Is My Data Safe? Predicting Instance-Level Membership Inference Success for White-box and Black-box Attacks

Tobias Leemann [1 2] **Bardh Prenkaj** [2] **Gjergji Kasneci** [2]

Abstract

We perform an extensive empirical investigation of three recent membership inference (MI) attacks on vision and language models. Our investigation includes the newly proposed Gradient Likelihood Ratio (GLiR) attack, a white-box attack with theoretical optimality guarantees. Prior research has suggested that white-box attacks cannot outperform black-box MI attacks. In this work, we challenge this hypothesis by running and evaluating this attack on real-world models with up to 53M parameters for the first time. We find that this white-box attack does indeed have the potential to outperform other attacks. We subsequently focus on the problem of MI susceptibility prediction, which is concerned with efficiently identifying individuals who are most susceptible to attack risk à priori. By doing so, we uncover which characteristics make instances susceptible to MI and whether the targeted instances are the same across attacks with different access (e.g., white-box or black-box) to the target model. We implement and study over 20 predictors of attack success. We find that GLiR mostly targets the same points as loss-based attacks and that the vulnerable instances can be efficiently predicted.

1. Introduction

With the adaptation of Machine learning (ML) in domains such as personalized AI assistance (Pataranutaporn et al., 2021), we often face sensitive data that cannot be publicly shared due to ethical or regulatory concerns. As machine learning penetrates these domains, preserving data privacy becomes essential. In particular, the trained model itself or its predictions may leak information about the training data (Shokri et al., 2017; Carlini et al., 2022a; Haim et al., 2022). For example, this is a critical problem for recent Large Language models (LLMs), where larger models have been observed to be even more prone to data leakage (Carlini et al., 2021). This work focuses on a privacy threat known as membership inference attacks (MIA), where attackers attempt to identify instances in the training data. Protecting against MIAs is crucial for enabling the secure and trustworthy deployment of machine learning, especially in domains dealing with sensitive personal data.

Prior work (Sablayrolles et al., 2019) has compared white-box (where an attacker has full access to the training pipeline and the model parameters) and black-box MIAs (allowing the attacker only to query the model) and produced evidence that nothing is to be gained through the white-box access in terms of attack success. However, a recent work (Leemann et al., 2023) proposed the Gradient-Likelihood Ratio attack (GLiR), which is theoretically optimal for single SGD-steps and showed that it outperforms loss-based attacks (Carlini et al., 2022a) drastically for small models. In this work, we are the first to put this attack to the test on real-world models and show that the attack can keep some of its advantages against the black-box attack in practice. As there is a need to guarantee *individual privacy* (Aerni et al., 2024) for every instance, we subsequently take an instance-level perspective. We contribute to better understanding individual attack susceptibility by studying the novel problem of MI *susceptibility prediction*, where the goal is to predict the success of MI attacks for identifying membership of individual data points based on their characteristics. By considering this task, we hope to reveal (1) what characteristics make points susceptible to attacks and (2) whether the diverse attacks considered exploit the same vulnerability patterns. Specifically, we make the following contributions:

- We consider three recent membership inference attacks with different access to the model: Counterfactual Distance (Pawelczyk et al., 2023), the Likelihood Ratio Attack on the loss (Carlini et al., 2022a), and the Gradient-Likelihood Ratio attack (Leemann et al., 2023), which the most powerful attack possible on single SGD steps from a theoretical standpoint.
- We are the first to apply the recent Gradient Likelihood

[*]Equal contribution [1]University of Tübingen, Tübingen Germany [2]Technical University of Munich, Munich, Germany. Correspondence to: Tobias Leemann <tobias.leemann@uni-tuebingen.de>.

ICML 2024 Workshop on the Next Generation of AI Safety, Vienna, Austria, 2024.

Ratio (GLiR) attack machine learning models of sizes up to 53M parameters, requiring practical solutions to numerical problems. We find that the white-box attack outperforms the black-box attacks in some cases.

- To determine which user's data faces the highest membership inference risk, we compute individual attack success rates for 10,000 instances across 3 datasets and implement 20 attack risk predictors. We find that the loss of models trained without the data point is most predictive, even for the non-loss-based attacks, highlighting that success for even the most complex attacks can be efficiently predicted.

In light of the above contributions, this work represents a major step forward in terms of better identifying points at risk and towards practical implementation of such checks suitable for industrial use.

2. Related Work

2.1. Membership Inference Attacks

MI attacks attempt to determine whether a given instance was part of the training dataset of a machine learning model. In the recent literature, many of these attacks have been proposed (Yeom et al., 2018; Shokri et al., 2017; Long et al., 2018; Sablayrolles et al., 2019; Haim et al., 2022; Carlini et al., 2023; Pawelczyk et al., 2023; Tan et al., 2022; 2023; Choquette-Choo et al., 2020; Leemann et al., 2023). Shokri et al. (2017) proposed a loss-based membership inference attack that determines if an instance is in the training set by testing if the loss of the model for that instance is below a specific threshold. Many recent membership inference attacks are also predominantly loss-based, where the calibration of the threshold varies from one proposed attack to the other and may be different for each instance (Carlini et al., 2022a; Ye et al., 2022; Watson et al., 2022). While there are theoretical claims that no information can be gained from white-box access over loss access (Sablayrolles et al., 2019), recent work has shown promising results with a training-gradient-based attack (Leemann et al., 2023). We investigate this claim further in this work.

