Estimating Model-Level Membership Inference Vulnerability Without Reference Models

Anonymous Authors[1]

Abstract

Membership inference attacks (MIAs) are the standard tool for evaluating the privacy risks of AI models, but state-of-the-art attacks require training tens to hundreds of expensive reference models. We present a framework for estimating model-level vulnerability to the Likelihood Ratio Attack (LiRA) directly from the train and test loss distributions of the target model, with no reference models required. We show that LiRA's per-sample signal decomposes into a variance-ratio term and a residual mean-shift term, placing models on a continuum of uncertainty collapse that is directly observable from loss distribution shape. At the heavy-tailed end (image classifiers), the LOSS attack TNR predicts LiRA TPR@FPR$= 10^{-3}$ with RMSE 0.03 across 9 architectures and 4 datasets, outperforming low-cost reference-model attacks such as RMIA. At the symmetric end (LLMs), the LOSS attack AUC predicts LiRA TPR with RMSE 0.01 across five GPT-2 sizes from 10M to 1B parameters.

1. Introduction

Large-scale machine learning models are increasingly fine-tuned on sensitive data, yet research has shown they may inadvertently memorize training samples (Carlini et al., 2022b; 2019). Membership inference attacks (MIAs) have become the primary tool to quantify this risk, measuring the True Positive Rate (TPR) at a low False Positive Rate (FPR) (Carlini et al., 2022a; Ye et al., 2022; Zarifzadeh et al., 2023)—the members an attacker can *confidently* identify. This metric is also aligned with legal standards such as the EU GDPR "reasonably likely" standard for singling out (European Data Protection Board, 2024; , ICO).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

The strongest available attack, LiRA (Carlini et al., 2022a), requires training tens to hundreds of reference models at full computational cost, with attack power continuing to grow with model count (Hayes et al., 2025). This makes SOTA privacy auditing prohibitive in practice, particularly in iterative workflows such as hyperparameter search or architecture selection, where many candidate models must be assessed (Ponomareva et al., 2023). Recent work has reduced this cost (RMIA (Zarifzadeh et al., 2023) requires as few as 2 reference models) or identified vulnerable *samples* for free (Pollock et al., 2025; Leemann et al., 2024), but no existing method estimates *model-level* risk without any reference models.

We propose a framework to estimate model-level vulnerability to LiRA directly from the shape of the target model's train and test loss distributions, with no reference model training required. We show that LiRA's per-sample signal decomposes into a variance-ratio term and a residual mean-shift term (Section 2), placing models on a continuum of uncertainty collapse that is directly observable from loss distribution shape. At the heavy-tailed end, the LOSS attack TNR predicts LiRA TPR; at the symmetric end, the LOSS attack AUC does. We confirm both empirically across 9 image classification architectures, 4 datasets, and five GPT-2 models from 10M to 1B parameters.

We believe this will substantially lower the cost of privacy risk assessment in practice, in particular for iterative workflows and foundation models where training reference models at scale has previously made routine auditing impractical.

2. Framework

Why the tail separates members from non-members. In both member and non-member loss distributions, most probability mass lies at low loss (the "head"), with a smaller fraction at high loss (the "tail"). For members, the head corresponds to learned or memorized examples; for non-members, the tail contains difficult or outlier cases the model fails to fit (Pollock et al., 2025). The two distributions overlap heavily at the head, and what separates them is the tail, dominated by non-members. Training has shrunk the member tail by pushing high-loss examples to low loss.

Figure 1. **Loss-distribution shape determines the best proxy for MIA risk.** As the fraction of hard-to-fit non-members increases, the non-member loss distribution becomes increasingly heavy-tailed **(B)**, and Pearson correlation with LiRA TPR transitions from LOSS AUC to LOSS TNR **(C)**.

Vulnerable members are those missing from the tail. Assuming i.i.d. sampling, the absence of a heavy tail for members suggests samples that would have been high-loss (if unseen) have shifted to the low-loss region after training. These *tail-to-head* samples are exactly those that LiRA targets: they exhibit low ℓ_{target} yet high $\bar{\ell}_{\text{out}}$, aligning with the heavy tail of the non-member distribution. Tail-to-head migration requires training to pin down model behavior at the trained sample. When the training set carries heavy near-duplication (as in large text corpora (Shilov et al., 2024)) including a specific (x, y) does not meaningfully change model behavior at x, and migration does not occur.

Theoretical analysis via Gaussian Process regression. GP regression provides a closed-form solution for LiRA as the number of shadow models grows. The log likelihood ratio decomposes as:

$$\Lambda = \underbrace{\frac{1}{2}\ln\left(\frac{\sigma^2_{\text{non}}}{\sigma^2_{\text{mem}}}\right)}_{\text{variance-ratio term}} + \underbrace{\frac{(y-\mu_{\text{non}})^2}{2\sigma^2_{\text{non}}} - \frac{(y-\mu_{\text{mem}})^2}{2\sigma^2_{\text{mem}}}}_{\text{residual term}}. \quad (1)$$

LiRA's advantage over the LOSS attack lives in the variance-ratio term, whose size depends on how much σ^2 collapses on the training sample. This gives two limits.

Strong-collapse ($\sigma^2_{\text{mem}} \ll \sigma^2_{\text{non}}$): the variance-ratio term dominates, samples with high σ^2_{non} receive a large LiRA bonus, and the non-member loss distribution develops a heavy tail. LOSS TNR at low FNR, the fraction of non-members correctly classified as such, is the natural summary.

Weak-collapse ($\sigma^2_{\text{mem}} \approx \sigma^2_{\text{non}}$): the variance-ratio term vanishes, Λ reduces to a mean-shift signal, and the loss distributions are near-symmetric. LOSS AUC, a measure of population-level mean separation, is the natural summary.

Real models lie on a continuum between these limits. As $\sigma^2_{\text{mem}}/\sigma^2_{\text{non}}$ shrinks on average, the non-member distribution grows heavier-tailed and TNR becomes a stronger proxy; as the ratio approaches 1, the distributions become symmetric and AUC overtakes TNR. Crucially, the loss-distribution shape itself, requiring no reference models, tells the practitioner which proxy applies. We confirm this continuum in a controlled GP toy (Figure 1): varying the fraction of hard-to-fit non-members from 0% to 100% slides the model from the symmetric to the heavy-tailed end, with Pearson correlation to LiRA TPR transitioning from AUC-dominated to TNR-dominated accordingly.

3. Reference-Free Proxies

We instantiate the framework with two proxies, computed solely from the target model's losses on its train and test sets.

LOSS TNR (for heavy-tailed distributions). The fraction of non-members correctly identified by the LOSS attack at threshold τ matched to achieve FNR = FPR:

$$\text{TNR}_{\text{LOSS}}(D_{\text{test}}, f_\theta, \tau) = \frac{|\{(x,y) \in D_{\text{test}} : -\ell(x,y) > \tau\}|}{|D_{\text{test}}|} \quad (2)$$

This directly estimates π_{non}, the non-member tail mass, which drives LiRA TPR in the strong-collapse regime.

LOSS AUC (for symmetric distributions). The area under the ROC curve for the LOSS attack, equal to $\Pr(\ell_{\text{mem}} < \ell_{\text{non}})$. This captures the standardized mean separation d that drives LiRA TPR in the weak-collapse regime.

Proxy selection. Several diagnostics can guide proxy selection (loss distribution shape, variance ratio $\sigma^2_{\text{non}}/\sigma^2_{\text{mem}}$, empirical TNR at low FNR); we find a KS goodness-of-fit test works well in practice. We fit a constrained Gaussian $N(\mu+\Delta, \sigma_{\text{mem}})$ to $\mathcal{L}_{\text{non}}$ and compute the KS statistic against the empirical CDF. A small KS indicates LOSS AUC is appropriate; a large KS (heavy tails or skew) indicates LOSS TNR. KS cleanly separates model families: GPT-2 runs yield 0.049–0.069, image classifiers 0.120–0.888.

4. Experiments

The GP analysis in Section 2 predicts that TNR tracks LiRA TPR at the heavy-tailed end and AUC at the symmetric end. We now confirm both empirically.

4.1. Image Recognition Models

We train 9 architectures (ResNet-20, WRN28-2, MobileNetV2, DenseNet121, WRN40-4, ResNet-18, WRN28-10, VGG11, VGG16; 60K–172M parameters) on 4 datasets (MNIST, CIFAR-10, CINIC-10, CIFAR-100), following Carlini et al. (2022a) with 64 reference models (32 IN, 32 OUT). Image classifiers lie in the strong-collapse regime, exhibiting heavy-tailed non-member loss distributions. For D_{test}, model f_θ, and threshold τ which we select to achieve a False Negative Rate equal to the FPR of the LiRA attack for which we are estimating the TPR:

$$\text{TNR}_{\text{LOSS}}(D_{\text{test}}, f_\theta, \tau) = \frac{|\{A_{\text{LOSS}}(f_\theta, x, y) > \tau \mid (x,y) \in D_{\text{test}}\}|}{|D_{\text{test}}|} \quad (3)$$

Figure 2 shows LOSS TNR to be a strong predictor of LiRA TPR@FPR$=10^{-3}$ (R^2=0.954, RMSE$=0.033$) across all 36 model-dataset combinations, with narrow bootstrapped confidence intervals. Table 1 shows it outperforms all baselines including RMIA with 2 reference models (RMSE$=0.046$), despite requiring no reference models at all. LOSS AUC is less effective here because memorization is concentrated in the tail; the train-test gap misses tail asymmetry entirely.

