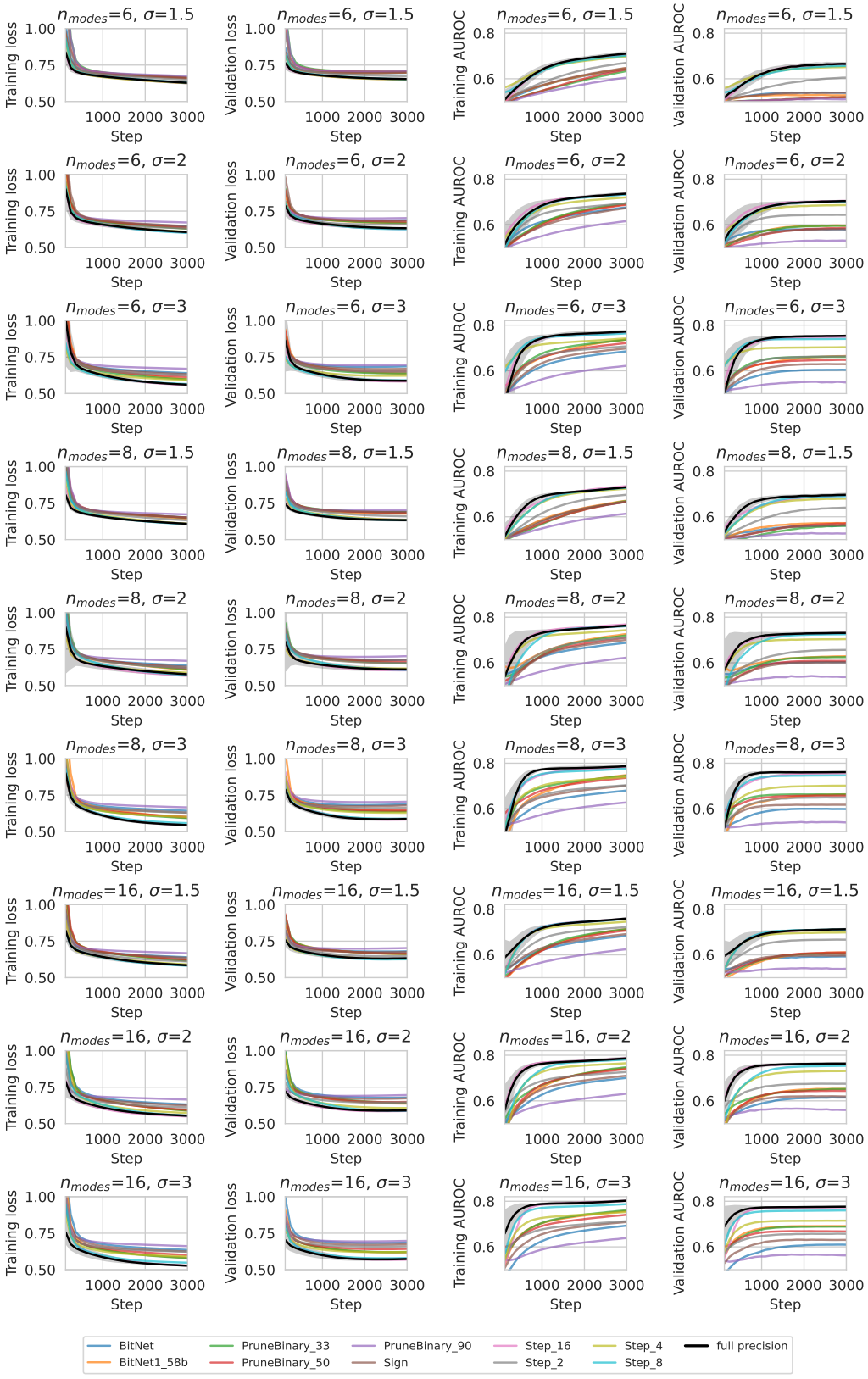
