Rare, Distinctive, Memorized: Auditing Memorization in Fine-Tuned Medical Foundation Models

Anonymous Authors[1]

Abstract

Medical foundation models fine-tuned on private patient data risk leaking individual training samples; in this regime a single recoverable image is a privacy violation. We audit per-class memorization in fine-tuned medical foundation models with a loss-difference memorization score: each canary's loss is compared between a model trained with the canary and an otherwise identical model trained without it. Memorization occurs in every architecture-dataset cell we test. Interestingly, the rarest class is not the most memorized: the rarity-monotonicity assumed by prior memorization work outside medical imaging is broken in this regime. Instead, visual distinctiveness contributes to memorization beyond rarity in our setting: the same controlled grayscale intervention applied to two classes at comparable rarity shifts memorization in opposite directions, and on a balanced dataset $M(x)$ rises from a near-zero baseline (0.004) to high memorization (0.480) under the same intervention. DP-LoRA at $\varepsilon = 1$ reduces leakage from the most-memorized class by 90% in our setup while preserving usable accuracy on focused diagnostic tasks, with the protection driven by the DP noise rather than the parameter restriction. These findings point towards more privacy-preserving adaptation of medical foundation models.

1. Introduction

Hospitals rarely train foundation models from scratch. Legal and ethical constraints prevent pooling patient data across institutions, so the deployment pattern in medical imaging is to take a publicly available foundation model (ResNet, ViT, MedSAM, BiomedCLIP) and fine-tune it on a private, institution-specific corpus. Two features distinguish this regime from general-purpose vision pipelines. First, rare-disease classes carry the diagnostic signal that practitioners most want the model to learn; they are invaluable training data, not outliers to discard or down-weight. Second, privacy is evaluated at the individual-patient level: a single leaked training image is a privacy violation regardless of population statistics, and recent work has shown that representations from medical foundation models alone enable patient re-identification at rates of 25 to 46% across imaging modalities (Nebbia et al., 2025).

Memorization is the underlying reason fine-tuning can leak individual training samples. Deep networks have been shown to fit random labels at zero training error (Zhang et al., 2017), larger models memorize faster and forget less (Tirumala et al., 2022), and individual samples can be extracted from production language models (Carlini et al., 2021); detection has matured around membership-inference attacks (Shokri et al., 2017; Yeom et al., 2018; Salem et al., 2019; Carlini et al., 2022). The corresponding picture for vision foundation models fine-tuned on patient data is far less developed. For vision classifiers, class rarity has been shown to drive memorization risk (Feldman, 2020; Carlini et al., 2023); we investigate whether this also holds in our medical fine-tuning setup and find that rarity is necessary but not sufficient: the rarest class is not the most memorized.

We show that rarity is necessary but not sufficient for memorization in this regime, and we test *visual distinctiveness* as one candidate mechanism for the residual variance. By distinctiveness we mean the visual basis on which a class differs from the training distribution. We do not propose an a priori quantitative metric for distinctiveness; instead, we manipulate it through a controlled colour-versus-structure intervention. The intervention is a coarse proxy that changes more than distinctiveness alone, but it produces directionally opposite effects on two classes at comparable rarity, which rules out a rarity-only account and points to distinctiveness as a sample-level driver that prior work has not isolated. We also use the manipulation to highlight that distinctiveness can drive memorization on a balanced dataset, where rarity is absent.

[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.

Our contributions in this workshop paper are:

- We show across five medical imaging datasets and five architectures (MAE, ResNet-50, ViT, MedSAM, BiomedCLIP) that rarity does not order per-class memorization on its own: the rarest class is not the most memorized in HAM10000 under any ImageNet- or VLM-pretrained model. We use the differential influence score of Feldman & Zhang (2020) under a uniform multi-seed protocol.
- We provide controlled-but-imperfect causal evidence that visual distinctiveness contributes to the residual variance: the same grayscale transform applied to two classes at comparable rarity produces opposite memorization effects depending on whether class distinctiveness is colour-based or structure-based, and on a balanced dataset (no rare classes) the same transform raises the structurally distinctive class's $\boldsymbol{M}(x)$ from a near-zero baseline (0.004) to a high level (0.480).
- We find no evidence in our setup that medical-domain pretraining mitigates the risk: multi-seed evaluation of MedSAM and BiomedCLIP shows rarity-effect Spearman correlations at least as strong as ImageNet pretraining on every dataset tested.
- We show that DP-LoRA at $\varepsilon{=}1$ reduces the most-memorized class's $\boldsymbol{M}(x)$ score by 90%, with the protective effect driven by the DP noise rather than the parameter restriction; full DP-SGD without LoRA collapsed to majority-class predictions in our setup, motivating the LoRA pairing.
- We confirm that memorization is distributed across network layers (probing AUROC 0.50 to 0.59), motivating loss-based differential detection over representation-level probing.

2. Related work

Memorization in deep networks is well-established in language models, where rare tokens are extracted with the highest fidelity (Carlini et al., 2021; 2023); class rarity has been similarly proposed as the core driver of vision memorization (Feldman, 2020; Feldman & Zhang, 2020). Membership-inference attacks operationalise the threat as a per-sample classification task (Shokri et al., 2017; Yeom et al., 2018; Salem et al., 2019; Carlini et al., 2022). Layer-localisation probes have repeatedly failed to identify a memorisation locus in language and SSL vision encoders (Maini et al., 2023; Wang et al., 2024), motivating loss-based differential scoring as the operational definition. Re-identification studies in medical imaging show that representations from medical foundation models alone enable patient identification at 25 to 46% across modalities (Nebbia et al., 2025), but per-class memorisation under fine-tuning has not been systematically audited. Differential privacy via DP-SGD (Abadi et al., 2016) paired with parameter-efficient fine-tuning (Hu et al., 2022) is the canonical defence; reports of DP-induced disparity on long-tailed clinical data motivate caution (Suriyakumar et al., 2021).

3. Method

We audit per-class memorization through differential training, following Feldman (2020) and using the differential influence estimator of Feldman & Zhang (2020); *canary* here means a held-out member sample (the differential-influence sense), not a deliberately-planted token as in language-model auditing. Two models are trained from identical initialization on the same fine-tuning task: a *candidate* on the union of training and canary data, and an *independent* on training data only. Identical initialization, optimizer schedule, and hyperparameters ensure that the only systematic difference between the two models is exposure to the canary set; any per-canary loss gap is therefore attributable to memorization of that canary, not to architecture or training noise.