2.2. Risks Quantification for Individual Instances

Theoretical Results. Azize and Basu (2024) consider the problem of private mean estimation on an instance-level perspective and find that attack susceptibility is mainly determined by the Mahalanobis distance. Considering averaged gradients in stochastic gradient descent (SGD), Leemann et al. (2023) discover a gradient susceptibility term to appear in their privacy bound that also depends on the sample gradient's Mahalanobis distance.

Empirical Observations. The overlap between points with associated privacy risks has been previously studied by Ye et al. (2022), who find that the strongest loss-based attack "R" used in their work can identify points with a high test loss. Other works have observed that outliers are most prone to attacks (Carlini et al., 2022a; Feldman and Zhang, 2020). Murakonda and Shokri (2020) present ML Privacy Meter, a tool that quantifies MI risk by running multiple MI attacks. In this work, we (i) investigate at instance level and are (ii) interested in finding good predictors for strong attacks that are more efficient than running the full attacks.

Attack	Query Access	Label Access	Minibatch Gradients
CFD (Pawelczyk et al., 2023)	✓	✗	✗
LiRA (Carlini et al., 2021)	✓	✓	✗
GLiR (Leemann et al., 2023)	✓	✓	✓

Table 1: MI attacks studied in this work and background knowledge required to run them (✓denotes required knowledge by the attacker, ✗ is not required)

3. Attacks and Risk Predictors

3.1. Attacks Considered

As the main goal of privacy is to protect personal data, the MI attack is a frequently employed method to evaluate the privacy of real-world models (Murakonda and Shokri, 2020). This attack aims to infer whether a specific instance was part of the model's training dataset and is defined as follows:

Definition 3.1 (Membership Inference Experiment (Yeom et al., 2018))**.** *Let $\mathcal{A}$ be an attacker, A be a learning algorithm, N be a positive integer, and $\mathcal{D}$ be a distribution over data points $\boldsymbol{x} \in D$, where the vector $\boldsymbol{x}$ may also be a tuple of data and labels. The MI experiment proceeds as follows: The model and data owner $\mathcal{O}$ samples $S \sim \mathcal{D}_n$ (i.e., sample n points i.i.d. from $\mathcal{D}$) and trains $A_S = A(S)$. They choose $b \in \{0, 1\}$ uniformly at random and draw $\boldsymbol{x}' \sim \mathcal{D}$ if $b = 0$, or $\boldsymbol{x}' \sim S$ if $b = 1$. Finally, the attacker is successful if $\mathcal{A}(\boldsymbol{x}', A_S, n, \mathcal{D}) = b$. $\mathcal{A}$ must output either 0 or 1.*

In this work, we consider three fundamentally different membership inference attacks requiring varying data and model access. We provide a summary in Table 1.

Counterfactual Distance Attack (CFD). This attack (Pawelczyk et al., 2023) is based on the distance to the decision boundary. No labels are needed to run this attack.

Loss Likelihood-Ratio Attack (LiRA). LiRA by Carlini et al. (2022a) is a common baseline MI attack that relies on the prediction loss and uses many shadow models that either include or exclude a point in the training set. These models are required to estimate an instance-specific loss threshold.

Gradient Likelihood Ratio Attack (GLiR). Leemann et al. (2023) study MI risk for SGD-trained models from a fundamental perspective. They inspect the information flow in SGD and devise an optimal Likelihood Ratio test for membership inference of a single SGD step. This, however, requires access to training gradients and stepwise model parameters, which can occur, for instance, in federated learning scenarios (Kairouz et al., 2021). While theoretically optimal – at least for single steps – their attack is only implemented and tested for models of up to around 2500 trainable parameters. One of our contributions is to adapt this attack to apply to large-scale models with up to 53M parameters.

3.2. Implementing GLiR for Practical Models

The original GLiR attack, as proposed in Leemann et al. (2023), requires some modifications to be applied in practice. The main bottleneck of this attack is the estimation of the inverse covariance matrix $\mathbf{\Sigma}^{-1} \in \mathbb{R}^{d \times d}$ of the parameter gradients at each step. Even storing the matrix is intractable for models with parameter counts d in the millions. Furthermore, using the common covariance estimator for gradients $\boldsymbol{\theta} \in \mathbb{R}^d, \hat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^\top$, where $\bar{\boldsymbol{\theta}}$ is the empirical mean, we require at least $n = d$ estimation samples even to have non-singular estimated covariance that can be inverted (and many more samples are required for this inversion to be numerically stable). An initial approach that we followed would be to assume independence between dimensions and only estimate a diagonal matrix. However, we observed that the independence assumption was harshly violated for the parameter gradients and obtained no promising results. As some information loss seems inevitable in the first place, we finally chose to consider only a limited number of parameter dimensions. In this work, we only use the parameters in the last layers of the models, e.g., the classification head of a language model (LM). We only estimate the covariance matrix for the gradients of these parameters. However, we still observed some dimensions that exhibited an extremely small variance, again making the numerical inversion of $\hat{\mathbf{\Sigma}}$ unstable. We excluded these dimensions with a variance below a certain threshold τ. In summary, we only chose a subset of the model parameters for the attack à priori and dynamically excluded dimensions with insufficient variance. This allowed us to run the GLiR attack on real-world models with promising results. We provide additional details in Appendix A.