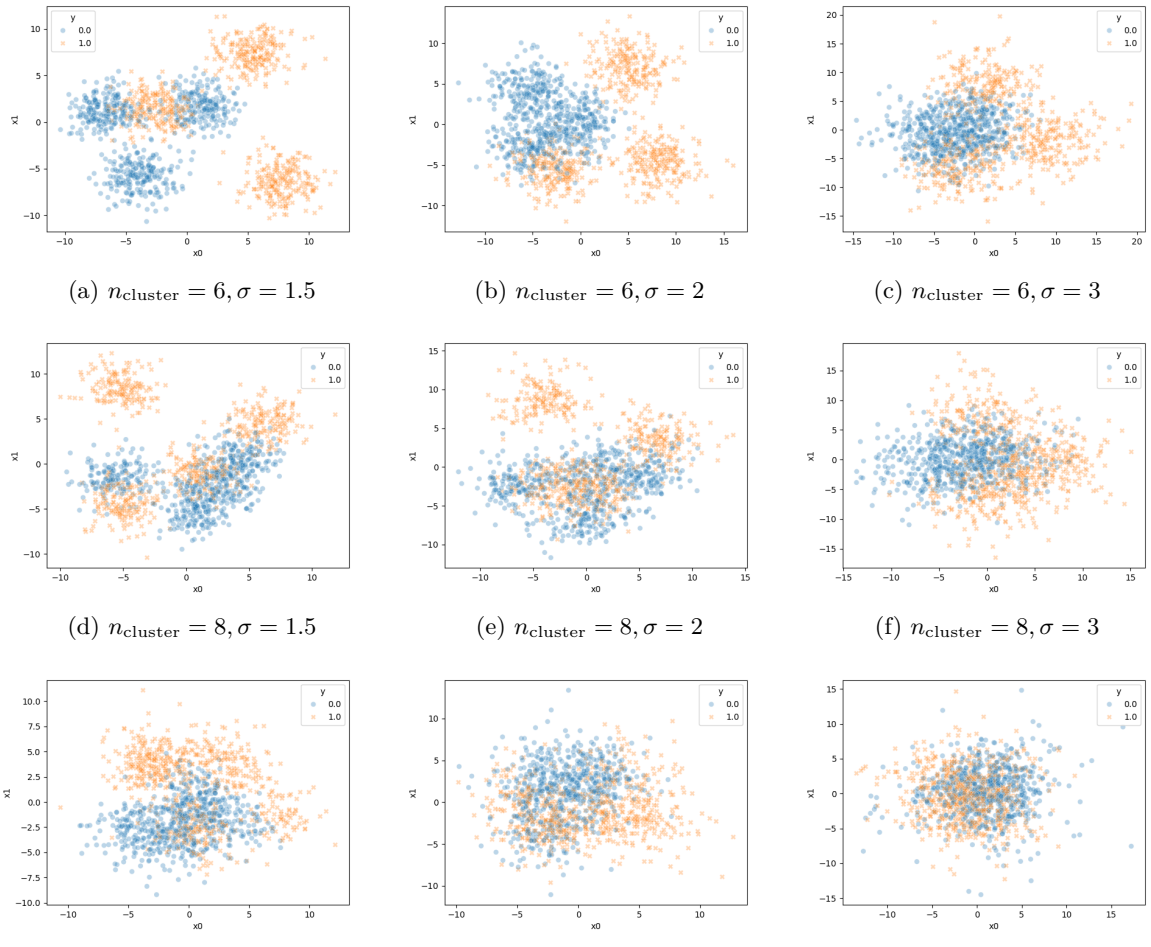