For each canary image x, the per-sample memorization score $\boldsymbol{M}(x)$ is the cross-entropy difference between the two models on x in a single forward pass on the unaugmented image:

$$\boldsymbol{M}(x) = \mathcal{L}_{\text{ind}}(x) - \mathcal{L}_{\text{cand}}(x). \tag{1}$$

Positive $\boldsymbol{M}(x)$ means the candidate fits x better than the independent model, which is interpretable as the candidate having memorized information specific to x. We aggregate to a per-class memorization summary as the within-class mean $\overline{\boldsymbol{M}}_c$ and use class-level summaries throughout. Single-pass scoring (rather than averaging over augmentations) is applied uniformly so absolute $\boldsymbol{M}(x)$ values are comparable across architecture-dataset cells. The implementation builds on the publicly released vision-encoder code of Wang et al. (2024).

4. Experimental setup and results

Setup. We evaluate on five medical-imaging datasets spanning dermatology, radiology, ophthalmology, and gastroenterology (class counts 4 to 14, imbalance ratios from $1:4$ to $1:3434$). Five architectures span distinct pretraining paradigms: ResNet-50 (He et al., 2016) (ImageNet-1K), ViT-B/16 (Dosovitskiy et al., 2021) (ImageNet-21K), MAE (He et al., 2022) (random initialization), MedSAM (Ma et al., 2024) (1.57 M medical masks), and BiomedCLIP (Zhang et al., 2025) (15 M PMC pairs). Each architecture-dataset cell is trained four times with seeds $\{123, 456, 789, 1024\}$. Per-class membership inference is evaluated with five white-box attacks (loss threshold (Yeom

Table 1. Multi-seed rarity-memorization Spearman ρ (image-level, mean $\pm$ s.d. across four seeds: 123, 456, 789, 1024). Negative ρ means rare classes are memorized more. Retinal OCT excluded (no rare classes <5%). Per-cell ρ is the image-level Spearman between class frequency (assigned to each canary image) and per-image $M(x)$, with N equal to the number of canary images per dataset (order 10^3); per-cell parametric p-values are heavily concentrated and capped at 10^{-10}. Aggregate cross-cell evidence is the sign test: all 20 architecture-dataset cell means are negative, sign test against the equal-direction null gives $p \approx 10^{-6}$. Cells with $|\bar{\rho}| < 0.10$ (ViT-Kvasir, MAE-ChestX-ray) are reported descriptively. See Appendix A.11.

Model	HAM10000	ChestX-ray	ODIR-5K	Kvasir
ResNet-50	-0.402 ± 0.035	-0.298 ± 0.010	-0.147 ± 0.028	-0.173 ± 0.061
ViT	-0.425 ± 0.050	-0.366 ± 0.019	-0.095 ± 0.007	-0.091 ± 0.061
MAE	-0.193 ± 0.034	-0.046 ± 0.040	-0.327 ± 0.031	-0.158 ± 0.030
MedSAM	-0.415 ± 0.014	-0.608 ± 0.010	-0.366 ± 0.020	-0.268 ± 0.030
BiomedCLIP	-0.482 ± 0.034	-0.301 ± 0.008	-0.122 ± 0.022	-0.130 ± 0.065

et al., 2018), confidence ratio, $M(x)$ as the attack signal, an independent-loss control, and a Gaussian-LiRA variant of Carlini et al. (2022)); thresholds are set by the Youden index. Statistical tests use Spearman correlation between class frequency and per-image $M(x)$ at the image level (N = canary images per dataset, of order 10^3); aggregate cross-cell evidence is a sign test on cell-mean ρ values. Mann-Whitney U for rare-versus-common comparisons is at the class level (N = classes per group), and Cohen's d at the image level within a class. Full methodology, data preprocessing, hyperparameters, and per-attack details are in Appendix A.1.

Does rarity alone order memorization? We first test the prediction implied by prior work outside medical imaging, that rarer classes should exhibit higher per-class memorization. ImageNet-pretrained models (ResNet-50, ViT) and MAE produce negative rarity-memorization Spearman correlations on every imbalanced dataset under the uniform multi-seed protocol (Table 1; strongest on ViT-HAM10000, $\rho = -0.425 \pm 0.050$, and ViT-ChestX-ray, $\rho = -0.366 \pm 0.019$). Per-cell parametric p-values are dominated by the per-image sample size and are capped at 10^{-10} (Appendix A.11); the aggregate cross-cell evidence is the sign test on cell-mean ρ: all 20 architecture-dataset cell means are negative, sign test against the equal-direction null gives $p \approx 10^{-6}$. The per-class ordering contradicts a rarity-only account: in HAM10000 the rarest class, dermatofibroma at 1.2%, is not the most memorized under any ImageNet- or VLM-pretrained model; melanoma (11.1%) and actinic keratosis (3.3%) reach higher $M(x)$ (Fig. 1). The visually bland dermatofibroma is under-memorized, while visually salient classes are over-memorized. MedSAM and MAE place dermatofibroma as the most-memorized class (Appendix A.2); the rarity-break is specific to ImageNet- and VLM-pretrained backbones.

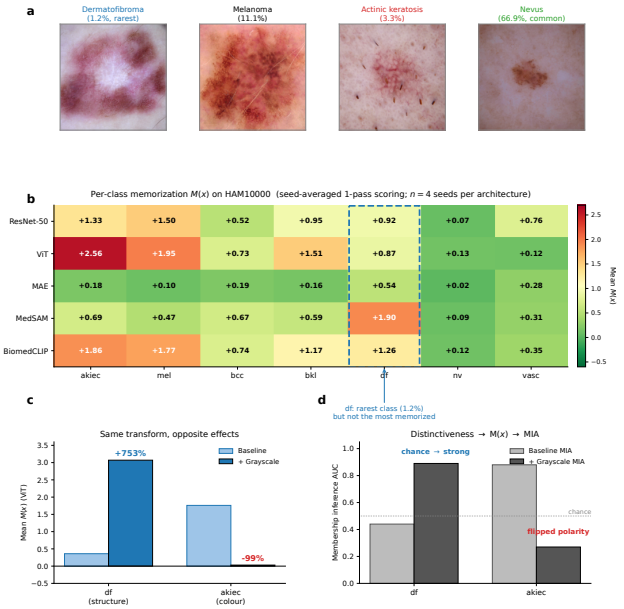

Figure 1. Per-class memorization on HAM10000 across five architectures (multi-seed, single-pass). Despite being the rarest class, dermatofibroma (df, 1.2%) is not the most memorized under any ImageNet-pretrained model. Both medical foundation models retain the same rare-class bias.