3.3. Instance-Wise Attack Risk Predictors

Ye et al. (2022) defined the per-sample MI attack which adapts Definition 3.1 such that the sample $\boldsymbol{x}'$ that the MI attack is run on is fixed, i.e., if $b = 1$ we insert $\boldsymbol{x}'$ in the training dataset. By carefully considering the related literature, we identify the following variables to be potentially indicative of MI risk:

Loss. We compute the loss of the ground truth label and the model output as the first predictor variable. Many popular MI attacks are based on the prediction loss (Shokri et al., 2017; Ye et al., 2022; Watson et al., 2022).

Confidence. We define confidence as the difference (in log-odds) between the most confident class and the second most confident class. In contrast to the loss, this estimator does not require a label. Exploiting extreme confidence has been previously used as a MIA itself (Salem et al., 2018).

Input-Grad. Following Shokri et al. (2021) who use model explanations for launching membership inference attacks, we investigate input gradients of the loss that can be interpreted as a simple type of feature attribution (Simonyan et al., 2013). Shokri et al. (2021) suggest that points with a high variance in feature importance may be members of the dataset, whereas points with rather uniform (low variance) feature importances can be interpreted as non-members. We, therefore, use the variance of input gradients as a predictor.

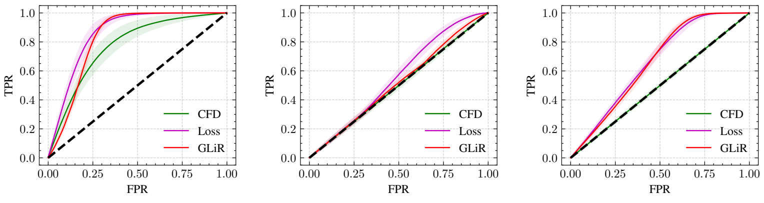
Parameter Gradients. We perform the same computation as above on the gradients w.r.t. the parameters. As the GLiR attack (Leemann et al., 2023) uses the parameter gradients, we are interested in determining whether their variance predicts attack risk.

SHAP values. A particular form of explanations that are considered by (Shokri et al., 2021) and is recurring in the literature on privacy risks of model explanations (Liu et al., 2024) are Shapley value explanations, in particular, the SHAP framework (Lundberg and Lee, 2017). We also compute the SHAP values' variance across features to predict attack risk.

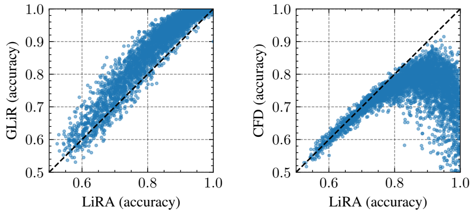
Loss Curvature. Following the intuition by Li et al. (2023b), points in the training set should result in dents in the loss landscape. Due to this effect, the model's curvature at this point should be high. The curvature is defined as the trace of the Hessian of the loss w.r.t. to the inputs. We use the stochastic Hutchinson Trace Estimator as described in Peebles et al. (2020) to estimate the curvature efficiently.

Mahalanobis Distance. We compute the Mahalanobis distance of the inputs w.r.t. the data distribution. Several theoretical works have identified this quantity (cf. Sec. 2.2), so we consider it a good candidate for risk prediction. As this distance is not defined for discrete inputs, we use the Mahalanobis distance of the latent representations on IMDB.

Outlier Detection (VAE). We use a Variational Autoencoder's (VAE) reconstruction loss as an outlier detector (Eduardo et al., 2020; Lai et al., 2023). Prior work has identified outliers to be most prone to MI attacks (Section 2.2).



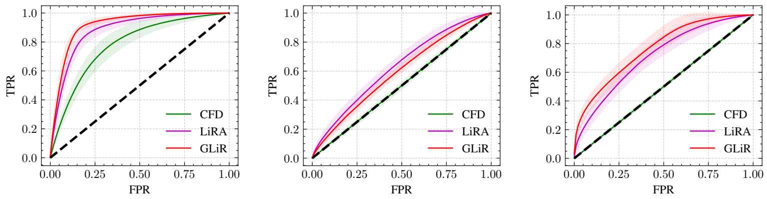
(a) CIFAR-10 (b) Skin Cancer (c) IMDB



(a) GLiR vs. LiRA (b) CFD vs. LiRA

Figure 1: **Success-rates for non-instance-calibrated attacks.** We show the success rates of the three attacks studied in this paper, when the scores are not recalibrated instance-wise using shadow models. Loss performs best on Cancer but on a similar level as GLiR on the other two datasets.

Figure 2: **Relating instance-wise success for different attacks.** We observe that the loss-based LiRA and GLiR attack similar points, with GLiR usually being more powerful.



(a) CIFAR-10 (b) Skin Cancer (c) IMDB

Figure 3: **Success-rates for instance-calibrated attacks.** We recalibrate the attack scores per instance to make the scores comparable over instances, using the same 200 shadow models for each attack. This results in the LiRA attack for the loss. We subsequently compute the trade-off curves. GLiR outperforms LiRA on CIFAR and IMDB.