Figure 7: Training curves of the models trained on synthetic datasets, with the corresponding quantized metrics

As discussed in the previous section, after 15 runs, the rankings obtained with $r_{\mathcal{Q}_n}$ are already highly correlated with those obtained with 300 runs. In contrast, the rankings obtained with the baseline MIS method require 150 runs to reach the same level of correlation. As a result, the time required to obtain stable rankings with $r_{\mathcal{Q}_n}$ is significantly lower ($\approx$ 1h) than with the baseline MIS method ($\approx$ 10h).

12.3 Visualization of the datasets

We provide in Figure 8 a visualization of the synthetic datasets used in the experiments, through a PCA projection in dimension 2. This visualization helps understand how different data distribution might result in different empirical results, as some datasets are more challenging than others, such as the dataset with $n_{\text{cluster}} = 6$ and $\sigma = 1.5$, for whom the labels of the datapoints are easily separable, while $n_{\text{cluster}} = 16$ and $\sigma = 3$ provides a more challenging dataset, with overlapping clusters.



(a) $n_{\text{cluster}} = 6, \sigma = 1.5$ (b) $n_{\text{cluster}} = 6, \sigma = 2$ (c) $n_{\text{cluster}} = 6, \sigma = 3$

(d) $n_{\text{cluster}} = 8, \sigma = 1.5$ (e) $n_{\text{cluster}} = 8, \sigma = 2$ (f) $n_{\text{cluster}} = 8, \sigma = 3$

(g) $n_{\text{cluster}} = 16, \sigma = 1.5$ (h) $n_{\text{cluster}} = 16, \sigma = 2$ (i) $n_{\text{cluster}} = 16, \sigma = 3$

Figure 8: Visualization of the synthetic datasets used in the experiments, through a PCA projection in dimension 2 (the original space is $\mathbb{R}^{128}$).

13 Text Experiments

13.1 Training curves

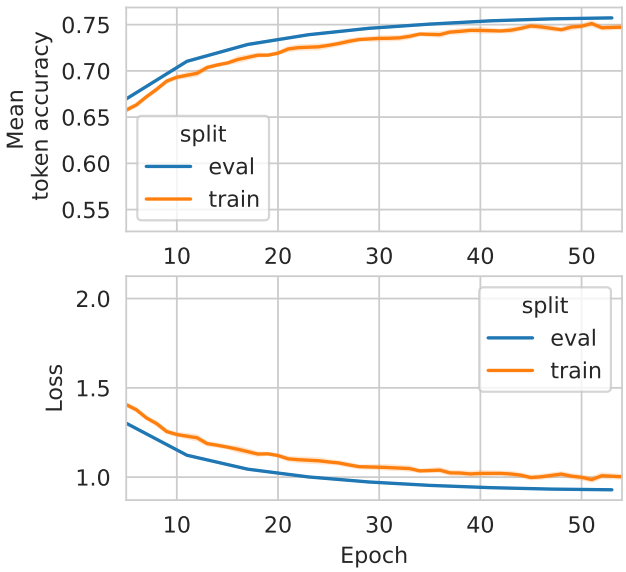


Figure 9: Training curves of the language models fine-tuned on SMolInstruct with a LORA adapter of rank 64.

We provide the training curves on the language models fine-tuned for our NLP experiments in Figure 9.

13.2 Black-Box Attacks

We used various black-box MIAs for LLMs on our language models. Each of them defines the probability of being a member of the model's training set as described below.

1. LOSS [Yeom et al., 2018]: where the prediction of the MIA solely relies on the loss of the model ($P(\text{member}) \propto -\mathcal{L}_\theta(x)$).
2. min-k% [Shi et al., 2024]: where the prediction of the MIA solely relies on the loss of 20% tokens predicted with the lowest probability ($P(\text{member}) \propto -\min_{J \subset \{1,..,N\}, |J|=\text{int}(|x|*0.2)} \sum_{j \in J} \mathcal{L}_\theta(x_i)$).
3. LOSS + zlib [Carlini et al., 2021]: where the loss is normalized with the size of the zlib compression of the text ($P(\text{member}) \propto -\mathcal{L}_\theta(x)/|zlib(x)|$).
4. LOSS + ref [Carlini et al., 2021]: where the loss of the model we attack is normalized by the loss of reference models (the ones obtained from the other runs) ($P(\text{member}) \propto -(\mathcal{L}_\theta(x) - \sum_i \mathcal{L}_{\theta_{ref_i}}(x))$)

Finally, these attacks can be combined (which we denote as LOSS + zlib + ref, for example). In practice, we found that most attacks performed no better than a random attack, an observation consistent with previous works on black-box attacks with LLMs [Duan et al., 2024]. The only competitive attacks are those using the pool of reference models, which are the most expensive ones. We believe this is because our models did not overfit on their training set, making the direct observation of their loss insufficient to evaluate the membership of a sample.