4.2. LLMs

At the symmetric end of the continuum, only the residual mean-shift term of Eq. 1 survives. LLMs sit here for two compounding reasons. First, mosaic memory (Shilov et al., 2024) means information at (x, y) is typically already represented through near-duplicates elsewhere in the training corpus, so including a specific sample does not meaningfully reduce per-sample uncertainty. Second, the sequence-level loss averages over many tokens that carry no membership signal, diluting whatever signal the relevant tokens do carry (Tao & Shokri, 2025). Together these effects mean $\sigma^2_{\text{mem}} \approx \sigma^2_{\text{non}}$ across samples, and the variance-ratio term of Eq. 1 contributes little to LiRA's score.

This is also precisely the setting where reference-model auditing is most computationally prohibitive. Hayes et al. (2025) show that LiRA TPR continues to improve with up to 256 reference GPT-2 models—each requiring a full pretraining run on the same data and architecture. Our framework sidesteps this entirely.

Following the setup proposed by Hayes et al. (2025), we evaluate five GPT-2 models (Radford et al., 2019) (10M–1018M parameters) trained on a 2^{19}-sample subset of C4 (Raffel et al., 2020), with LiRA using 256 reference models. Per-sample variance plots confirm that σ^2_{in} and σ^2_{out} concentrate close to $y{=}x$ across all model sizes, with visibly less collapse than in any image setup (see Appendix). Loss distributions are near-symmetric

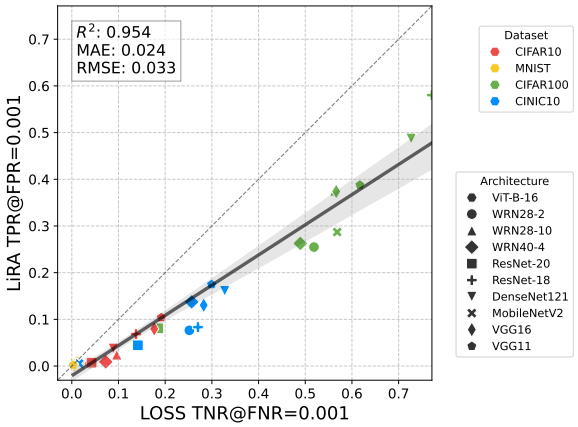

Figure 2. LOSS TNR reliably predicts LiRA TPR@FPR$=0.001$ across varied architectures and datasets (RMSE$=0.033$, 97.5% CI shaded).

Table 1. Predictors of LiRA TPR@0.001, averaged across 9 architectures and 4 datasets.

Metric	R^2	RMSE	MAE
LOSS TNR (Ours)	**0.954**	**0.032**	**0.024**
RMIA (2 ref. models)	0.910	0.046	0.036
Loss AUC	0.844	0.051	0.052
LT-IQR AUC (Pollock et al., 2025)	0.806	0.067	0.046
Train-Test Gap	0.717	0.081	0.060

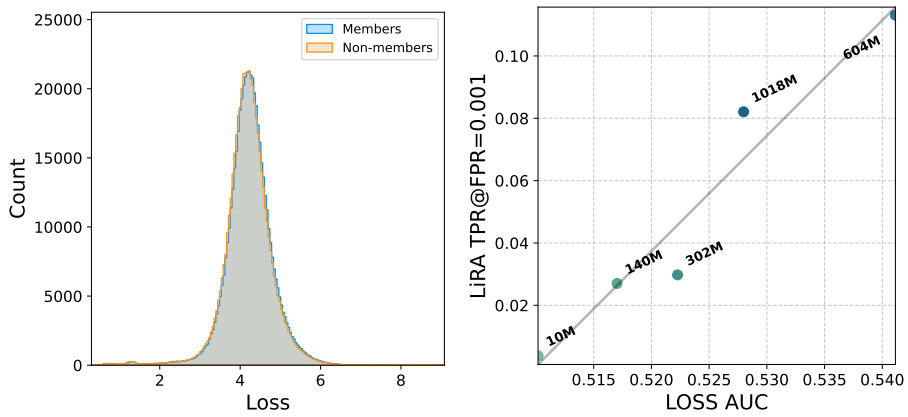

Figure 3. **Loss distributions and predictive relationship for LLMs.** Left: near-symmetric member/non-member overlap, unlike image classifiers. Right: LOSS AUC predicts LiRA TPR@FPR$=0.001$ across five GPT-2 sizes (RMSE$=0.01$).

with high member/non-member overlap (Figure 3, left)—no heavy tail exists for TNR to detect. LOSS AUC predicts LiRA TPR@FPR$=10^{-3}$ across all five model sizes with RMSE$=0.01$ (Figure 3, right), and the same fitted line holds across the full range of model sizes, suggesting the relationship is not sensitive to scale within this range.

4.3. Generalization Across Tasks

We further validate on text classification (IMDb, AGNews; TextCNN (Kim, 2014)), tabular classification (Census, Texas Hospital, Purchase (Shokri et al., 2017)), and tabular regression (California Housing, Bike Sharing). Fitting the linear predictor on image-only data and evaluating on these new domains yields R^2=0.97, RMSE$=0.03$, MAE$=0.03$, confirming the framework is not narrowly tied to a single task family.

5. Conclusion

We present a framework for estimating model-level vulnerability to SOTA membership inference attacks without reference models. We show that LiRA's per-sample signal decomposes into a variance-ratio term and a residual mean-shift term, with the relative contribution of each determined by how much training collapses model uncertainty at the trained sample. This places models on a continuum, with the shape of the loss distribution itself acting as a reference-free diagnostic for where a given model sits and which loss-based proxy of LiRA TPR is appropriate.

We instantiate the framework with two natural proxies. At the heavy-tailed end, where image classifiers sit, the LOSS attack TNR predicts LiRA TPR@FPR$=10^{-3}$ with RMSE 0.032 across 9 architectures and 4 datasets, outperforming the train-test accuracy gap, LT-IQR, and low-cost reference-model attacks. It predicts the TPR of RMIA and Attack R with similar accuracy, confirming it captures signal shared across this family of MIAs rather than a LiRA-specific artifact. At the symmetric end, where LLMs sit, the LOSS attack AUC predicts LiRA TPR with RMSE 0.01 across five GPT-2 sizes from 10M to 1B parameters.

While our estimator is empirically calibrated, it is not an arbitrary regression. Both LOSS TNR and LOSS AUC arise as population-level summaries of the same quantities that LiRA thresholds in the two limiting regimes. The observed linear relationship reflects shared dependence rather than incidental correlation.

Where SOTA attacks require training tens to hundreds of reference models, our framework uses only the target model's loss values on its train and test sets. We believe this will substantially help enable privacy risk assessments in practice, in particular during iterative development workflows and scenarios where multiple candidate models require assessment.

Limitations and future work. While we evaluate three reference-model MIAs (LiRA, RMIA, Attack R), we have not assessed transferability to attacks from fundamentally different families or to broader privacy threats such as reconstruction or inversion. Our experiments span diverse setups across the continuum, but the framework remains an estimator and may not generalize to all scenarios. Our LLM study covers five mid-sized architectures due to computational constraints; evaluating scaling to larger models and varied datasets is a natural next step. The framework is also proxy-agnostic, and other cheap MIA signals such as Min-K%++ (Zhang et al., 2024) are natural candidates for additional instances.

References

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Carlini, N., Terzis, A., Jagielski, M., Tramer, F., Papernot, N., and Zhang, C. The privacy onion effect: memorization is relative. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022b. Curran Associates Inc. ISBN 9781713871088.

Chen, D., Yu, N., and Fritz, M. Relaxloss: Defending membership inference attacks without losing utility. In *International Conference on Learning Representations*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of ai models. Adopted on 17 December 2024, December 2024. URL https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf. Version 1.0.

Fanaee-T, H. and Gama, J. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014.

Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

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. Strong membership inference attacks on massive datasets and (moderately) large language models, 2025. URL https://arxiv.org/abs/2505.18773.

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

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

(ICO), I. C. O. How should we assess security and data minimisation in ai?, 2025. URL https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/. Accessed: 2025-09-17.

Kelley Pace, R. and Barry, R. Sparse spatial autoregressions. *Statistics Probability Letters*, 33(3):291–297, 1997. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(96)00140-X. URL https://www.sciencedirect.com/science/article/pii/S016771529600140X.

Kim, Y. Convolutional neural networks for sentence classification. In Moschitti, A., Pang, B., and Daelemans, W. (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181/.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Leemann, T., Prenkaj, B., and Kasneci, G. Is my data safe? predicting instance-level membership inference success for white-box and black-box attacks. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R. (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Pollock, J., Shilov, I., Dodd, E., and de Montjoye, Y.-A. Free {Record-Level} privacy risk evaluation through {Artifact-Based} methods. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5525–5544, 2025.

Ponomareva, N., Vassilvitskii, S., Xu, Z., McMahan, B., Kurakin, A., and Zhang, C. How to dp-fy ml: A practical tutorial to machine learning with differential privacy. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 5823–5824, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599561. URL https://doi.org/10.1145/3580305.3599561.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Shilov, I., Meeus, M., and de Montjoye, Y.-A. The mosaic memory of large language models. *arXiv preprint arXiv:2405.15523*, 2024.