4. Experiments

Setup and Datasets. We use the CIFAR-10 (Krizhevsky et al., 2009), which has been used in previous privacy studies (Carlini et al., 2022a). However, we are interested in evaluating the attacks on a real-world dataset of sensitive nature, which is why we also select a dataset of Skin Cancer images (Tschandl et al., 2018). Finally, to connect to recent language model applications, we use the IMDB dataset of movie reviews (Maas et al., 2011). We use ResNet-18 (He et al., 2016) models with 11M parameters for images and a DistillBERT model with four layers (Sanh et al., 2019) and a total of around 53M parameters for the IMDB dataset. We provide source code for our experiments online.[1]

Prediction Dataset. We split each dataset into an *attack assessment* dataset and a *background* dataset, which will be used by the attacker as an estimate of the background distribution $\mathcal{D}$. We train $N{=}200$ models for each task using the same training routine on randomly sampled 50%-splits of the attack assessment dataset. This results in around 100 models where each instance is a member and the same number of models are trained without each instance. We can compute the attack scores for each model and instance. Having access to a large number of models with or without each instance and the corresponding attack scores, we can compute the empirical trade-off curve between False Positives and True Positives *on an instance level*. We aggregate this curve to a single score. In this work, we use the maximum attack accuracy that is possible for a particular instance. Note that while accuracy should not be used for evaluating MI success on a dataset level (Carlini et al., 2022a), the accuracy is meaningful on an instance level. An overview of the individual score distribution is given in Figure 4 (Appx.). The goal will be to predict this individual accuracy score (dependent variable) from a data point's own characteristics or the model behavior at the point (independent variable).

Independent Models. Some predictors, e.g., loss, require additional models to be trained. To this end, we further train 50 more models not used in the attacks, where each instance from the assessment dataset has a 50% of being included. We average the model characteristics either for all models ("all"), the models with an instance ("in"), or the models trained without the instance ("out") as predictors, e.g., *loss ("out")* denotes the average loss of the independent models trained without a certain instance.

4.1. Attack Efficacy

We plot trade-off curves obtained for the standard attack scores in Figure 1 where the same threshold is used across all instances. We observe that the loss-based attack is the most powerful on CIFAR and Skin Cancer without recalibration.

[1] `https://github.com/tleemann/gaussian_mip`

(a) CIFAR-10

Predictor	CFD	LiRA	GLiR
loss (in)	0.12	0.29	0.28
loss (out)	**0.48**	**0.91**	**0.89**
loss (all)	0.47	0.89	0.89
confidence (in)	0.43	0.41	0.36
confidence (out)	0.36	0.52	0.60
confidence (all)	0.23	0.61	0.64
param-grad (in)	0.08	0.23	0.22
param-grad (out)	0.38	0.71	0.79
param-grad (all)	0.37	0.70	0.77
input-grad (in)	0.04	0.15	0.14
input-grad (out)	0.37	0.52	0.61
input-grad (all)	0.36	0.52	0.60
curvature (in)	0.18	0.21	0.21
curvature (out)	0.28	0.53	0.60
curvature (all)	0.22	0.52	0.60
shap (in)	0.04	0.07	0.07
shap (out)	0.21	0.17	0.20
shap (all)	0.14	0.14	0.15
mahalanobis	0.06	0.04	0.05
vae-loss	0.06	0.04	0.05

(b) IMDB

Predictor	CFD	LiRA	GLiR
loss (in)	0.03	0.55	0.56
loss (out)	**0.04**	**0.90**	**0.89**
loss (all)	0.03	0.90	0.89
confidence (in)	0.03	0.67	0.66
confidence (out)	0.03	0.73	0.79
confidence (all)	**0.04**	0.78	0.82
param-grad (in)	0.03	0.36	0.34
param-grad (out)	0.03	0.65	0.66
param-grad (all)	0.03	0.64	0.64
input-grad (in)	0.01	0.14	0.13
input-grad (out)	0.02	0.89	0.89
input-grad (all)	0.03	0.88	0.88
curvature (in)	0.02	0.40	0.39
curvature (out)	0.03	0.88	0.87
curvature (all)	0.03	0.87	0.87
shap (in)	0.04	0.57	0.56
shap (out)	0.04	0.74	0.80
shap (all)	0.04	0.92	0.92
mahalanobis-latent (in)	0.03	0.55	0.54
mahalanobis-latent (out)	0.03	0.59	0.64
mahalanobis-latent (all)	**0.04**	0.63	0.65

(c) Skin Cancer

Predictor	CFD	LiRA	GLiR
loss (in)	0.06	0.72	0.42
loss (out)	**0.07**	**0.82**	**0.45**
loss (all)	0.06	0.79	0.44
confidence (in)	0.06	0.60	0.39
confidence (out)	**0.07**	0.52	0.37
confidence (all)	0.05	0.57	0.39
param-grad (in)	0.06	0.69	0.41
param-grad (out)	0.05	0.78	0.48
param-grad (all)	0.05	0.76	0.46
input-grad (in)	0.00	0.04	0.01
input-grad (out)	0.01	0.11	0.06
input-grad (all)	0.00	0.07	0.03
curvature (in)	0.06	0.63	0.40
curvature (out)	0.05	0.69	0.46
curvature (all)	0.06	0.68	0.45
shap (in)	0.06	0.18	0.20
shap (out)	**0.07**	0.18	0.21
shap (all)	0.06	0.19	0.21
mahalanobis	0.05	0.19	0.14

Table 2: Evaluation of the attack predictors to predict MI success (R^2-Score of RF-Regressor).