13.3 Additional Results

We report in Figure 10 the dependence of the quantized security $r_{\mathcal{Q}_n}$ with the best metric achieved by an MIA for each quantizer. We provide results for: AUROC, average precision, TPR at 10, 5, and 1% FPR. Overall, our metric correlates with all the attacks, including the reference models, and we observe that the quantized security

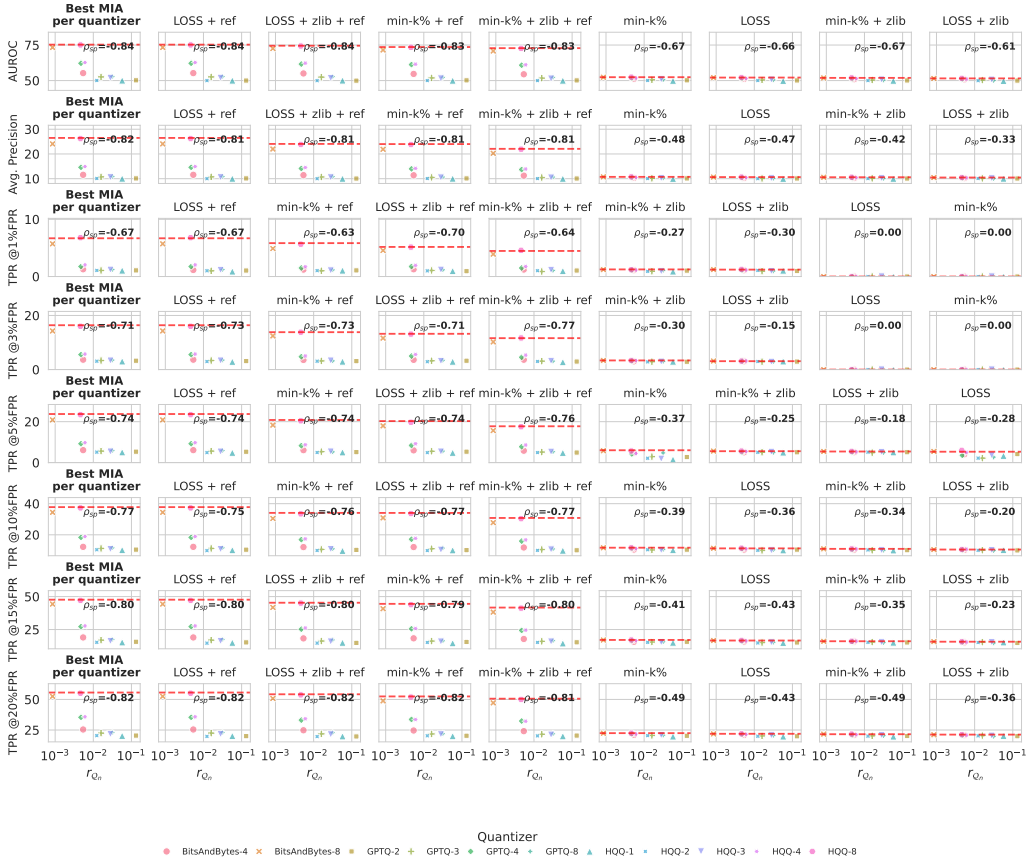


Figure 10: Quantized security against the performance of various MIAs for LLMs. The first figure displays the best metric achieved by an MIA for each quantizer against the quantized security $r_{\mathcal{Q}_n}$. We report the average-precision, auroc, and TPR at 20, 15, 10, 5, 3, and 1% FPR.

MINISTRAL-7B:

attack	average_precision		roc_auc		tpr@10		tpr@15	
	Spearman corr.	Avg perf.	Spearman corr.	Avg perf.	Spearman corr.	Avg perf.	Spearman corr.	Avg perf.
LOSS	-0.63	22.75	-0.63	54.69	-0.65	12.78	-0.64	18.67
LOSS + ref	**-0.77**	**39.84**	**-0.77**	**69.18**	**-0.76**	**32.00**	**-0.76**	**40.22**
LOSS + zlib	-0.59	22.17	-0.63	53.53	-0.61	11.98	-0.57	17.51
LOSS + zlib + ref	-0.77	38.39	-0.76	68.84	-0.76	30.79	-0.75	39.35
min-k%	-0.65	22.99	-0.65	55.26	-0.66	12.84	-0.67	18.88
min-k% + ref	-0.77	39.33	-0.77	69.01	-0.77	31.70	-0.76	40.12