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

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

Tao, J. and Shokri, R. (token-level) informia: Stronger membership inference and memorization assessment for llms. *arXiv preprint arXiv:2510.05582*, 2025.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 3093–3106, 2022.

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

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zarifzadeh, S., Liu, P., and Shokri, R. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H. F., and Li, H. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In *NeurIPS*, 2015.

A. Appendix

B. Why each proxy is appropriate at its end of the continuum

Section 2 places models on a continuum between two limits, with the loss-distribution shape as the reference-free diagnostic for where a model sits. Figure 4 confirms this directly: image classifiers exhibit substantial member variance collapse and sit near the full-collapse end of the continuum, while LLMs exhibit much less and sit near the no-collapse end.

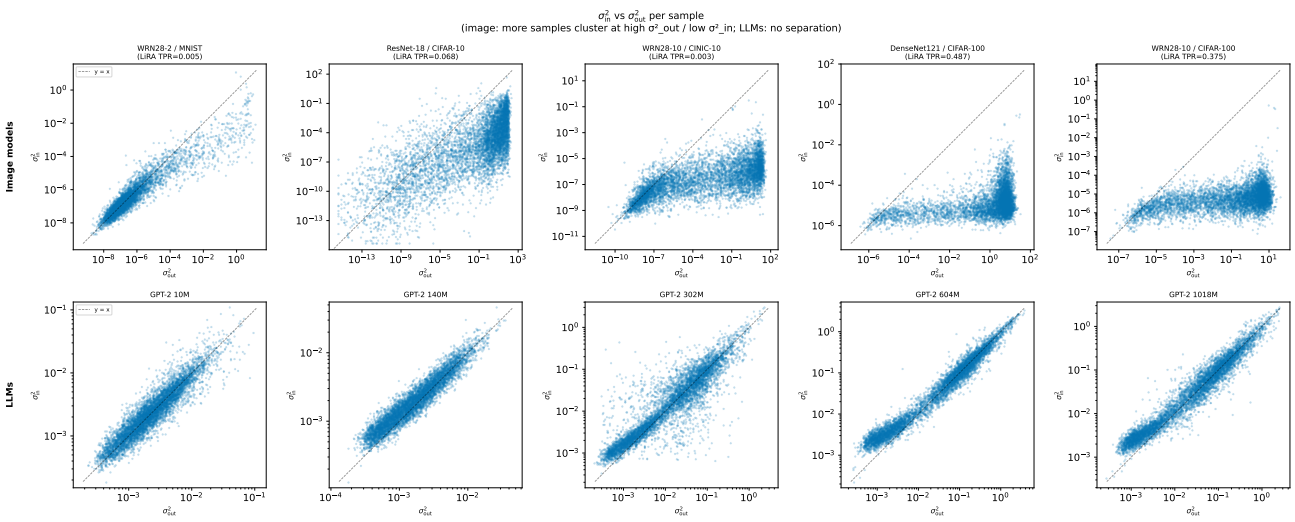

Figure 4. **Training collapses per-sample uncertainty in image models but not in LLMs.** Each panel plots σ^2_{in} against σ^2_{out} per sample. **Top row (image models):** members scatter visibly below $y = x$, indicating training has collapsed predictive variance at trained samples. **Bottom row (GPT-2):** points sit close to $y = x$ at all scales. σ^2_{out} itself spans several orders of magnitude, so the contrast is not that LLM uncertainty is uniformly low; training on a specific sample simply does not meaningfully reduce it. This places image models near the full-collapse end of the continuum (Section 2) and LLMs near the no-collapse end.

This appendix formalizes the corresponding proxy selection. The argument has the same shape in each regime. The LiRA decomposition (Eq. (1)) splits the per-sample score into a variance-ratio term and a residual term, with one or the other dominating at each end of the continuum. Whichever term dominates determines a population-level parameter that drives LiRA TPR. Both LOSS TNR and LOSS AUC are functions of the same parameter; they are different empirical estimators of the quantity LiRA itself thresholds. The question of which proxy is more reliable reduces to which estimates that parameter most efficiently from a finite sample.

We work at the population level. $\mathcal{L}_{\text{mem}}$ and $\mathcal{L}_{\text{non}}$ denote the population distributions of ℓ_{target} on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ respectively. The two proxies are

$$\text{AUC} = \Pr_{X\sim\mathcal{L}_{\text{mem}},\, Y\sim\mathcal{L}_{\text{non}}}(X < Y), \qquad \text{TNR}(\beta) = \Pr_{\ell\sim\mathcal{L}_{\text{non}}}(\ell > \tau_\beta), \tag{4}$$

with τ_β chosen so that $\Pr_{\ell\sim\mathcal{L}_{\text{mem}}}(\ell > \tau_\beta) = \beta$ (the LOSS attack false negative rate). Both are computed from ℓ_{target} on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ alone and require no reference models. We relate each proxy to LiRA TPR at fixed FPR β, in each of the two regimes derived in Section 2, and close with the unifying principle.

B.1. Weak-collapse limit

In the no-collapse limit of Eq. (**??**), per-sample variance is unaffected by training and the population-level loss distributions are single-mode and near-symmetric. We model them as $\mathcal{L}_{\text{mem}} \sim \mathcal{N}(\bar{\mu}_{\text{mem}}, \bar{\sigma}^2)$ and $\mathcal{L}_{\text{non}} \sim \mathcal{N}(\bar{\mu}_{\text{non}}, \bar{\sigma}^2)$ with shared variance, and write

$$d = \frac{\bar{\mu}_{\text{non}} - \bar{\mu}_{\text{mem}}}{\bar{\sigma}} \tag{5}$$

for the standardized mean separation. (Bars distinguish these population-level moments from the per-sample moments $\mu_{\text{mem}}(x), \sigma^2_{\text{mem}}(x)$ of Section 2.)

The FNR constraint $\Pr_{\ell\sim\mathcal{L}_{\text{mem}}}(\ell > \tau_\beta) = \beta$ gives $\tau_\beta = \bar{\mu}_{\text{mem}} + z_{1-\beta}\bar{\sigma}$, where Φ is the standard normal CDF and $z_q = \Phi^{-1}(q)$. Substituting into the TNR expression, splitting the fraction, and applying $1 - \Phi(x) = \Phi(-x)$:

$$\text{TNR}(\beta) = 1 - \Phi\left(\frac{\tau_\beta - \bar{\mu}_{\text{non}}}{\bar{\sigma}}\right) = 1 - \Phi(z_{1-\beta} - d) = \Phi(d - z_{1-\beta}). \tag{6}$$

Variances add for independent random variables, so $Y - X \sim \mathcal{N}(\bar{\mu}_{\text{non}} - \bar{\mu}_{\text{mem}}, 2\bar{\sigma}^2)$. Then

$$\text{AUC} = \Pr(Y - X > 0) = \Phi(d/\sqrt{2}). \tag{7}$$

The $\sqrt{2}$ comes from the variance of $Y - X$ being $2\bar{\sigma}^2$ rather than $\bar{\sigma}^2$.

LiRA is a per-sample attack: at each x it computes Λ from Eq. (1) and thresholds it. Eq. (**??**) reduces Λ to a function of the per-sample mean shift $\delta(x) = \mu_{\text{mem}}(x) - \mu_{\text{non}}(x)$. With the standard noise model $y = \mu_{\text{mem}}(x) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2(x))$ for member samples,

$$\Lambda = \frac{\delta(x)^2}{2\sigma^2(x)} + \frac{\delta(x)}{\sigma^2(x)}\epsilon.$$

Λ is conditionally Gaussian at each x, with mean and variance both proportional to $\delta(x)^2/\sigma^2(x)$. Define $d^2 := \mathbb{E}_x[\delta(x)^2/\sigma^2(x)]$. Under shared variance this reduces to the d of Eq. (5), and the marginal distribution of Λ on members is approximately $\mathcal{N}(d^2/2, d^2)$. Computing the same way for non-members gives $\mathcal{N}(-d^2/2, d^2)$. The standardized separation between these two distributions is d, and the same Gaussian-tail computation as for TNR yields

$$\text{TPR}_{\text{LiRA}}(\beta) \approx \Phi(d - z_{1-\beta}). \tag{8}$$

Comparing Eqs. (6), (7), and (8): TNR and TPR_{LiRA} are the *same function* of d, and AUC is a different but strictly monotone function of the same d. All three encode the same population parameter through different empirical statistics. In the population limit either proxy is equally informative.

Signal-to-noise. The proxies separate in finite samples. LiRA TPR is itself a function of d, and any proxy noise propagates through the calibration into the predicted LiRA TPR.

TNR is a binomial proportion. The empirical estimator counts $\mathcal{D}_{\text{test}}$ samples above τ_β, with standard error

$$\text{SE}(\widehat{\text{TNR}}) = \sqrt{\frac{\text{TNR}(1 - \text{TNR})}{|\mathcal{D}_{\text{test}}|}} \approx \sqrt{\frac{\text{TNR}}{|\mathcal{D}_{\text{test}}|}} \tag{9}$$

when TNR is small. At the low FNR of interest ($\beta = 10^{-3}$) and $d = 0.5$, Eq. (6) gives $\text{TNR}(10^{-3}) \approx 5 \times 10^{-3}$. With $|\mathcal{D}_{\text{test}}| = 10^4$ the standard error is roughly 7×10^{-4}. Relative noise is about 14%.