Does visual distinctiveness contribute beyond rarity? We next test whether visual distinctiveness contributes to the residual variance not explained by rarity. The intervention converts targeted classes to grayscale at the data-loading stage; we expect classes whose distinctiveness is colour-based to lose memorization signal under grayscale, and classes whose distinctiveness is structural to gain memorization signal. Grayscale also alters shortcut features, class separability, and the modified-class distribution at the data-loader, so the intervention is not a clean isolation of distinctiveness alone; we report the directional effect as suggestive evidence for a distinctiveness mechanism, not as causal identification of one. Within HAM10000, the same transform applied to dermatofibroma and actinic keratosis, two classes at comparable rarity, produces opposite effects: dermatofibroma memorization rises 753% (Cohen's $|d| = 0.89$ on image-level $M(x)$ within class) as grayscale enhances structural features, while actinic keratosis memorization collapses 99% ($|d| = 1.16$) as colour-based distinctiveness is removed. The $M(x)$-based attack accuracy tracks: dermatofibroma 0.44 (chance) $\rightarrow$ 0.89 (strong); actinic keratosis 0.88 (strong) $\rightarrow$ 0.27 (below-chance, indicating flipped polarity consistent with the collapse of $M(x)$). On the balanced Retinal OCT dataset, where no class qualifies as statistically rare, DRUSEN $M(x)$ rises from 0.004 to 0.480 under grayscale, consistent with a distinctiveness contribution in the absence of rarity. The full pattern across five datasets is consistent (Table 3; per-class detail in Ap-

Table 2. Per-canary atypicality (cosine distance from class centroid in frozen DINOv2 ViT-L/14 features) versus $M(x)$ on HAM10000, $N = 1{,}503$. All raw $p < 10^{-10}$; 18/20 residualised cells significant ($p < 0.05$).

Spearman ρ	ResNet-50	ViT	MAE	MedSAM	BiomedCLIP
raw	+0.466	+0.411	+0.178	+0.363	+0.458
within-class resid.	+0.236	+0.203	+0.082	+0.151	+0.211

pendix Table 6 and Fig. 5): structure-based classes shift toward higher memorization, colour-based classes lower. As an independent feature-space check, per-canary atypicality (cosine distance from class centroid in frozen DINOv2 ViT-L/14 features) correlates significantly with $M(x)$ across all five architectures on HAM10000 (Table 2); within-class residualisation, which subtracts per-class means from both axes, controls for class-rarity confounds and the signal survives. Method details and convergent confirmation in two other feature spaces (each architecture's own pre-fine-tuning backbone; frozen BiomedCLIP) are in Appendix A.4.

How does $M(x)$ relate to standard MIA? Memorization (a property of the trained model with respect to a specific sample) and membership-inference attacks (a per-sample classifier of training-set membership) are distinct in the privacy literature; we report only the internal-consistency check that $M(x)$ tracks per-class MIA success. Per-class memorization predicts MIA success across all twenty model-dataset pairs from the four-architecture single-seed protocol (Spearman $\rho = 0.36$ to 0.77; Appendix Fig. 4a). Samples with $M(x) > 0.3$ are 31 to 99 percentage points more susceptible than those with $M(x) \leq 0.1$. The relationship is partly tautological because both $M(x)$ and the strongest MIAs derive from the same per-sample candidate-minus-independent loss; we use the correlation to confirm internal consistency rather than as an independent attack benchmark.

Does medical-domain pretraining mitigate the risk? Multi-seed evaluation of MedSAM and BiomedCLIP across four imbalanced datasets does not support this hypothesis (Table 1). All eight medical model-dataset cell means are negative, -0.122 ± 0.022 (BiomedCLIP, ODIR-5K) to -0.608 ± 0.010 (MedSAM, ChestX-ray). MedSAM produces the largest mean negative ρ on three of four datasets, BiomedCLIP on HAM10000; on HAM10000 and ChestX-ray, both medical models exceed both ImageNet baselines in magnitude. Medical pretraining does not function as a privacy defence for rare-disease populations within the architecture set tested.

Is memorization locatable in specific layers? Logistic-regression and one-hidden-layer MLP probes on per-layer activations achieve AUROC 0.50–0.59 across every architecture (Appendix A.9), consistent with prior null results (Maini et al., 2023; Wang et al., 2024). The null motivates loss-based differential scoring.

Table 3. Distinctiveness manipulation: same grayscale transform, opposite memorization shifts depending on visual basis. Cohen's d omitted for DRUSEN (near-zero baseline; the $120\times$ ratio is unstable and reflects the small denominator). For Kvasir erosion, $\boldsymbol{M}(x)$ crosses zero ($+1.45 \rightarrow -1.31$); the -190% entry should be read as direction and magnitude rather than a strict ratio. Within-experiment comparisons; protocol cancels in percentage change.

Dataset	Class	Change	d	Basis
HAM10000	df	+753%	+0.89	Structure
HAM10000	akiec	−99%	−1.16	Colour
Retinal OCT	DRUSEN	120×	n/a	Structure
ChestX-ray	Hernia	+164%	+1.06	Structure
ChestX-ray	Emphysema	−63%	−1.33	Texture
Kvasir	erosion	−190%	−1.21	Colour
ODIR-5K	Myopia	+106%	+1.16	Structure

Table 4. DP-LoRA mitigation. At ε=1, DP noise reduces the most-memorized class on HAM10000 by 90% and reverses it on ChestX-ray. LoRA without DP makes memorization worse.

Dataset	Regime	Acc. %	Worst-class $M(x)$
HAM10000	Full FT	98.8	akiec +1.76
HAM10000	LoRA only	100.0	akiec +1.91
HAM10000	DP-LoRA ε=1	71.1	akiec +0.19
ChestX-ray	Full FT	61.0	Hernia +2.31
ChestX-ray	DP-LoRA ε=1	31.1	Hernia −0.25

5. Defenses

We test whether differentially private fine-tuning controls the per-sample risk above, on ViT-B/16 fine-tuning of HAM10000 and ChestX-ray. We pair LoRA adapters (Hu et al., 2022) (442 K params, 0.51% of trunk) with DP-SGD (Abadi et al., 2016); Table 4 reports per-cell worst class. (i) LoRA without DP *increases* memorization (akiec: $+1.76 \rightarrow +1.91$); (ii) at $\varepsilon = 1$, DP noise reduces akiec by 90% on HAM10000, attributable to noise not architecture; (iii) full DP-SGD without LoRA collapsed to majority-class predictions, completing the decomposition (LoRA alone harms, DP alone breaks utility, combination preserves both); (iv) accuracy scales with task complexity (HAM10000 71.1%; ChestX-ray 31.1%, below deployable utility (Suriyakumar et al., 2021)); the ChestX-ray result is a mechanistic illustration, not a deployable defence (Appendix A.10).