LLAMA3-8B

attack	average_precision		roc_auc		tpr@10		tpr@15	
	Spearman corr.	Avg perf.	Spearman corr.	Avg perf.	Spearman corr.	Avg perf.	Spearman corr.	Avg perf.
LOSS	-0.73	21.9	-0.72	53.55	-0.77	12.16	-0.75	17.69
LOSS + ref	**-0.85**	**39.12**	**-0.87**	**68.55**	**-0.89**	**32.35**	**-0.87**	**40.31**
LOSS + zlib	-0.73	21.45	-0.73	52.83	-0.76	11.39	-0.77	17.13
LOSS + zlib + ref	-0.84	37.82	-0.87	68.2	-0.87	30.92	-0.87	39.61
min-k%	-0.71	22.15	-0.7	54.11	-0.74	12.35	-0.72	18.07
min-k% + ref	-0.84	38.52	-0.85	68.28	-0.85	31.71	-0.84	39.87

Figure 11: Average correlation between the quantized security $r_{\mathcal{Q}_n}$ and the efficiency of various MIAs for LLMs. The first table displays the result for MINISTRAL-7B and the second table for LLAMA3-8B. We report for the various MIAs, the correlation, averaged over different quantizers, between $r_{\mathcal{Q}_n}$ and the precision, auroc, TPR at 15 and 10 % FPR.

$r_{\mathcal{Q}_n}$ is a good indicator of the quantization robustness of the learning procedure against MIAs. For other attacks, $r_{\mathcal{Q}_n}$ does not seem to correlate with their performances, which can be explained by the fact that these attacks perform close to random guesses.

14 Molecular experiments

14.1 Experimental setup and Comprehensive results

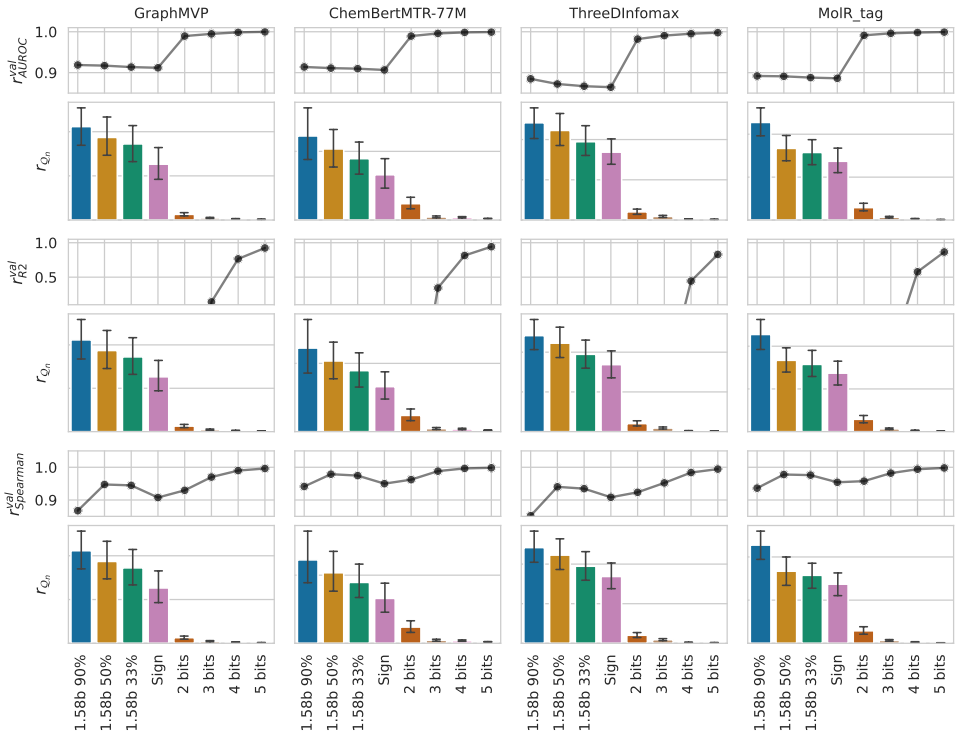


Figure 12: **Impact of quantization on classification tasks.** Evolution of the privacy of each downstream model $r_{\mathcal{Q}_n}$ along with relative performances of the quantized models compared to the original on classification/regression tasks.