On IMDB, GLiR performs on par. The CFD attack only performs substantially better than random on the CIFAR-10 dataset. We then recalibrate the attacks per instance. To this end, we use the 200 models trained and computing quantiles for the attack scores per instance. We then run the attack using the quantiles as a score. We show the corresponding curves in Figure 3. However, when recalibrating scores instance-wise with shadow models (resulting in the LiRA attack for the loss), we observe that GLiR outperforms LiRA on CIFAR-10 and IMDB across the entire trade-off curve. LiRA maintains its advantage for Skin-Cancer, although success rates for both attacks have substantially improved. We provide log-log plots in Appendix B.

4.2. Predicting Individual Attack Susceptibility

Evaluating Risk Predictors. We fit a simple non-linear random forest-regressor, which we constrain to have a max-depth of 5 to prevent overfitting to the prediction dataset to predict attack accuracy. We then evaluate this regressor's R^2 score. The R^2 score corresponds to the share of variance that can be explained by the predictor and report the results in Table 2 (cf. Table 6 for rank-correlation). We find that the loss of models trained without the instance is most predictive of MI success. This is not unexpected for the loss-based attack because all points have a relatively small loss when they are in the training dataset. MI risk is therefore determined by their behavior when in the test set, if the points have a low loss, they will be relatively safe to attacks. Surprisingly, the loss is also most predictive for the other attacks, suggesting the most vulnerable points stay rather similar across attacks. We show that a small number of "out" models can reliably predict MI success in Figure 7 (Appx.), even for the most complex GLiR attack, making it possible to warn users of risks without implementing and running the complex attack.

4.3. Relating Vulnerabilities Across Attacks

Inspired by Ye et al. (2022), we are interested in relating the success of the attacks across different instances. This is important to answer the question of whether the same points are at risk for all the attacks or whether the attacks actually use different characteristics of the instances. We show scatter plots in Figure 2. We find that GLiR and LiRA target mostly the same instances, while CFD is most successful for the instances that were mildly certain for LiRA (i.e., medium test loss). CFD's success decreases again for high test loss points, while LiRA's still increases.

5. Discussion and Conclusion

In this work, we obtained several insights that have interesting implications for future research.

Scaling Training Gradient Attacks. First, we found that even when using only a small subset of the model parameters (e.g., <0.0001% for DistilBERT), the GLiR attack already outperforms the LiRA attack. This suggests that the attack is even more powerful with more computational resources and a larger background dataset. Recalibrating attack scores instance-wise is however required to obtain this result. The most brittle part of the attack remains the computation of the inverse covariance and gradient product. We leave an investigation of better estimators to compute this quantity to future work. Nevertheless, our finding does not confirm prior work's hypothesis (Sablayrolles et al., 2019) on the equivalence of black-box and white-box attacks, possibly due to overly stringent assumptions on the parameters' distribution. Therefore, we argue that additional research on more resource efficient white-box MIAs is required.

Predicting Attack Risk In Practice. Overall, our results highlight that points with high test loss are still most vulnerable to MI attacks. Therefore, a practical strategy to identify points at risk could be training a model on public data from the same domain and subsequently testing this model on the private data points. Model developers should especially consider the privacy risks associated with the points that incur high losses on the public data-trained model.

Protecting Vulnerable Instances. While complete removal of the instances at risk often results in the vulnerability merely shifting to other points (Carlini et al., 2022b), the identification of vulnerable instances can be crucial to better understanding risk factors and to developing adaptive defenses while keeping the overall utility of the models high. One proposed approach is to artificially increase the training loss for these points (Li et al., 2023a). Our observation that the three MI attacks all target points with high test loss highlights that defenses that focus on these points may also be effective protection against non-loss-based attacks.

In conclusion, we hope this work provides further insights into the structure of instances vulnerable to MI attacks that can be useful in developing better defenses against this prominent privacy threat.

Acknowledgments

The authors would like to thank Martin Pawelczyk for his valuable feedback on this work.

References

M. Aerni, J. Zhang, and F. Tramèr. Evaluations of machine learning privacy defenses are misleading. *arXiv preprint arXiv:2404.17399*, 2024.

A. Azize and D. Basu. How much does each datapoint leak your privacy? quantifying the per-datum membership leakage. *arXiv preprint arXiv:2402.10065*, 2024.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. *IEEE Symposium on Security and Privacy (SP)*, 2022a.

N. Carlini, M. Jagielski, C. Zhang, N. Papernot, A. Terzis, and F. Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022b.

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

C. A. Choquette-Choo, F. Tramèr, N. Carlini, and N. Papernot. Label-only membership inference attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume abs/2007.14321, 2020.

S. Eduardo, A. Nazábal, C. K. Williams, and C. Sutton. Robust variational autoencoders for outlier detection and repair of mixed-type data. In *International Conference on Artificial Intelligence and Statistics*, pages 4056–4066. PMLR, 2020.

V. Feldman and C. Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

N. Haim, G. Vardi, G. Yehudai, O. Shamir, et al. Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

C. Lai, D. Zou, and G. Lerman. Robust variational autoencoding with wasserstein penalty for novelty detection. In F. J. R. Ruiz, J. G. Dy, and J. van de Meent, editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 3538–3567. PMLR, 2023. URL `https://proceedings.mlr.press/v206/lai23a.html`.

T. Leemann, M. Pawelczyk, and G. Kasneci. Gaussian membership inference privacy. *Advances in Neural Information Processing Systems*, 36, 2023.