AUC is a U-statistic over pairs. Every (member, non-member) pair contributes, so

$$\text{SE}(\widehat{\text{AUC}}) \sim \frac{1}{\sqrt{|\mathcal{D}_{\text{train}}| \cdot |\mathcal{D}_{\text{test}}|}} \tag{10}$$

is independent of β (Hanley & McNeil, 1982). For $|\mathcal{D}_{\text{train}}| = |\mathcal{D}_{\text{test}}| = 10^4$ and $d = 0.5$, AUC is around 0.638 and the standard error is order 10^{-4}. Relative noise is well under 1%.

The two estimators have different effective sample sizes. TNR uses only the samples that fall above τ_β, which at low FNR is a tiny fraction of the test set. AUC uses every pair. When both estimate the same parameter, the bulk-based estimator wins because it makes use of more samples. AUC is the more reliable predictor of LiRA TPR in this limit.

B.2. Strong-collapse limit

In the full-collapse limit of Eq. (**??**), training collapses per-sample variance on members. The non-member distribution develops a heavy tail. The member distribution loses part of that tail to training-induced migration into the head. A mean-shift model is no longer appropriate: the difference between the two distributions is a movement of mass between head and tail.

We model this with a two-component mixture. $\mathcal{L}_{\text{bulk}}$ and $\mathcal{L}_{\text{tail}}$ are supported on $[0, \tau_{\text{tail}}]$ and $[\tau_{\text{tail}}, \infty)$ respectively, and we assume the two components are well-separated:

$$\mathcal{L}_{\text{non}}(\ell) = (1 - \pi_{\text{non}})\, \mathcal{L}_{\text{bulk}}(\ell) + \pi_{\text{non}}\, \mathcal{L}_{\text{tail}}(\ell), \tag{11}$$

$$\mathcal{L}_{\text{mem}}(\ell) = (1 - \pi_{\text{mem}})\, \mathcal{L}_{\text{bulk}}(\ell) + \pi_{\text{mem}}\, \mathcal{L}_{\text{tail}}(\ell), \tag{12}$$

with $\pi_{\text{mem}} \ll \pi_{\text{non}}$. The mixture weights π_{non} and π_{mem} represent the probability that a randomly drawn non-member (resp. member) loss exceeds τ_{tail}. Their difference $\pi_{\text{non}} - \pi_{\text{mem}}$ is the population that has migrated from tail to head under training.

In the full-collapse limit, the variance-ratio term of Eq. (1) dominates LiRA's score and is largest exactly on samples where $\sigma^2_{\text{mem}}(x) \ll \sigma^2_{\text{non}}(x)$. These samples are the tail-to-head fraction. When training collapses σ^2 on a sample, that sample's loss drops sharply, and the population-level effect is to remove mass from the non-member tail (where it would have sat unseen) and add it to the member head (where it now sits trained). At low FPR, LiRA's threshold selects these samples with high probability and rejects the bulk, so

$$\text{TPR}_{\text{LiRA}}(\beta) \;\propto\; \pi_{\text{non}} - \pi_{\text{mem}} \tag{13}$$

up to a small error and a threshold-dependent constant.

Setting $\tau_\beta = \tau_{\text{tail}}$, the FNR constraint gives $\beta = \pi_{\text{mem}}$ (the member tail mass is π_{mem} by construction) and

$$\text{TNR}(\pi_{\text{mem}}) = \pi_{\text{non}}. \tag{14}$$

TNR is a direct estimate of π_{non}. π_{mem} enters only through the threshold β.

For AUC, condition on which mixture component each draw came from. Bulk-vs-tail comparisons are decided by the component, since bulk and tail have disjoint support: a draw from the bulk is always less than a draw from the tail. The four cases are:

- Both in bulk (prob. $(1 - \pi_{\text{mem}})(1 - \pi_{\text{non}})$): $\Pr(X < Y) = 1/2$.
- Member in bulk, non-member in tail (prob. $(1 - \pi_{\text{mem}})\pi_{\text{non}}$): $\Pr(X < Y) = 1$.
- Member in tail, non-member in bulk (prob. $\pi_{\text{mem}}(1 - \pi_{\text{non}})$): $\Pr(X < Y) = 0$.
- Both in tail (prob. $\pi_{\text{mem}}\pi_{\text{non}}$): $\Pr(X < Y) = 1/2$.

Summing the contributions:

$$\text{AUC} = \tfrac{1}{2}(1 - \pi_{\text{mem}})(1 - \pi_{\text{non}}) + (1 - \pi_{\text{mem}})\pi_{\text{non}} + \tfrac{1}{2}\pi_{\text{mem}}\pi_{\text{non}} = \tfrac{1}{2} + \tfrac{1}{2}(\pi_{\text{non}} - \pi_{\text{mem}}). \tag{15}$$

AUC depends on $\pi_{\text{non}} - \pi_{\text{mem}}$, scaled by $1/2$ and constrained to $[\frac{1}{2}, 1]$.

Comparing Eqs. (13), (14), and (15): all three quantities are linear in $\pi_{\text{non}} - \pi_{\text{mem}}$, with different slopes and dynamic ranges. As before, they are different empirical estimators of the same theoretical quantity.

Signal-to-noise revisited. The question reduces to which proxy estimates the shared parameter most efficiently. The answer reverses for two reasons.

TNR is no longer a small-probability event. π_{non} is typically much larger in image classifiers (order 0.1 to 0.5), so by Eq. (9) the standard error is roughly $\sqrt{\pi_{\text{non}}/|\mathcal{D}_{\text{test}}|}$, which is small relative to its value. The variance penalty that crippled TNR in the weak-collapse limit is gone.

AUC has lost most of its signal. In the weak-collapse limit AUC was a strictly monotone function of d with no inherent ceiling besides Φ's asymptotes. In the strong-collapse limit it moves at slope $1/2$ in $\pi_{\text{non}} - \pi_{\text{mem}}$ and is bounded above by 1 (Eq. (15)). As vulnerability scales up, TNR can move from β to 1. AUC can only move from $1/2$ to 1. Per unit of underlying signal, AUC's response is half of TNR's, and AUC's range is half of TNR's.

AUC retains a small efficiency advantage from using every pair of samples. TNR's signal per unit of $\pi_{\text{non}} - \pi_{\text{mem}}$ is twice AUC's, and TNR's range is twice as wide. The signal in this regime is concentrated in the tail, and AUC dilutes it by averaging over the entire distribution. The slope and dynamic-range advantages outweigh AUC's residual efficiency advantage. Empirically (Table 1), TNR outperforms AUC at this end of the continuum.

B.3. Summary

Across both regimes, LiRA's decomposition isolates a single dominant term whose population-level value drives TPR, and both LOSS-based proxies are different empirical estimators of that same value. AUC integrates over the entire loss distribution. TNR localizes to the tail. Which estimator has higher signal-to-noise depends on where in the loss distribution the dominant term's signal is concentrated.

When the dominant term is a mean shift spread across the whole distribution, AUC's whole-distribution integration uses every sample efficiently while TNR's tail localization wastes most of the data on a small-probability event. When the dominant term is concentrated in the tail, TNR's localization captures it undiluted while AUC averages it down by mixing with the rest of the low-signal population. The shape of the loss distribution shows which term dominates, and therefore which estimator is appropriate.

Selecting a proxy in practice. A practitioner has access to the target model and its losses on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, but no ground-truth LiRA TPR to validate against. Proxy selection must therefore rely on diagnostics computed directly from these loss distributions. We suggest the following signals:

1. **KS-based goodness-of-fit diagnostic (primary).** We test whether the non-member loss distribution is well-approximated by a mean-shifted version of the member distribution. Concretely, we fit a constrained Gaussian model $N(\mu + \Delta, \sigma_{\text{mem}})$ to the non-member losses, where μ and σ_{mem} are estimated from $\mathcal{L}_{\text{mem}}$ and Δ is a fitted shift. We then compute the Kolmogorov–Smirnov (KS) statistic between the empirical CDF of $\mathcal{L}_{\text{non}}$ and this constrained Gaussian.

   A small KS value indicates that the mean-shift assumption is adequate (weak-collapse regime), and LOSS AUC is appropriate. A large KS value indicates a systematic deviation from this model e.g., heavy tails, skew, or multimodality. This suggests the presence of a tail component (strong-collapse regime), where LOSS TNR is preferred. In our experiments, KS cleanly separates model families: GPT-2 runs have KS 0.049–0.069, while image-classifier runs have KS 0.120–0.888.

2. **Shape of the loss distributions.** As a qualitative diagnostic, a visible high-loss tail on $\mathcal{L}_{\text{non}}$ that is absent from $\mathcal{L}_{\text{mem}}$ suggests the strong-collapse regime. In contrast, near-symmetric distributions that differ primarily in mean indicate the weak-collapse regime.

3. **Variance ratio $\sigma^2_{\text{non}}/\sigma^2_{\text{mem}}$.** A one-number summary of distributional differences. Ratios meaningfully above 1 indicate a tail component; ratios near 1 indicate the weak-collapse regime.

4. **Empirical TNR at low FNR.** If $\widehat{\text{TNR}}(\beta)$ is substantially above β, the LOSS attack is detecting tail samples and TNR is in regime. If $\widehat{\text{TNR}}(\beta)$ is at or near β, no detectable tail signal is present and AUC is the fallback.