We show in Table 3, Table 4, and Figure 12 the comprehensive results of the quantized models on the classification and regression tasks, respectively.

We observe that while the quantized models are generally less accurate than the original models, they still achieve reasonable performance on the classification tasks. In regression examples, the performance of quantized models is significantly lower than that of the original models. In particular, when the quantization is on less than 4 bits, the prediction of the molecular properties is almost consistently lower than a simple mean prediction. As explained in subsection 4.3, this result is expected, as while classification tasks rely on the definition of the boundary between classes, regression tasks require a fine-grained prediction of the target value. However, while the direct predictions of the quantized models do not provide a good estimate of the target value, the ordering of the predictions is still preserved, as shown by the Spearman correlation between the quantized models' predictions and the labels in Table 5 and Figure 12.

14.2 Details on the evaluation of the quantizers' privacy

Our hypothesis in the estimation of $r_{\mathcal{Q}_n}$ is that the quantized weights with the lowest average loss dominate the maximum value of $\lambda_k = \Lambda_{k,k} = \lim_{n\to\infty} (\delta_2^n/\delta_k^n)^2 (\sigma_k^n)^2$. We show in Figure 13 the evolution of $\Lambda_{k,k}$ with the k, where the indexes are sorted with decreasing values of average loss, and the histogram of the index of the

Table 3: AUROC performance of the quantized models on the classification tasks, averaged over all embedders.

dataset	Sign	1.58b 33%	1.58b 50%	1.58b 90%	2 bits	3 bits	4 bits	5 bits	original
AMES	0.773 (89%)	0.781 (90%)	0.784 (90%)	0.748 (86%)	0.843 (97%)	0.853 (98%)	0.859 (99%)	**0.861 (99%)**	**0.862 (100%)**
BBB Martins	0.800 (89%)	0.803 (89%)	0.807 (90%)	0.804 (89%)	0.890 (99%)	**0.895 (99%)**	**0.895 (99%)**	**0.896 (100%)**	**0.896 (100%)**
Bioavailability Ma	0.586 (94%)	0.588 (94%)	0.590 (95%)	0.579 (93%)	0.619 (99%)	**0.622 (100%)**	**0.622 (100%)**	**0.623 (100%)**	**0.622 (100%)**
CYP2C9 Substrate CarbonMangels	0.558 (86%)	0.557 (86%)	0.561 (86%)	0.589 (90%)	0.642 (99%)	0.646 (99%)	**0.647 (99%)**	**0.647 (99%)**	**0.648 (100%)**
CYP2C9 Veith	0.787 (89%)	0.793 (90%)	0.799 (91%)	0.827 (94%)	0.868 (99%)	0.873 (99%)	**0.876 (99%)**	**0.877 (99%)**	**0.877 (100%)**
Carcinogens Lagunin	0.766 (91%)	0.765 (91%)	0.772 (92%)	0.791 (94%)	0.824 (98%)	0.830 (99%)	0.831 (99%)	**0.832 (99%)**	**0.833 (100%)**
ClinTox	0.561 (80%)	0.555 (79%)	0.557 (79%)	0.589 (84%)	0.698 (98%)	0.702 (99%)	0.704 (99%)	**0.706 (99%)**	**0.707 (100%)**
DILI	0.827 (92%)	0.830 (93%)	0.831 (93%)	0.843 (94%)	0.888 (99%)	**0.891 (99%)**	**0.892 (99%)**	**0.892 (99%)**	**0.892 (100%)**
HIA Hou	0.805 (91%)	0.805 (91%)	0.804 (91%)	0.790 (89%)	**0.882 (99%)**	**0.883 (99%)**	**0.883 (99%)**	**0.883 (100%)**	**0.883 (100%)**
PAMPA NCATS	0.585 (81%)	0.583 (81%)	0.584 (81%)	0.583 (81%)	0.708 (99%)	0.711 (99%)	**0.713 (99%)**	**0.714 (99%)**	**0.714 (100%)**
Pgp Broccatelli	0.856 (92%)	0.858 (92%)	0.859 (92%)	0.864 (93%)	0.921 (99%)	**0.924 (99%)**	**0.924 (100%)**	**0.924 (100%)**	**0.925 (100%)**
Skin Reaction	0.664 (89%)	0.667 (89%)	0.668 (90%)	0.670 (90%)	0.735 (99%)	0.740 (99%)	0.741 (99%)	**0.742 (99%)**	**0.743 (100%)**
hERG	0.728 (91%)	0.726 (91%)	0.726 (91%)	0.739 (92%)	0.793 (99%)	0.795 (99%)	**0.796 (99%)**	**0.797 (99%)**	**0.797 (100%)**
hERG (k)	0.767 (88%)	0.783 (90%)	0.789 (91%)	0.757 (87%)	0.819 (94%)	0.837 (96%)	0.855 (98%)	**0.862 (99%)**	**0.866 (100%)**