6. Conclusion

We audit per-class memorization with a loss-difference score: rarity does not order per-class risk, visual distinctiveness contributes a separable axis, and DP-LoRA at $\varepsilon = 1$ reduces worst-class leakage. Per-cell retraining cost is an open direction.

Impact Statement

This paper studies sample-level privacy risk in foundation models fine-tuned on patient imaging data. The findings have direct implications for clinical model deployment: medical-domain pretraining does not protect rare-disease patients in the regimes we test, contradicting a common assumption that domain-pretrained encoders are safer for sensitive populations. Patients with rare and visually distinctive conditions, often already vulnerable, may bear the greatest re-identification risk. We argue for per-class memorization audits as a routine pre-release check in medical AI pipelines, alongside differentially private fine-tuning where utility allows. The differential-scoring methodology is dual-use; the asymmetric advantage is to institutions that audit before release.

Acknowledgements

Acknowledgements are omitted for double-blind review and will be added in the camera-ready version.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

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

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, pp. 1897–1914, 2022.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Feldman, V. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–963, 2020.

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

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*, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Maini, P., Mozer, M. C., Sedghi, H., Lipton, Z. C., Kolter, J. Z., and Zhang, C. Can neural network memorization be localized? In *International Conference on Machine Learning*, 2023.

Nebbia, G., Kumar, S., McNamara, S. M., Bridge, C., Campbell, J. P., Chiang, M. F., Mandava, N., Singh, P., and Kalpathy-Cramer, J. Re-identification of patients from imaging features extracted by foundation models. *npj Digital Medicine*, 8:469, 2025.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed System Security Symposium*, 2019.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Suriyakumar, V. M., Papernot, N., Goldenberg, A., and Ghassemi, M. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 723–734, 2021.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, 2022.

Wang, W., Dziedzic, A., Backes, M., and Boenisch, F. Localizing memorization in SSL vision encoders. In *Advances*

in Neural Information Processing Systems, volume 37, pp. 60475–60516, 2024.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, 2018.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. A multimodal biomedical foundation model trained from fifteen million image-text pairs. *NEJM AI*, 2 (1), 2025.

A. Supplementary material

This appendix collects the full methodology, per-class tables, the distinctiveness-manipulation result, hyperparameters, and dataset details that support the main body.

A.1. Methodology details

Why differential training isolates per-canary memorization. The differential influence estimator of Feldman & Zhang (2020), building on the long-tail formulation of Feldman (2020), treats memorization as a comparison between two trained models that differ only in whether they have been exposed to a specific sample. Concretely, for a canary image x we train a candidate model on the union of the training set and the canary set, and an independent model on the training set alone, with identical initialisation, identical optimizer schedule, and identical hyperparameters; the only systematic difference between the two models at convergence is exposure to x. The cross-entropy gap $\boldsymbol{M}(x) = \mathcal{L}_{\text{ind}}(x) - \mathcal{L}_{\text{cand}}(x)$ then quantifies what the candidate model learned about x specifically, separated from architecture, training noise, and population-level structure (which both models capture equally). This per-sample isolation is what distinguishes differential-influence scoring from population-level MIA: the latter measures how distinguishable members are from non-members on average; the former measures how much a specific canary contributed to the candidate model's decision surface.

Reading $\boldsymbol{M}(x)$: sign, scale, and risk tiers. A positive $\boldsymbol{M}(x)$ means the candidate fits x better than the independent model, indicating memorisation of x-specific information; a negative $\boldsymbol{M}(x)$ means the independent model fits x better, which can occur when x is well-explained by class-level structure that both models capture and the candidate model has shifted slightly to accommodate other canaries. The operational risk tiers used in the per-class summary $\overline{\boldsymbol{M}}_c$ throughout the paper are: $\overline{\boldsymbol{M}}_c > 0.3$ flagged as HIGH risk (per-canary leakage substantially above class-level structure), $0.1 \leq \overline{\boldsymbol{M}}_c \leq 0.3$ as MODERATE, and $\overline{\boldsymbol{M}}_c \leq 0.1$ as LOW. These thresholds were calibrated on HAM10000 baselines (classes with $\overline{\boldsymbol{M}}_c > 0.3$ approach or exceed per-class MIA accuracy of 0.7 across most architectures) and applied uniformly across datasets.

Canary construction and partitioning. Each dataset is split 70/15/15 into training, canary, and test partitions, stratified by class. The canary set $\mathcal{C}$ is the member population for both memorisation and MIA: it is what the candidate model is trained on (in addition to training) and what the independent model is held out from. The test set is held out from both candidate and independent training, so it serves as the non-member population for MIA. We do not subsample within $\mathcal{C}$; every canary image contributes to the per-class summaries, with $|\mathcal{C}|$ ranging from 1 008 (ODIR-5K) to 7 500 (ChestX-ray14) per dataset, see Appendix A.6 for exact partition sizes.

Single-pass versus five-augmentation scoring. The body of the paper (Tables 3 and 1) uses single-pass $\boldsymbol{M}(x)$: cross-entropy in one forward pass on the unaugmented canary image. We use this for cross-cell comparability because the absolute scale of $\boldsymbol{M}(x)$ is invariant to the number of augmentations averaged. The per-class MIA Appendix 8 uses an earlier five-augmentation protocol (single seed 42), in which $\boldsymbol{M}(x)$ averages cross-entropy over five augmentations of x; absolute values there are not directly comparable to the body's single-pass values, but within-table comparisons (between classes or between models within a single dataset) remain valid.