5. **Empirical TNR at low FNR.** If $\widehat{\text{TNR}}(\beta)$ is substantially above β, the LOSS attack is detecting a tail signal and TNR is in regime. If $\widehat{\text{TNR}}(\beta)$ is close to β, no detectable tail signal is present and AUC is the fallback.

In practice, the KS diagnostic provides the most general test, as it captures any deviation from the mean-shift Gaussian model, not only variance inflation. The empirical regimes (image classifiers near strong collapse, LLMs near weak collapse) are consistent across all signals. For models in the middle of the continuum, the cleanest approach is to compute both proxies on a small calibration set and select the one with lower calibration error (e.g., RMSE).

C. Metric performance along the no-collapse–collapse continuum

We study how the predictive performance of LOSS TNR and LOSS AUC varies as the member and non-member loss distributions of the target model become more similar. To simulate this, we train target models using RelaxLoss (Chen et al.), a privacy defense that explicitly encourages the member and non-member loss distributions to converge during training, across 5 architectures (ResNet-20, DenseNet-121, WRN-28-2, VGG-16, VGG-11) and 2 datasets (CIFAR-10, CIFAR-100). We vary the defense strength $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ and instantiate LiRA against each resulting target model.

Since the same α value has differing effects across datasets, we use the KS value described in X to measure the distance between the member and non-member loss distributions. Specifically, we fit a constrained Gaussian $N(\mu + \Delta, \sigma_{\text{mem}})$ to the

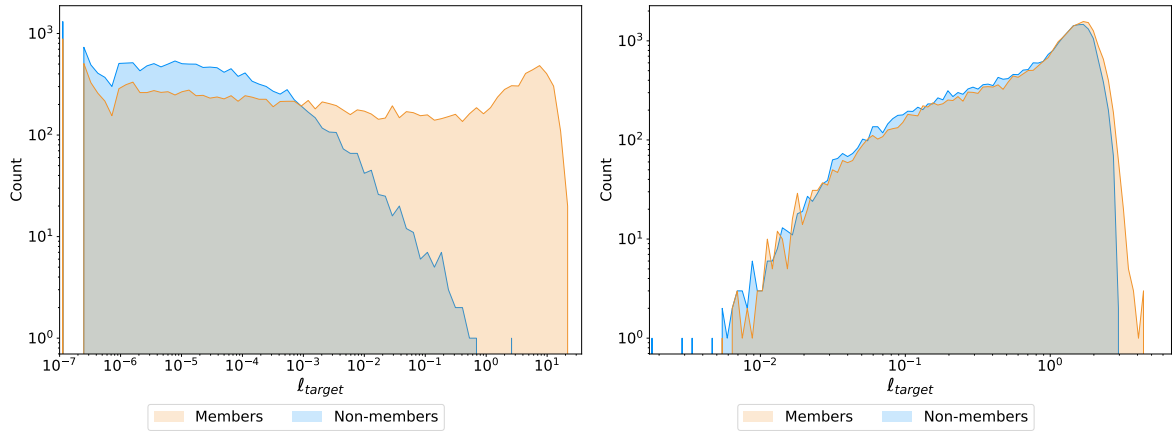

Figure 5. Member and non-member loss distributions for VGG-16 trained on CIFAR-10 with (a) no defense, $s(\mathcal{L}_{\text{mem}}, \mathcal{L}_{\text{non}}) = 0.87$, and (b) RelaxLoss with $\alpha = 1.0$, $s(\mathcal{L}_{\text{mem}}, \mathcal{L}_{\text{non}}) = 0.14$.

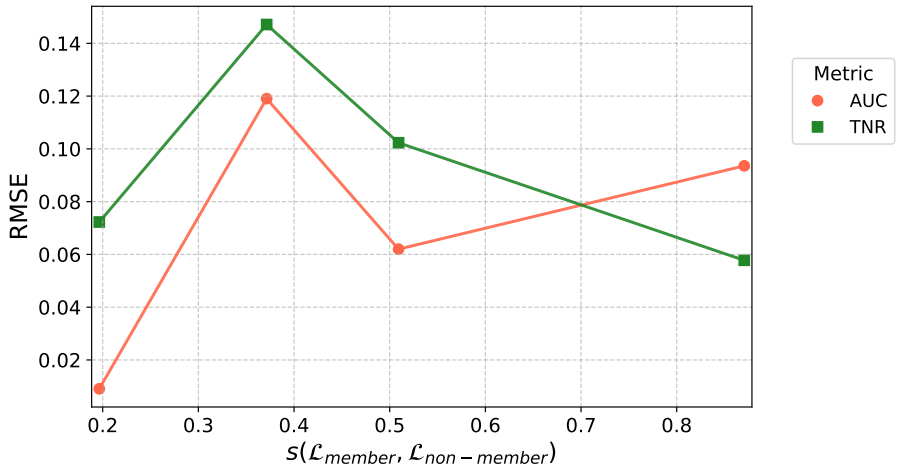

Figure 6. RMSE for predicting LiRA TPR@FPR=10^{-2} with LOSS AUC and LOSS TNR across different member–non-member training distributions. Higher s indicates that the two distributions are more different (strong-collapse limit), and lower values indicate that they are more similar (low-collapse limit).

non-member losses. μ and σ_{mem} are estimated from $\mathcal{L}_{\text{mem}}$ and $\Delta = E[\mathcal{L}_{\text{non}}] - \mu$ is a fitted shift. We use the KS statistic between the empirical CDF of $\mathcal{L}_{\text{non}}$ and this Gaussian directly as a distance measure,

$$s(\mathcal{L}_{\text{mem}}, \mathcal{L}_{\text{non}}) = \sup_{x \in \mathbb{R}} \left| \hat{F}_{\text{non}}(x) - \Phi\left(\frac{x - (\mu + \Delta)}{\sigma_{\text{mem}}} \right) \right|,$$

where $\hat{F}_{\text{non}}$ is the empirical CDF of $\mathcal{L}_{\text{non}}$ and Φ is the standard normal CDF. A low value of s indicates highly similar distributions (weak-collapse limit), while higher values indicate distributions with different shapes (strong-collapse limit).

Figure 6 shows the RMSE of LOSS TNR and LOSS AUC as predictors of LiRA TPR@FPR=10^{-2} across this range. When the non-member distribution is substantially heavier-tailed than the member distribution (high s), LOSS TNR is the stronger predictor. As the two distributions converge, LOSS AUC takes over.

D. Loss distributions across all setups

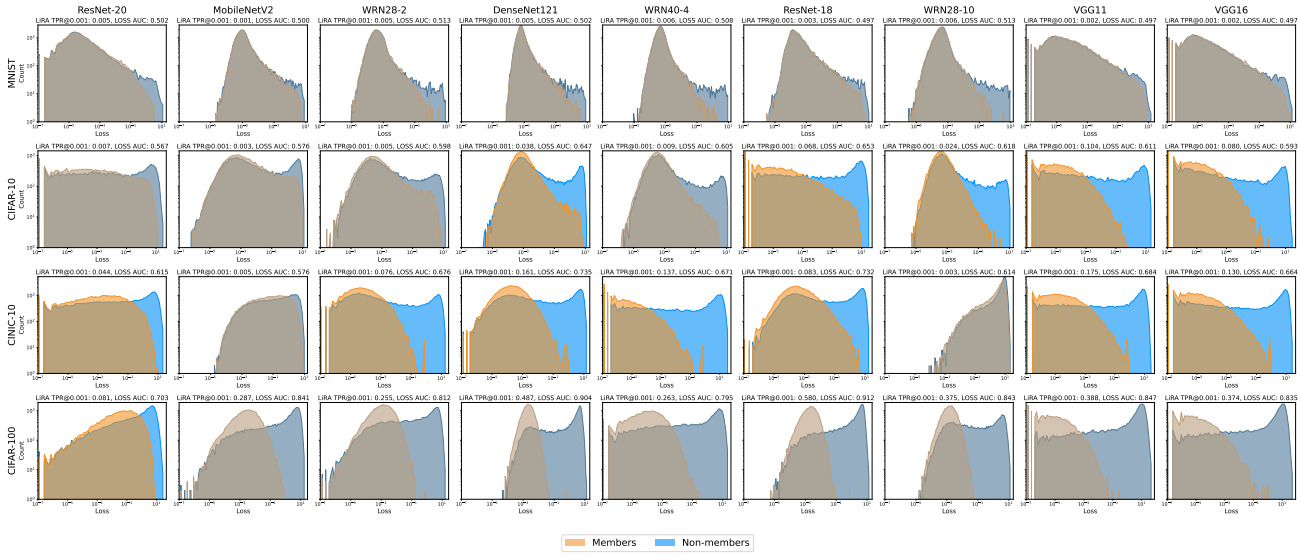

Figure 7. Histograms showing ℓ_{target} distributions for training set members (orange) and non-members (blue) in log-log scale across all setups. The distributions demonstrate clear separation between members and non-members, with members typically having lower loss values.