Table 4: R2 performance of the quantized models on the regression tasks, averaged over all embedders. If the R2 score is less than -1, we display -inf for clarity.

dataset	Sign	1.58b 33%	1.58b 50%	1.58b 90%	2 bits	3 bits	4 bits	5 bits	original
Caco2 Wang	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-364%)	0.008 (1%)	0.458 (75%)	**0.567 (93%)**	**0.609 (100%)**
HydrationFreeEnergy FreeSolv	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-0.466 (-70%)	0.412 (55%)	0.669 (91%)	**0.715 (98%)**	**0.725 (100%)**
LD50 Zhu	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-578%)	-0.129 (-25%)	0.339 (67%)	**0.454 (89%)**	**0.505 (100%)**
Lipophilicity (az)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-539%)	-0.037 (-7%)	0.404 (72%)	**0.508 (91%)**	**0.552 (100%)**
PPBR AZ	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-0.480 (-233%)	0.022 (2%)	**0.156 (65%)**	**0.229 (100%)**
Solubility AqSolDB	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-792%)	-0.081 (-10%)	0.575 (72%)	**0.740 (93%)**	**0.792 (100%)**
VDss Lombardo	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-inf (-inf%)	-0.859 (-287%)	-0.010 (6%)	0.175 (73%)	**0.224 (91%)**	**0.248 (100%)**

Table 5: Spearman correlations between the labels and the predictions of the quantized models on the regression tasks, averaged over all embedders.

dataset	Sign	1.58b 33%	1.58b 50%	1.58b 90%	2 bits	3 bits	4 bits	5 bits	original
Caco2 Wang	0.673 (89%)	0.713 (95%)	0.720 (95%)	0.679 (90%)	0.690 (92%)	0.733 (97%)	0.745 (99%)	**0.748 (99%)**	**0.750 (100%)**
HydrationFreeEnergy FreeSolv	0.905 (98%)	0.907 (99%)	0.906 (98%)	0.893 (97%)	0.910 (99%)	0.913 (99%)	**0.915 (99%)**	**0.916 (99%)**	**0.916 (100%)**
LD50 Zhu	0.572 (83%)	0.611 (89%)	0.618 (90%)	0.533 (77%)	0.596 (87%)	0.636 (92%)	0.669 (97%)	**0.679 (99%)**	**0.685 (100%)**
Lipophilicity (az)	0.658 (87%)	0.700 (92%)	0.705 (93%)	0.637 (84%)	0.669 (88%)	0.713 (94%)	0.741 (98%)	**0.749 (99%)**	**0.754 (100%)**
PPBR AZ	0.561 (98%)	0.563 (98%)	0.564 (99%)	0.555 (97%)	0.565 (99%)	0.568 (99%)	**0.569 (99%)**	**0.570 (99%)**	**0.570 (100%)**
Solubility AqSolDB	0.838 (94%)	0.853 (96%)	0.852 (96%)	0.756 (85%)	0.842 (94%)	0.864 (97%)	0.879 (99%)	**0.884 (99%)**	**0.887 (100%)**
VDss Lombardo	0.570 (99%)	0.572 (99%)	0.572 (99%)	0.560 (97%)	0.572 (99%)	0.573 (99%)	**0.575 (99%)**	**0.575 (99%)**	**0.576 (100%)**