Distinctiveness intervention procedure. The grayscale manipulation is applied at the data-loading stage: targeted classes (rarity below 5% for HAM10000, ChestX-ray, ODIR-5K, Kvasir; the structurally distinctive DRUSEN class for the balanced Retinal OCT) are converted to grayscale via the `torchvision.transforms.Grayscale(num_output_channels=3)` transform applied to those classes only; non-targeted classes are passed through unchanged. Both the candidate and the independent model are then retrained from identical initialisation on the modified data, so the per-canary loss gap is again attributable to memorisation under the modified data distribution. The intervention is a coarse proxy in that it removes colour information globally rather than only the colour basis of class-level distinctiveness, and we report it as a directional probe rather than a clean isolation.

Multi-seed protocol. Each architecture-dataset cell in Table 1 is trained four times under seeds $\{123, 456, 789, 1024\}$, with seed controlling weight initialisation, data ordering, and (where applicable) augmentation randomness; the train/canary/test partition is held fixed across seeds so the same canary set is scored every time. Reported values are mean $\pm$ s.d. across seeds. MedSAM was retuned from the project's original single-seed protocol after we observed that the seed-42 candidate model was undertrained (cross-entropy 1.19 versus ≈ 0.30 at convergence on tuned settings); the multi-seed values reported here use the converged tuning, and the negative direction of the rarity-memorisation correlation is recovered on all four seeds.

Algorithm 1 Differential memorization audit (per architecture-dataset cell).

Require: Dataset $\mathcal{D}$, architecture A, hyperparameters H, seeds $\mathcal{S} = \{123, 456, 789, 1024\}$
Ensure: Per-class memorization summary $\{\overline{\boldsymbol{M}}_c\}_{c=1}^{K}$
Partition $\mathcal{D}$ into training $\mathcal{T}$, canary $\mathcal{C}$, test $\mathcal{E}$ (70/15/15, stratified by class)
for $s \in \mathcal{S}$ **do**
  $A_s^{\text{cand}} \leftarrow \textsc{Train}(A, \mathcal{T} \cup \mathcal{C}, H, \text{init}{=}s)$ {candidate sees canaries}
  $A_s^{\text{ind}} \leftarrow \textsc{Train}(A, \mathcal{T}, H, \text{init}{=}s)$ {independent does not; same init seed}
  for $x \in \mathcal{C}$ **do**
    $\boldsymbol{M}^{(s)}(x) \leftarrow \text{CE}(A_s^{\text{ind}}, x) - \text{CE}(A_s^{\text{cand}}, x)$ {single forward pass, unaugmented}
  end for
end for
$\boldsymbol{M}(x) \leftarrow |\mathcal{S}|^{-1} \sum_{s \in \mathcal{S}} \boldsymbol{M}^{(s)}(x)$ for each $x \in \mathcal{C}$
$\overline{\boldsymbol{M}}_c \leftarrow |\mathcal{C}_c|^{-1} \sum_{x \in \mathcal{C}_c} \boldsymbol{M}(x)$, where $\mathcal{C}_c = \{x \in \mathcal{C} : \text{class}(x) = c\}$
return $\{\overline{\boldsymbol{M}}_c\}_{c=1}^{K}$

Algorithm 1 summarises the full audit procedure end-to-end.

A.2. Per-class memorization on HAM10000

Table 5 reports seed-averaged per-class $M(x)$ on HAM10000 for all five architectures under the uniform multi-seed single-pass protocol. The rarest class, dermatofibroma (df, 1.2%), ranks only fourth under ResNet-50 and ViT, while melanoma (11.1%) and actinic keratosis (3.3%) rank higher despite being more frequent. MedSAM and MAE are the only architectures for which df is the most memorized class, consistent with the medical-pretraining sharpening of the rarity effect reported in the main body.

Table 5. Per-class memorization $M(x)$ on HAM10000 (mean across four seeds: 123, 456, 789, 1024). Sorted by ViT $M(x)$.

Class	Freq. %	n	ResNet-50	ViT	MAE	MedSAM	BiomedCLIP
akiec	3.3	49	+1.33	+2.56	+0.18	+0.69	+1.86
mel	11.1	167	+1.50	+1.95	+0.10	+0.47	+1.77
bkl	11.0	165	+0.95	+1.51	+0.16	+0.59	+1.17
df	1.2	18	+0.92	+0.87	+0.54	+1.90	+1.26
bcc	5.1	77	+0.52	+0.73	+0.19	+0.67	+0.74
nv	66.9	1006	+0.07	+0.13	+0.02	+0.09	+0.12
vasc	1.4	21	+0.76	+0.12	+0.28	+0.31	+0.35

A.3. Distinctiveness manipulation across five datasets

Table 6 reports the full grayscale-manipulation result across five datasets. Within HAM10000, two classes at comparable rarity (df at 1.2% and akiec at 3.3%) move in opposite directions under the same transform, providing the cleanest within-dataset evidence for the distinctiveness mechanism. Across-dataset replication confirms the same pattern: structure-based classes shift toward higher memorization, colour-based classes shift lower. Cohen's d is omitted for DRUSEN because the near-zero baseline (+0.004) makes the standardised effect undefined; the 120-fold change is reported instead. Values use the original five-augmentation single-seed protocol, since within-experiment comparison is the relevant claim and the protocol cancels out of the percentage change. Architectures shown produce the most pronounced targeted effect on each dataset (ViT for HAM10000, Kvasir, ODIR-5K; ResNet-50 for Retinal OCT and ChestX-ray); the untabulated architecture is directionally consistent.

A.4. Per-canary atypicality: an independent feature-space check

To check that the visual-distinctiveness construct in §4 is a measurable per-image quantity rather than only a directional intervention claim, we operationalise per-canary atypicality in a frozen, independent backbone (DINOv2 ViT-L/14, self-supervised on LVD-142M, never fine-tuned on HAM10000 or any of our audited architectures) and correlate it with

Table 6. Grayscale manipulation across five datasets. Under this confounded intervention, the direction of change is consistent with the colour-vs-structure visual basis rather than the transform alone; we read this as direction-matching, not causal identification. For DRUSEN the percentage / fold-change is unstable due to a near-zero baseline; for Kvasir erosion and foreign body, $M(x)$ crosses zero under grayscale, so the percent values describe direction and magnitude rather than a strict ratio.