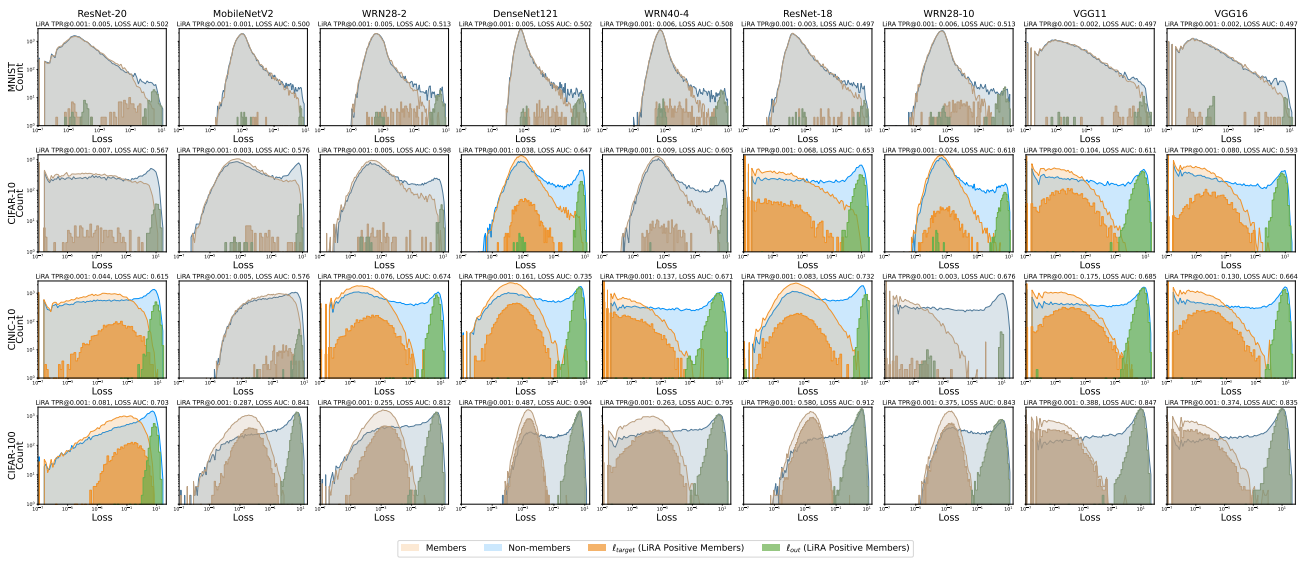

Figure 8. ℓ_{target} distributions for training set members and non-members across all setups. Orange histogram shows member losses, blue shows non-member losses, with density curves overlaid. Green bars indicate the OUT model mean (ℓ_{out}) for points identified by LiRA at FPR=0.001.

E. Varying FPRs

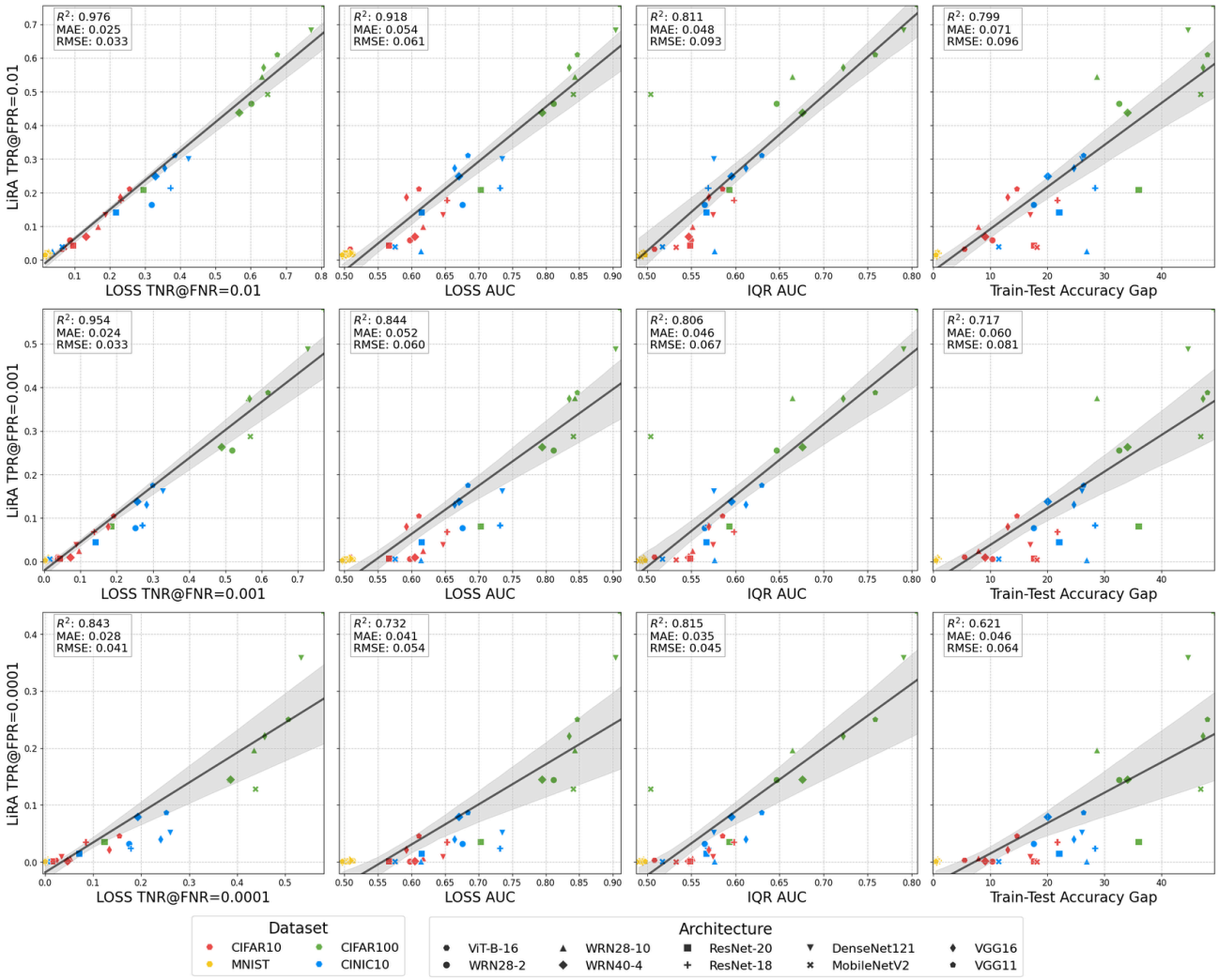

Figure 9. Comparison of different metrics for predicting LiRA TPR at FPRs of 0.01, 0.001, 0.0001. Each panel shows the relationship between a predictor metric (x-axis) and LiRA TPR (y-axis) across all model-dataset combinations. Linear fits (solid lines) with 97.5% confidence intervals (shaded regions) show goodness-of-fit.

F. Sensitivity to selection of LOSS FNR threshold

We select FPR=FNR across the paper, as it is a natural choice. We here study the sensitivity of LOSS TNR's predictive capability to the choice of FNR threshold. Following the experimental setup described in Section **??**, we evaluate how well LOSS TNR predicts LiRA's TPR across four different false positive rate thresholds: 0.1, 0.01, 0.001, and 0.0001. For each LiRA FPR setting, we fit a linear function to predict the corresponding LiRA TPR from LOSS TNR values computed at varying FNR thresholds, and measure prediction quality using the RMSE.

Figure 10 shows that RMSE remains relatively stable across a broad range of FNR values, demonstrating that LOSS TNR's predictive power is robust to the specific threshold selection. Notably, setting the FNR equal to the target LiRA FPR (green stars) achieves a similar performance compared to the FNR that minimizes RMSE (yellow stars), showing it to be a good and principled choice. The sharp degradation in RMSE only occurs at very high FNR values (approaching 1.0), where the LOSS threshold becomes too relaxed to provide a meaningful signal.

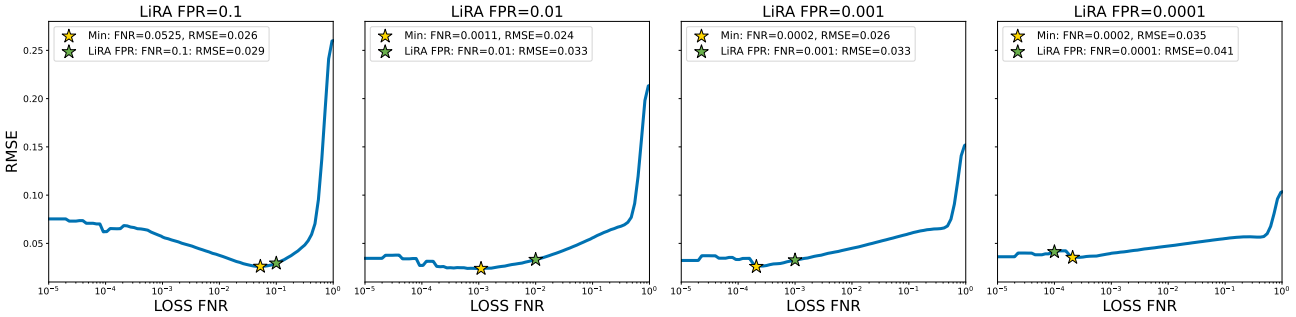

Figure 10. RMSE of LOSS TNR@different FNR values as a predictor of LiRA TPR at a fixed FPR. The yellow star indicates the FNR that achieves the lowest RMSE, while the green star shows performance when setting the FNR to equal to the LiRA FPR. Across all settings, the RMSE difference between the optimal selection and FNR=FPR is minimal.