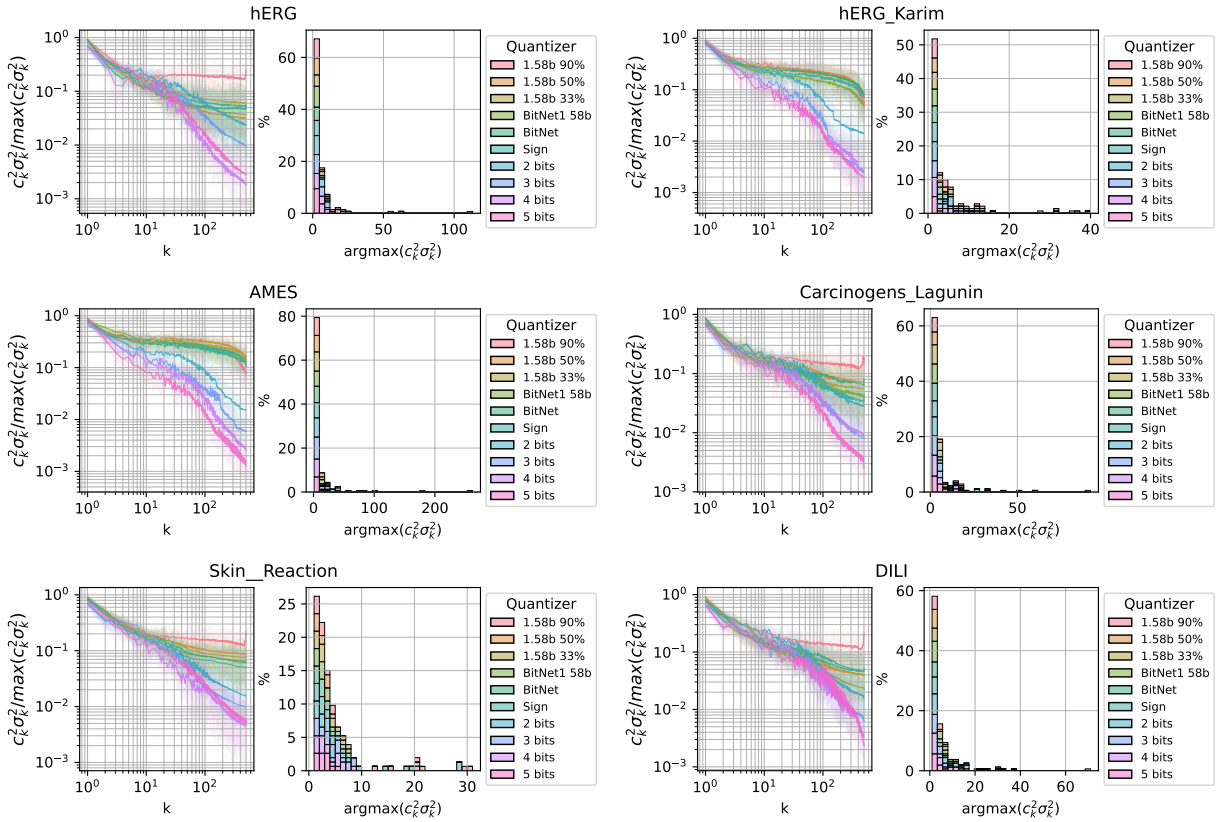


Figure 13: Evolution of $\Lambda_{k,k}$ with the k, where the indices are sorted with decreasing values of average loss, and the histogram of the index of the maximum value of $\Lambda_{k,k}$ for each quantizer.

maximum value of $\Lambda_{k,k}$ for each quantizer, on four different datasets (trained on 500 epochs, hence $k \leq 500$).

The maximum value of $\Lambda_{k,k}$ is consistently reached for low k values, which seems to confirm our hypothesis that the estimation of $r_{\mathcal{Q}_n}$ mainly revolves around the low loss values achieved by the quantizer, thereby validating our sampling strategy for estimating $r_{\mathcal{Q}_n}$.