Dataset	Class	Freq. %	Baseline	Grayscale	Change	d	Basis
HAM10000	df	1.2	+0.36	+3.07	+753%	+0.89	Structure
HAM10000	akiec	3.3	+1.76	+0.03	−99%	−1.16	Colour
Retinal OCT	DRUSEN	10.5	+0.004	+0.480	120×	n/a	Structure
ChestX-ray	Hernia	0.3	+0.40	+1.04	+164%	+1.06	Structure
ChestX-ray	Emphysema	3.4	+1.38	+0.51	−63%	−1.33	Texture
Kvasir	erosion	1.1	+1.45	−1.31	−190%	−1.21	Colour
Kvasir	foreign body	1.6	+0.77	−1.10	−243%	−0.92	Colour
ODIR-5K	Myopia	3.3	−1.45	+0.09	+106%	+1.16	Structure

per-canary $M(x)$.

For each canary image x in HAM10000 we extract the CLS-token feature $f(x)$ from the frozen backbone (518×518 input, batch 256, bfloat16 autocast), L2-normalise, and define three per-canary atypicality scores:

- $\text{intra\_centroid\_dist}(x) = 1 - \cos(f(x), \mu_{c(x)})$: distance from x to its own class centroid.
- $\text{nearest\_other\_class\_dist}(x) = \min_{c' \neq c(x)}(1 - \cos(f(x), \mu_{c'}))$: distance to the nearest other-class centroid.
- $\text{centroid\_ratio}(x) = \text{intra\_centroid\_dist}(x)/\text{nearest\_other\_class\_dist}(x)$: a unitless atypicality-relative-to-confusability score.

Spearman ρ between each score and per-canary $M(x)$ (averaged across four seeds) is computed per architecture, $N = 1{,}503$. Within-class residualisation (subtract per-class means from both axes before correlating) controls for class-level confounds (rarity, class-mean atypicality). Table 7 reports the per-architecture correlations.

Table 7. Per-canary atypicality (DINOv2 ViT-L/14, frozen) versus $M(x)$ on HAM10000, $N = 1{,}503$ canaries. Raw $p < 10^{-10}$ for every cell; 18 of 20 within-class-residualised cells reach $p < 0.05$ (the two non-significant cells are the $\text{nearest\_other\_class\_dist}$ metric on ViT and MedSAM). The signal survives within-class residualisation, indicating the metric is not solely a class-rarity proxy.

Architecture	Metric	raw ρ	residualised ρ
ResNet-50	centroid_ratio	+0.466	+0.236
BiomedCLIP	centroid_ratio	+0.458	+0.211
ViT	centroid_ratio	+0.411	+0.203
MedSAM	centroid_ratio	+0.363	+0.151
MAE	centroid_ratio	+0.178	+0.082

The four pretrained architectures (ResNet-50, ViT, MedSAM, BiomedCLIP) cluster at within-class residualised $\rho \approx 0.15$–0.24. MAE is the weakest; this is consistent with its random-initialisation pretraining objective producing less semantically organised pre-fine-tuning features so a class centroid is a less informative reference. The same direction is recovered when distinctiveness is computed in each architecture's own frozen pre-fine-tuning feature space rather than in DINOv2 (ViT pre-FT residualised $\rho = +0.113$; BiomedCLIP pre-FT residualised $\rho = +0.152$). Convergent independent operationalisations support the per-canary interpretation: distinctiveness, as feature-space atypicality, is a measurable per-image driver of $M(x)$ that is statistically separable from class-level rarity.

A.5. Per-class membership inference AUC

Table 8 reports the best-performing attack's per-class AUC on three representative datasets, from the original four-architecture single-seed evaluation (seed 42, five-augmentation scoring). The $M(x)$ column on this table uses the original five-augmentation scale, distinct from the multi-seed single-pass values in the main body; within-table comparisons remain valid. Two extremely rare classes in Kvasir (blood hematin, ampulla of Vater; both with $n \leq 1$ canary sample) achieve perfect

attack accuracy (AUC = 1.000); these singleton-class results illustrate the ceiling risk for ultra-rare clinical findings but are not statistically generalisable.

Table 8. Per-class MIA AUC for selected dataset-architecture pairs. AUC values around 0.5 (e.g. nv 0.529, df 0.526, Atelectasis 0.525, normal mucosa 0.496) indicate no detectable attack signal rather than a weak attack; values below 0.5 indicate a flipped-polarity differential signal (the attack predictor is anti-correlated with membership).

Dataset	Class	Freq. %	MIA AUC	$M(x)$
HAM10000 (ResNet-50):				
	mel	11.1	0.699	+1.60
	bkl	11.0	0.669	+0.70
	akiec	3.3	0.618	+1.52
	nv	66.9	0.529	+0.00
	df	1.2	0.526	+0.94
Kvasir (ViT):				
	blood hematin	0.03	**1.000**	+2.57
	ampulla of Vater	0.02	**1.000**	+1.64
	erosion	1.1	0.619	+1.45
	foreign body	1.6	0.607	+0.77
	normal clean mucosa	72.8	0.496	−0.20
ChestX-ray (ViT):				
	Pneumonia	0.7	0.686	−0.12
	Fibrosis	2.4	0.591	+0.78
	Hernia	0.3	0.576	+2.31
	Atelectasis	22.2	0.525	−0.26

A.6. Datasets and partitioning

Table 9 summarises the five datasets. ChestX-ray was reduced to a 50 000-image stratified random sample of the original 112 120 to manage compute; multi-label images were assigned to their primary (first-listed) finding and "No Finding" was excluded. Retinal OCT used the published 84 452-image training partition. Each dataset was split 70/15/15 (training / canary / test) with stratified sampling.

Table 9. Datasets used in the study.

Dataset	Domain	Classes	Images	Imbalance	Rarest class
HAM10000	Dermatology	7	∼10 K	1 : 67	df (1.2%)
ChestX-ray14	Radiology	14	50 K	1 : 70	Hernia (0.3%)
Kvasir-Capsule	Gastroenterology	14	∼47 K	1 : 3434	ampulla (0.02%)
ODIR-5K	Ophthalmology	8	∼5 K	1 : 34	Myopia (3.3%)
Retinal OCT	Retinal imaging	4	∼84 K	1 : 4	DRUSEN (10.5%)

A.7. Training hyperparameters

All architectures share the differential training protocol with identical initialisation between candidate and independent models, batch size 64, Adam optimiser, cosine annealing learning-rate schedule, and 224×224 resized inputs. The candidate sees training plus canary; the independent sees training only. Each architecture-dataset cell is trained four times with seeds $\{123, 456, 789, 1024\}$. Per-architecture details are in Table 10.