G. Varying the number of reference models

We study the effectiveness of LOSS TNR at estimating the TPR@FPR of attacks of varying strength. We instantiate LiRA with 4, 8, 16, 32 and 64 reference models and report the resulting linear model for each value of K, along with the slopes α.

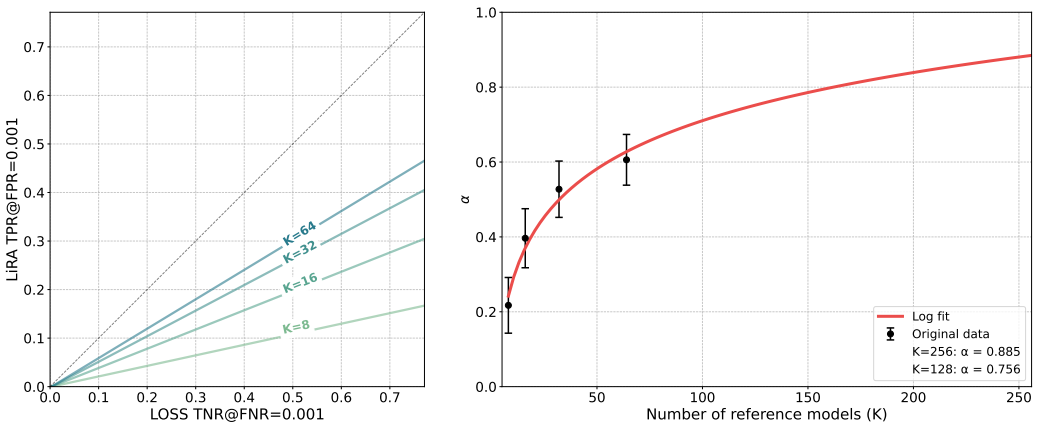

Figure 11. Performance of linear models for predicting LiRA with varying numbers of reference models (K). Left panel shows the correlation between LOSS TNR (x-axis) and LiRA TPR@0.001 (y-axis) across different K values, with linear regression fits shown for each condition. Right panel shows slope parameter (α) for different K. The error bars show the 97.5% confidence intervals obtained with boostrapping.

Figure 11 shows the risk to steadily increase with K as the attack becomes more confident at identifying members. As noticed in previous work though, TPR@low FPR experiences decreasing returns as the number of reference models increase which we can now model to estimate the risk against stronger attackers.

H. Fitting different functions

So far we have estimated LiRA TPR with a simple linear model. We now study whether non-linear models lead to better risk estimates. Identifying non-members is indeed a matter of finding a good threshold to separate the tail of the loss distribution from the rest, while identifying members is dependent on reference models and is likely to be a more difficult task. We indeed find empirically that the LOSS TNR is typically higher than LiRA TPR at the same FPR.

We thus compute goodness-of-fit metrics for convex functions such as a two-parameter polynomial, power-law, and exponential functions.

Figure 12 shows that the exponential function $a(e^{bx}-1)$ achieves the best fit, outperforming the linear model. An exponential fit would imply that as LOSS TNR increases, member identification becomes easier as the model memorizes more difficult samples. Initial gains in LOSS TNR produce modest improvements in LiRA TPR, but these improvements accelerate as

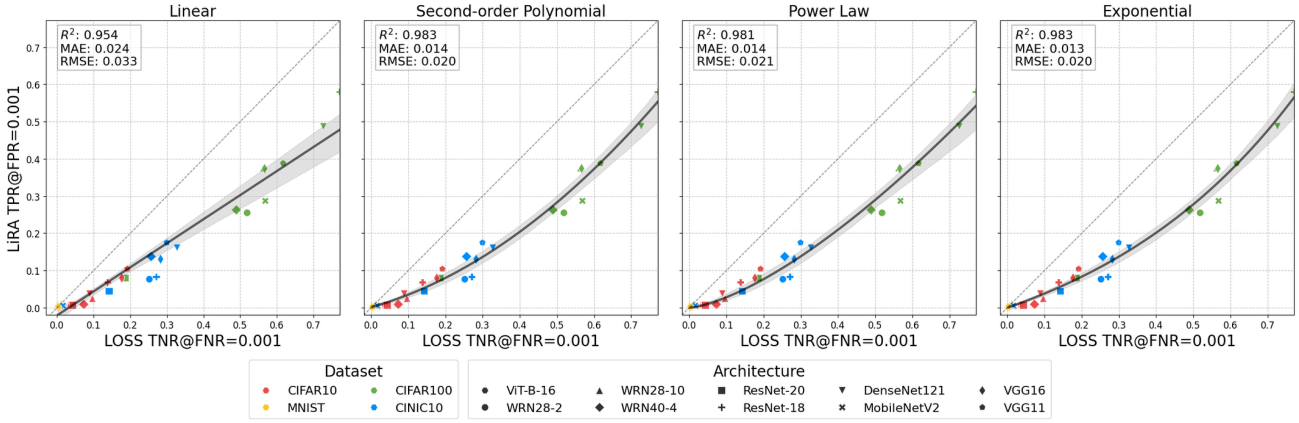

Figure 12. Comparison of regression models for predicting TPR@0.001 with LOSS TNR@0.001 evaluated on image recognition setups. The exponential function achieves the lowest RMSE of 0.02, compared to 0.033 achieved by the linear fit.

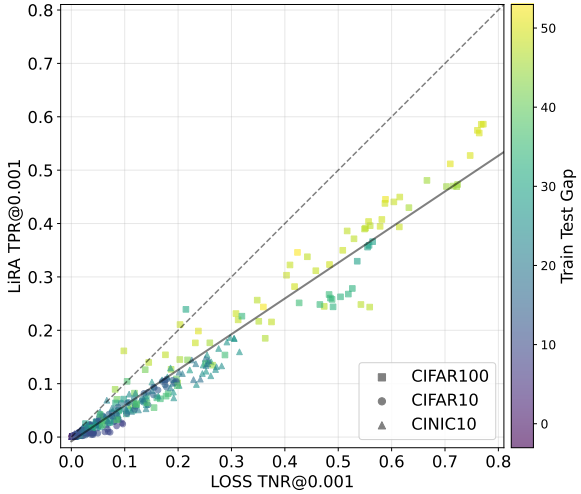

Figure 13. LOSS TNR reliably predicts LiRA TPR across varying overfitting levels.

memorization increases.

I. Robustness across levels of generalization gaps

Prior work has identified overfitting, typically measured by the train-test accuracy gap, as a key factor influencing membership inference vulnerability (Yeom et al., 2018; Carlini et al., 2022a).

We now study the effectiveness of our method for predicting the vulnerability of models with widely varying train-test gaps, ranging from near-zero (minimal overfitting) to over 50 percentage points (severe overfitting). Specifically, we treat intermediate training checkpoints from each setup in our main experiment as additional target models, instantiate LiRA against each target, and assess how well our method predicts its vulnerability. Figure 13 shows that LOSS TNR remains a reliable predictor of model vulnerability across this full range of generalization gaps.

This consistency demonstrates that LOSS TNR captures memorization patterns much richer than that of the overall generalization performance. While models with larger train-test gaps tend to have higher vulnerability (evidenced by the color gradient), TNR provides accurate risk estimates across a wide spectrum.