J. Li, N. Li, and B. Ribeiro. Mist: Defending against membership inference attacks through membership-invariant subspace training. *arXiv preprint arXiv:2311.00919*, 2023a.

M. Li, J. Wang, J. Wang, and S. Neel. Mope: Model perturbation-based privacy attacks on language models. *arXiv preprint arXiv:2310.14369*, 2023b.

H. Liu, Y. Wu, Z. Yu, and N. Zhang. Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 120–120. IEEE Computer Society, 2024.

Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

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

P. Pataranutaporn, V. Danry, J. Leong, P. Punpongsanon, D. Novy, P. Maes, and M. Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

M. Pawelczyk, H. Lakkaraju, and S. Neel. On the privacy risks of algorithmic recourse. In *International Conference on Artificial Intelligence and Statistics*, pages 9680–9696. PMLR, 2023.

W. Peebles, J. Peebles, J.-Y. Zhu, A. Efros, and A. Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 581–597. Springer, 2020.

A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

R. Shokri, M. Strobel, and Y. Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241, 2021.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

J. Tan, B. Mason, H. Javadi, and R. Baraniuk. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17488–17500, 2022.

J. Tan, D. LeJeune, B. Mason, H. Javadi, and R. G. Baraniuk. A blessing of dimensionality in membership inference through regularization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 10968–10993. PMLR, 2023.

P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

L. Watson, C. Guo, G. Cormode, and A. Sablayrolles. On the importance of difficulty calibration in membership inference attacks. *International Conference on Learning Representations*, 2022.

J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.

S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

A. Experimental Details

We integrate source code to run the attacks in this project in the Github repository of the GLiR attack[2]. We provide some more details on the implementation below:

Datasets. Recall that we rely on CIFAR-10, Skin Cancer, and IMDB datasets in this paper. Table 3 summarizes their main characteristics. Note that the final images of Skin Cancer were cropped manually to fit the 800 × 600 pixels. Furthermore, since IMDB is a language dataset, its data shape cannot be determined à priori.

Table 3: Dataset characteristics.

	Datasets		
	CIFAR-10	Skin Cancer	IMDB
Type	Image	Image	Text
Num. instances	60,000	10,015	50,000
Assess Dataset Size	4,000	2,000	4,000
Num. classes	10	7	2
Data shape	$32 \times 32 \times 3$	$800 \times 600 \times 3$	-
Model	Resnet-18	Resnet-18	DistilBert (4 layers)
Training Batch Size	32	64	64
Training Epochs	30	40	8
Training LR (Adam)	1e-3	1e-3	5e-5
Main applications	Image classification	Medical image analysis	Sentiment analysis

Hyperparameters and Details for GLiR Attack. Table 4 shows each dataset's hyperparameters used in the GLiR attack. We do not use every updates batch gradient but instead store a fixed number of training minibatch gradients alongside with the model parameters for each training run. We later use these samples to perform the attack. The background samples are used to estimate the gradients' distribution, while the number of parameters is used to estimate the CDF scores: i.e., compute the p-values under the null hypotheses "x' is a test point". We chose to use the parameters in the last layers of the networks for our attack as we observed their gradients to have the highest variance and therefore provided a clearer signal. Despite only using a small fraction of the parameters and the minibatch gradients due to computational constraints, the attack performance was quite high, especially after recalibration. We leave it to future work to assess the maximum performance of this attack, when all available information is processed.

Table 4: GLiR hyperparameters for each dataset.

	Datasets		
	CIFAR-10	Skin Cancer	IMDB
Training batch size	32	32	64
Num. parameters used	5120	3584	2306
Share of parameters used	0.00046%	0.00032%	0.00004%
Bg. samples	45000	6500	5000
Num of Batch Grads.	30	40	40
Var. limit τ	10^{-1}	10^{-2}	10^{-3}

Additional Details on Risk Predictors.

Shapley Values. We use the `shap` libarary to compute the Shapley values. For the Resnet models, we use the `shap.DeepExplainer`. For the language model, we use the `shap.KernelExplainer`. As for input grad and suggested by (Carlini et al., 2022b), we take the variance of the SHAP values as risk predictor.

VAE-Loss. For CIFAR-10, we also use an autoencoder available online [3]. However, as the results were not promising and we did not find available VAE implementations for the remaining two datasets, we decided to skip this predictor on IMDB and Skin Cancer.

Mahalanobis Distance. The Mahalanobis distance for an instance $\boldsymbol{x}$ w.r.t. a distribution D with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is defined as

$$d(\boldsymbol{x}, D) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \qquad (1)$$

As we need to estimate the covariance to compute it, this predictor can only be used if dimensionality is small enough. To this end, for the LLM which has large embedding matrices as inputs, we compute the Mahalanobis distance on the latent embeddings (it is therefore dependent on trained models). For the Cancer dataset, we use a random subportion of 2500 input dimensions to estimate Mahalanobis distance.

B. Additional Results

Individual success rate histograms. We show histograms of the individual success rates in Figure 4 for CIFAR-10. This dataset's median and mean accuracy are the highest for the GLiR attack.

Log-Log-Plots. We show the logarithmic plots of the trade-off curves in Figure 5 and Figure 6 with similar results as in the main plots.

Spearman Rank-Correlation. The Spearman rank correlation for the risk predictors is given in Table 6. From this metric we can also see the signs of the relation between the predictors and the attack risk. Loss (out) still maintains the highest correlation. As the CFD attack does not substantially outperform random guesses there is only weak correlation.