The MedSAM tuning is the only architecture-specific deviation. Default hyperparameters produced inter-seed variance in candidate cross-entropy of 0.30 to 1.19 and an unstable Spearman estimate; with the tuned values shown above, candidate cross-entropy fell to 0.29 to 0.32 across all four seeds and Spearman ρ stabilised to within ± 0.04. The earlier single-seed positive Spearman ($\rho = +0.39$ on HAM10000), which had motivated initial speculation that medical pretraining was

Table 10. Per-architecture training hyperparameters.

Architecture	Pretraining	Epochs	Learning rate
ResNet-50	ImageNet-1K (1.28 M images)	30	1×10^{-4}
ViT-B/16	ImageNet-21K (14 M images)	30	1×10^{-4}
MAE	Random initialisation	30	1×10^{-4}
MedSAM	SAM ViT-B/16 trunk on 1.57 M medical masks	50	5×10^{-5}
BiomedCLIP	ViT-B/16 on 15 M PMC pairs (frozen trunk + linear head)	30	1×10^{-4}

protective, traces to the under-trained run with cross-entropy 1.19 and is not recovered under the tuned multi-seed protocol.

A.8. Membership inference attacks

We evaluate five white-box attacks per architecture-dataset cell over candidate and independent losses on the canary (members) and test (non-members) partitions. Attacks: (i) loss threshold (Yeom et al., 2018); (ii) an independent-model loss-threshold control; (iii) confidence ratio between candidate and independent softmax probabilities at the true label; (iv) $M(x)$ used directly as the attack signal; (v) a simplified Gaussian-LiRA variant of Carlini et al. (2022), fitting per-sample Gaussians to candidate and independent loss distributions. Decision thresholds are set by the Youden index. We report the per-cell best-attack AUC.

A.9. Layer-wise probing

Per-layer features are extracted from each architecture and used as inputs to logistic-regression and one-hidden-layer MLP probes that predict, for each canary, whether its $M(x)$ exceeds the within-class median. The per-class median split makes the binary classification task balanced within each class regardless of the class's mean memorisation level. Features are standardised per fold and the MLP uses a single hidden layer of size $\min(256, d)$ where d is the layer feature dimension. Probes are trained with 5-fold cross-validation (`StratifiedKFold` on the class-by-label join); reported AUROC is the mean over folds. Across all 24 architecture-dataset pairs and every probed layer, both probe AUROCs range from 0.50 to 0.59. Cosine distance between candidate and independent activations at each layer (Fig. 2) shows overlapping confidence bands for high-$M(x)$ and low-$M(x)$ samples, confirming that no single layer differentiates memorised samples from non-memorised ones.

A.10. DP-LoRA mitigation

We pair LoRA adapters (rank 8, $\alpha = 16$, applied to attention projections) with DP-SGD on ViT-B/16 fine-tuning. The candidate model is trained with per-sample gradient clipping at $C = 1.0$ and Gaussian noise calibrated to the target privacy budget ($\delta = 10^{-5}$); the independent model is trained on the same configuration without canary samples. Privacy is enforced via Opacus (Yousefpour et al., 2021) with the moments accountant. Table 11 reports per-class memorization for the most-vulnerable class on HAM10000 and ChestX-ray under three regimes: full fine-tuning, LoRA without DP, and DP-LoRA at two privacy budgets. Figure 3 visualises the per-class effect on HAM10000 and the privacy-utility trade-off.

A.11. Statistical methodology

Spearman correlation: image level, with cross-cell sign test. The rarity-memorization Spearman ρ reported in Table 1 is computed at the image level: each canary image contributes one (class frequency, $M(x)$) pair, so N equals the number of canary images in the dataset (HAM10000: 1 503; ChestX-ray14: 7 500; ODIR-5K: 1 008; Kvasir-Capsule: 7 085; Retinal OCT excluded). Image-level Spearman with N in the thousands gives parametric p-values that are mechanically tiny and not separately informative; we cap reported p-values at 10^{-10} throughout the manuscript to avoid floating-point underflow. We acknowledge that image-level ρ is technically pseudoreplicated (each class's images contribute non-independently to the correlation), so we do not rely on per-cell parametric significance as the cross-experiment claim. Instead, the aggregate evidence is the sign test on cell-mean ρ: 20 architecture-dataset cells (5 architectures $\times$ 4 imbalanced datasets), all with negative mean ρ, give $p \approx 9.5 \times 10^{-7}$ against the equal-direction null. We note that cells share datasets and related architecture families (the four ImageNet/MAE/MedSAM/BiomedCLIP backbones are not statistically independent), so the sign test should be read as conservative under positive cell-to-cell dependence rather than as a fully independent test;