J. Predicting vulnerability to different attacks

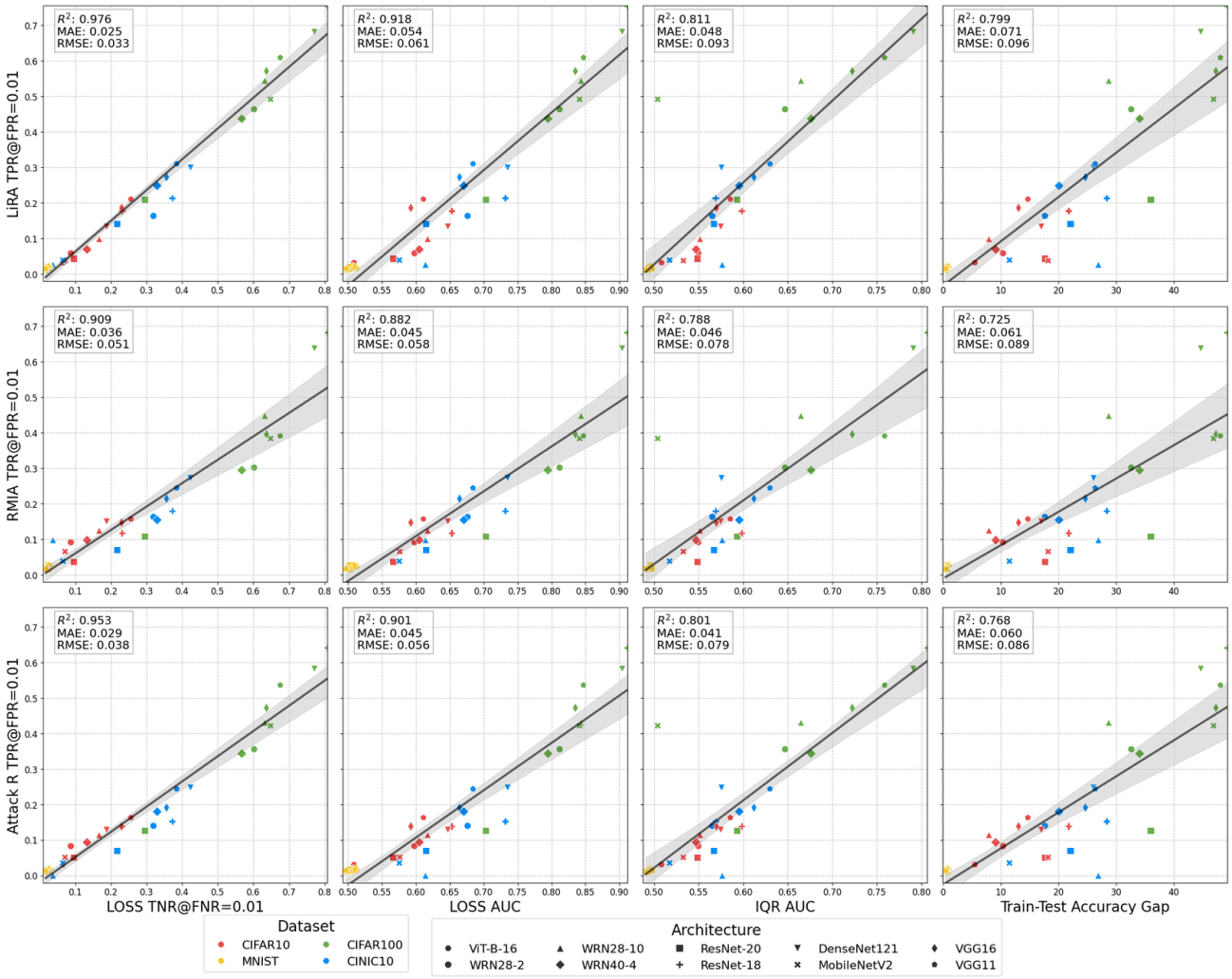

Figure 14. Comparison of different metrics for predicting the TPR at FPR=0.01 for LiRA, RMIA and Attack-R with 64 reference models. Each panel shows the relationship between a predictor metric (x-axis) and the attack TPR (y-axis) across all model-dataset combinations. Linear fits (solid lines) with 97.5% confidence intervals (shaded regions) show goodness-of-fit.

To evaluate whether our method generalizes to other attacks within the same family, we estimate the TPR@0.01 for RMIA and Attack R using 64 reference models (for RMIA: 32 IN and 32 OUT; for Attack R: 64 OUT). Figure 14 demonstrates that TNR serves as an effective predictor for both RMIA and Attack R (two reference-model-based MIAs). Since these attacks share a common approach of using reference models to estimate counterfactuals for individual samples, it follows logically that TNR captures the same underlying signal across attacks in this family.

K. Predicting vulnerability across different tasks

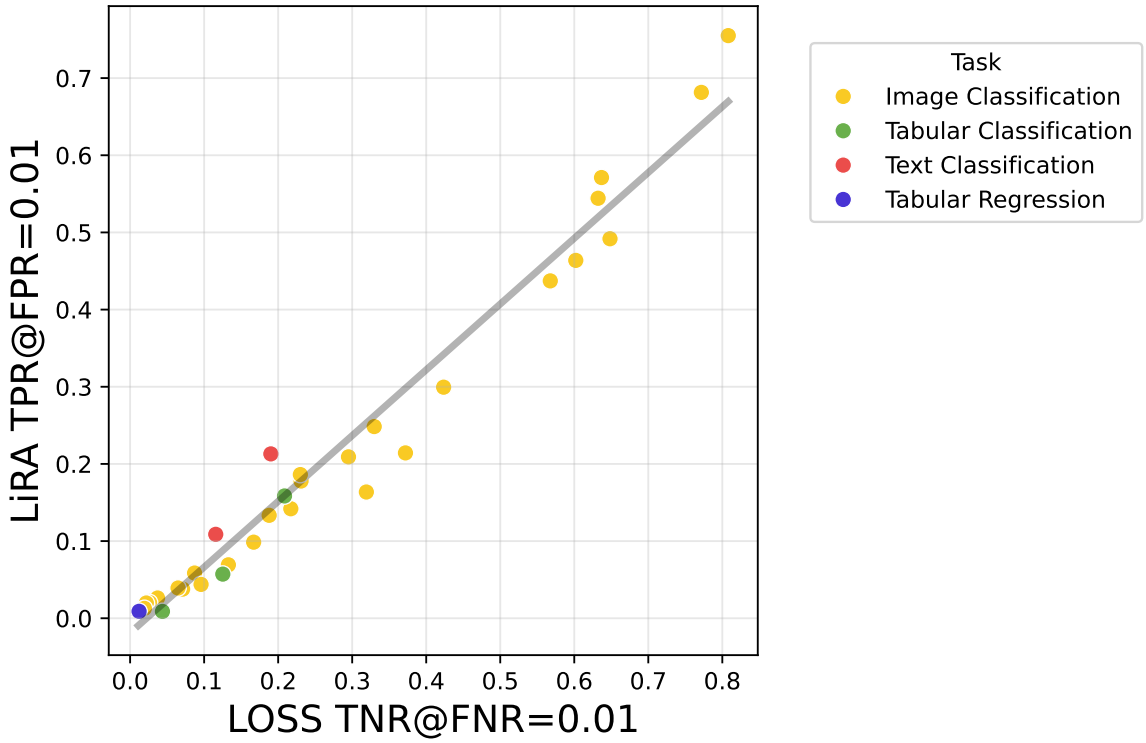

Figure 15. LOSS TNR successfully predicts model-level vulnerability to LiRA across tasks. Our metric achieves $R^2 = 0.97$. Fitting the prediction line on image-only tasks leads to RMSE=0.03 and MAE=0.03 when predicting vulnerability of non-image tasks.

To evaluate how our method performs for tasks beyond image classification, we instantiate it for 3 distinct tasks, specifically:

1. **Text classification**
   (a) IMDb movie reviews dataset (Maas et al., 2011), TextCNN following Kim (2014).
   (b) AGNews dataset (Zhang et al., 2015), TextCNN following Kim (2014).
2. **Tabular classification**
   (a) US Census dataset (Becker & Kohavi, 1996), 3-layer feedforward neural network with ReLU activation
   (b) Texas hospital stays dataset [1], 4-layer feedforward neural network with Tanh activation
   (c) Purchase dataset [2] as processed by Shokri et al. (2017), 4-layer feedforward neural network with Tanh activation
3. **Tabular regression**
   (a) California housing prices dataset, 3-layer feedforward neural network with ReLU activation (Kelley Pace & Barry, 1997)

[1] https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm
[2] https://kaggle.com/c/acquire-valued-shoppers-challenge/data

(b) Bike sharing dataset (Fanaee-T & Gama, 2014), 3-layer feedforward neural network with ReLU activation

Figure 15 shows that LOSS TNR remains a strong predictor of LiRA TPR, and, importantly, that the same fitted line is valid for these setups as for the image classification ones ($R^2 = 0.97$, RMSE$= 0.03$, MAE$= 0.03$). This suggests that the mapping is not narrowly tied to a single task family and remains informative under moderate task shift.

L. Assets

Dataset / Model	Details	Citation & URL	License
Image Classification			
CIFAR-10	Standard	(Krizhevsky et al., 2009) https://www.cs.toronto.edu/~kriz/cifar.html	MIT
CIFAR-100	Standard	(Krizhevsky et al., 2009) https://www.cs.toronto.edu/~kriz/cifar.html	MIT
MNIST	Handwritten digits	(LeCun et al., 2002) http://yann.lecun.com/exdb/mnist/	Custom (permissive, research use)
ResNet / WRN / ViT / DenseNet / MobileNetV2 / VGG	PyTorch impl.	(He et al., 2016); (Zagoruyko & Komodakis, 2016); (Dosovitskiy et al., 2020); (Huang et al., 2017); (Sandler et al., 2018); (Simonyan & Zisserman, 2015) https://pytorch.org/vision	Apache 2.0
Text Classification			
IMDb	Sentiment	(Maas et al., 2011) https://ai.stanford.edu/~amaas/data/sentiment/	Custom
AG News	Topic classification	(Zhang et al., 2015) https://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html	CC BY-SA
TextCNN	Model	(Kim, 2014) https://github.com/yoonkim/CNN_sentence	Open-source
Tabular Classification			
Adult	Income prediction	(Becker & Kohavi, 1996), UCI https://archive.ics.uci.edu/ml/datasets/adult	CC BY 4.0
Texas Hospital	Discharge data	Texas DSHS https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm	Restricted
Purchase	Retail dataset	(Shokri et al., 2017) https://kaggle.com/c/acquire-valued-shoppers-challenge/data	Kaggle terms
MLP (Tabular)	3–4 layer NN	PyTorch https://pytorch.org	BSD-style
Tabular Regression			
California Housing	Regression	(Kelley Pace & Barry, 1997) https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html	CC BY 4.0
Bike Sharing	Regression	(Fanaee-T & Gama, 2014) https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset	CC BY 4.0
MLP (Regression)	3-layer NN	PyTorch https://pytorch.org	BSD-style
Language Modeling			
C4	219-sample subset	(Raffel et al., 2020) https://www.tensorflow.org/datasets/catalog/c4	ODC-BY
GPT-2	Fine-tuned	(Radford et al., 2019) https://github.com/openai/gpt-2	MIT

Table 2. Datasets and model assets used in this work.