How many models are required for reliable risk prediction? We consider the number of models we require to predict risk reliably. Using the most powerful predictor from our previous results ("loss, out"), we use more models trained without specific points and average their loss successively. We show the results in Figure 7. We find that even a single model can already be quite predictive, whereas stable results can be obtained using ten or more models.

[2] https://github.com/tleemann/gaussian_mip

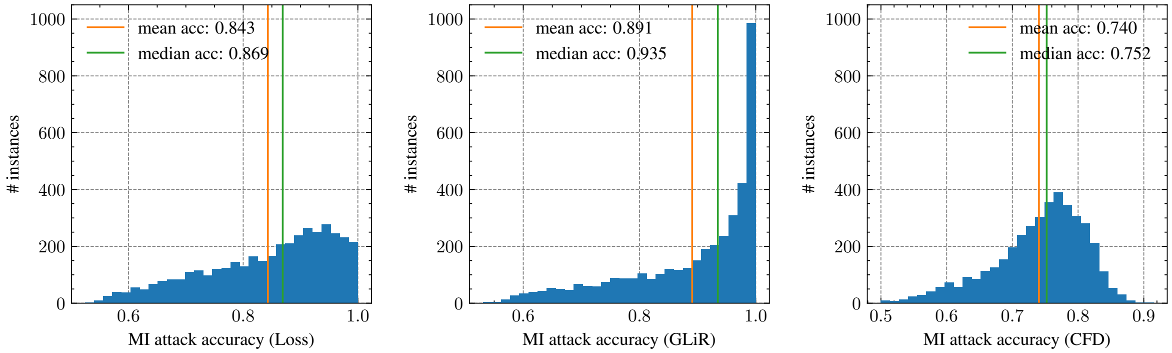
[3] https://github.com/o-tawab/Variational-Autoencoder-pytorch

Table 5: Coefficients of predictors in a linear regression model. We combine the predictors to a linear regression model for risk prediction on CIFAR-10. As the out scores where usually most predictive and as "in" "out" and "all" are linearly related, we use only the "out" versions of the predictors. We obtain the following coefficients (predictor scores have been normalized):

Dataset	VAE	mahal.	Loss	Confid.	shap	curva.	inp-grad	grad	tot. R^2
LiRA	0.000	0.003	**0.075**	-0.037	-0.007	0.002	0.013	-0.001	0.88
GLiR	0.002	0.003	**0.059**	-0.045	-0.008	0.001	0.022	0.002	0.84
CFD	0.000	0.007	-0.022	**-0.046**	0.004	-0.006	0.009	0.016	0.54

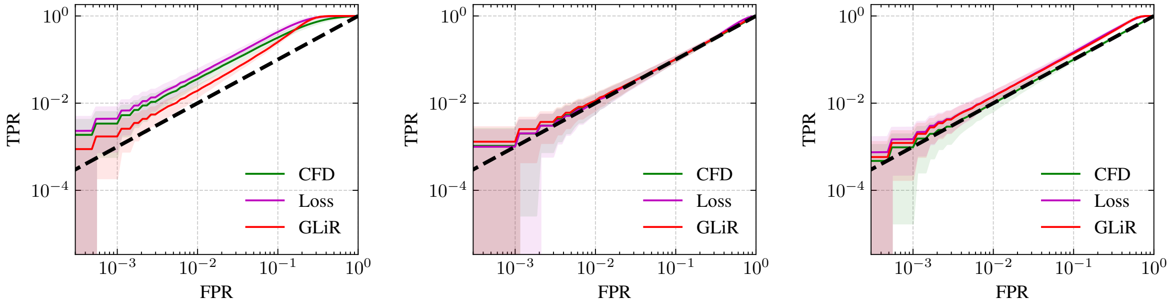
This highlights that even for computationally challenging attacks like GLiR, the points with an attack surface can be identified at low costs.

Combining predictors. We investigate combining the predictors to a more powerful predictor with linear regression in Table 5 and report the coefficient. However, we observed that at least the linear combined model cannot outperform the single scores on their own. Loss has the highest predictive coefficients for Loss and GLiR, whereas the label-independent confidence has the highest coefficient for the, also label-independent, CFD attack.



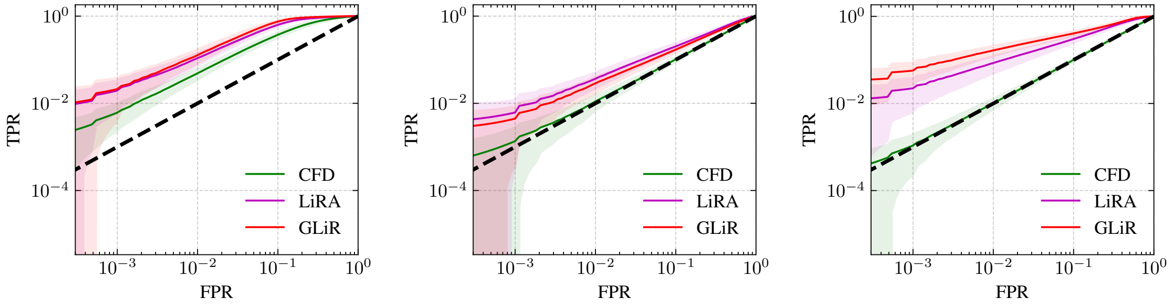
(a) LiRA (b) GLiR (c) CFD

Figure 4: **Success-rates for individual instances.** We show the distribution of the individual success rates for the three attacks on the CIFAR-10 dataset. On this dataset, GLiR can identify many instances with high certainty.



(a) CIFAR-10 (b) Skin Cancer (c) IMDB

Figure 5: **Success-rates for non-calibrated attacks.** We show the success rates of the three attacks studied in this paper when the scores are not recalibrated instance-wise using shadow models. Log-Log plot corresponding to Figure 1.



(a) CIFAR-10 (b) Skin Cancer (c) IMDB

Figure 6: **Success-rates for recalibrated attacks.** We recalibrate the attack scores per instance to make the scores comparable over instances and subsequently compute the trade-off curves. We use empirical quantiles for each instance as surrogate scores. Log-Log plot corresponding to Figure 3.

(a) CIFAR-10

Predictor	CFD	LiRA	GLiR
loss (in)	0.11	0.44	0.42
loss (out)	0.28	0.95	0.95
loss (all)	0.28	0.94	0.93
confidence (in)	0.01	-0.60	-0.57
confidence (out)	0.34	-0.55	-0.57
confidence (all)	0.22	-0.65	-0.65
input-grad (in)	0.15	0.39	0.37
input-grad (out)	0.50	0.51	0.54
input-grad (all)	0.48	0.52	0.54
curvature (in)	0.14	0.38	0.36
curvature (out)	0.44	0.49	0.51
curvature (all)	0.43	0.49	0.51
grad (in)	0.13	0.39	0.37
grad (out)	0.53	0.70	0.72
grad (all)	0.51	0.69	0.71
shap (in)	0.13	0.20	0.21
shap (out)	0.40	0.34	0.37
shap (all)	0.32	0.30	0.32
vae-reconstruction-loss	0.12	0.04	0.05
mahalanobis	0.11	0.03	0.04

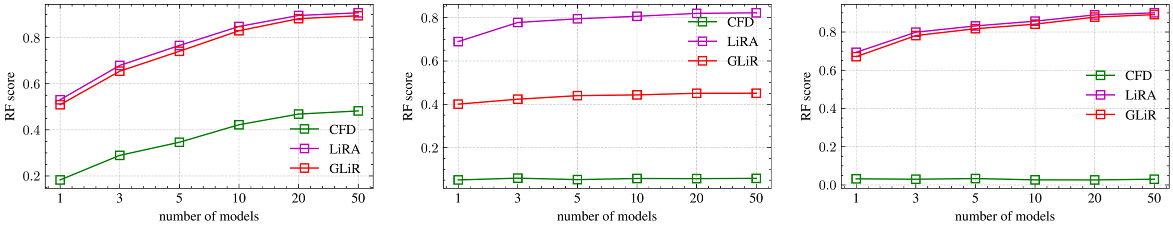
(b) IMDB

Predictor	CFD	LiRA	GLiR
loss (in)	0.01	0.76	0.75
loss (out)	0.00	0.90	0.90
loss (all)	0.00	0.90	0.90
confidence (in)	-0.00	-0.81	-0.81
confidence (out)	-0.01	-0.87	-0.88
confidence (all)	-0.01	-0.89	-0.90
input-grad (in)	-0.00	0.77	0.77
input-grad (out)	0.00	0.90	0.91
input-grad (all)	-0.00	0.90	0.91
curvature (in)	0.01	0.71	0.70
curvature (out)	-0.00	0.89	0.89
curvature (all)	-0.00	0.89	0.89
grad (in)	-0.01	0.64	0.62
grad (out)	-0.01	0.80	0.80
grad (all)	-0.01	0.79	0.78
shap (in)	-0.00	0.01	0.01
shap (out)	-0.00	0.02	0.03
shap (all)	-0.01	0.03	0.03
mahalanobis-latent (in)	0.01	0.73	0.73
mahalanobis-latent (out)	0.01	0.80	0.80
mahalanobis-latent (all)	0.01	0.80	0.81

(c) Skin Cancer

Predictor	CFD	LiRA	GLiR
loss (in)	0.03	0.85	0.59
loss (out)	0.02	0.88	0.64
loss (all)	0.02	0.88	0.63
confidence (in)	-0.02	-0.79	-0.56
confidence (out)	-0.01	-0.74	-0.57
confidence (all)	-0.01	-0.77	-0.57
input-grad (in)	0.03	0.82	0.62
input-grad (out)	0.02	0.85	0.66
input-grad (all)	0.02	0.85	0.66
curvature (in)	0.03	0.80	0.60
curvature (out)	0.01	0.83	0.65
curvature (all)	0.01	0.83	0.64
grad (in)	0.03	0.83	0.60
grad (out)	0.02	0.87	0.66
grad (all)	0.02	0.86	0.65
shap (in)	0.00	0.37	0.38
shap (out)	0.02	0.38	0.39
shap (all)	0.01	0.39	0.40
mahalanobis	0.03	0.36	0.27

Table 6: Prediction scores as Spearman-rank correlation. The results confirm our prior findings.



(a) CIFAR-10

(b) Skin Cancer

(c) IMDB

Figure 7: **Ablation study for the number of models used compute "out"-loss.** RF-Score corresponds to the R^2 score used in the main table. The results show that with 5 or 10 models, good MI risk predictions are possible.