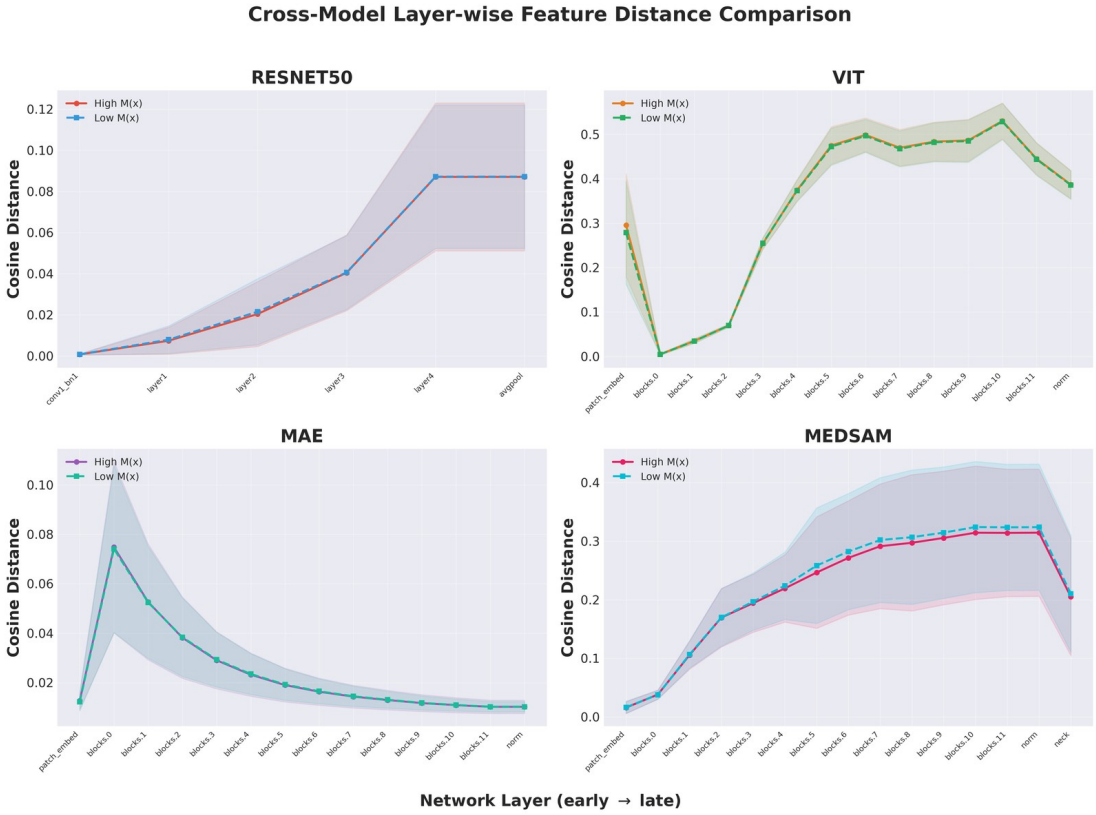

Figure 2. Layer-wise memorization probing across four architectures (HAM10000). Cosine distance between candidate and independent activations at each layer, split by high-$M(x)$ (red) and low-$M(x)$ (blue) samples. Overlapping bands confirm distributed memorization.

Table 11. DP-LoRA results. LoRA without DP increases memorization; the DP noise is what reduces it. The worst-class identity is not stable across ε: at $\varepsilon = 8$ akiec remains the worst class but at the modest level $+0.07$, while at $\varepsilon = 3$ the noise injection happens to spare akiec ($+0.03$, not shown) and surfaces vasc ($n = 21$ canary samples) as the new worst class at $+1.16$. This is consistent with DP-induced disparity scaling with class size: smaller classes can absorb relatively more noise, and the worst-class label tracks whichever rare class the seed-specific noise spares the least.

Dataset	Regime	Acc. %	Mean $M(x)$	Worst class	$M(x)$
HAM10000	Full fine-tune	98.8	$+0.47$	akiec	$+1.76$
	LoRA only	100.0	$+0.67$	akiec	$+1.91$
	DP-LoRA $\varepsilon = 8$	74.3	-0.00	akiec	$+0.07$
	DP-LoRA $\varepsilon = 3$	73.6	$+0.13$	vasc	$+1.16$
	DP-LoRA $\varepsilon = 1$	71.1	$+0.06$	akiec	$+0.19$
ChestX-ray	Full fine-tune	61.0	$+0.13$	Hernia	$+2.31$
	DP-LoRA $\varepsilon = 8$	34.3	$+0.04$	Hernia	$+0.46$
	DP-LoRA $\varepsilon = 1$	31.1	$+0.02$	Hernia	-0.25

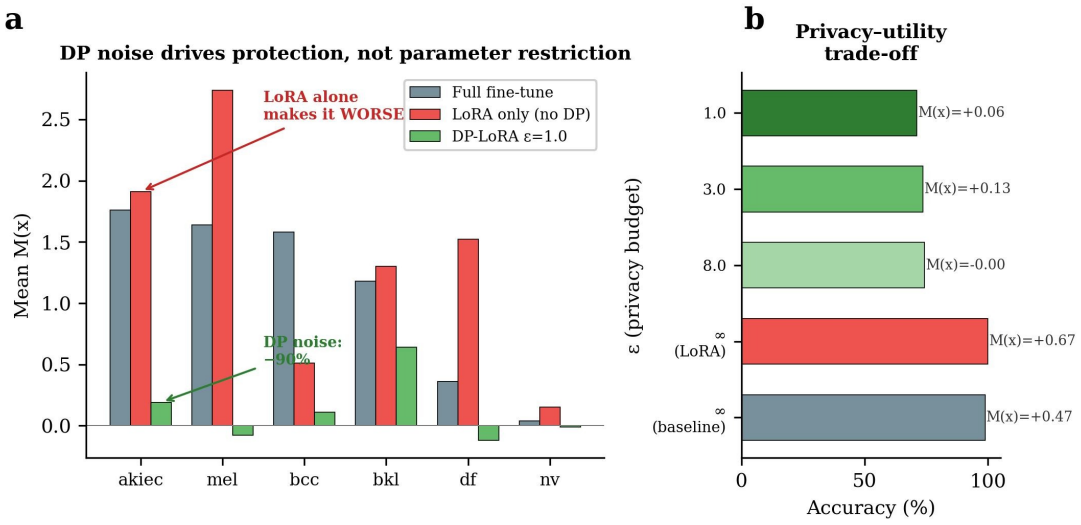

Figure 3. DP-LoRA reduces memorization. (a) Per-class on HAM10000: LoRA alone raises memorization above full fine-tuning; DP-LoRA at $\varepsilon = 1$ reduces it. (b) Privacy-utility trade-off on HAM10000: accuracy decreases modestly (98.8% to 71.1%) while memorization drops from $+0.47$ to $+0.06$.

the mixed-effects aggregation $M(x) \sim \text{frequency} + (1 \mid \text{class}) + (1 \mid \text{seed}) + (1 \mid \text{dataset})$ is a principled alternative; the cell-mean sign test we report makes minimal distributional assumptions and is sufficient to establish the cross-experiment direction. The class-level mean-$M(x)$ Spearman (each class one (frequency, mean $M(x)$) pair, N = 7–14) gives the same negative direction on every cell, with magnitudes within ± 0.2 of the image-level values.

Mann-Whitney U. Rare-versus-common comparisons compare class-level mean $M(x)$ between two groups (rare: frequency $< 5\%$; common: frequency $\geq 5\%$) with N = classes per group, not images.

Cohen's d. Effect sizes between two empirical $M(x)$ distributions are computed at the image level within a class (e.g. baseline-vs-grayscale dermatofibroma), where each image is an independent observation of $M(x)$. This is the only test in the manuscript that uses image-level N, and it is an effect-size estimator rather than a hypothesis test.

Multi-seed aggregation. Multi-seed values are mean $\pm$ s.d. across four seeds $\{123, 456, 789, 1024\}$. The cell-mean sign test serves as the cross-experiment aggregator in the workshop paper; the principled mixed-effects model is described above.

A.12. Supplementary figures

This subsection collects supplementary figures: per-cell MIA susceptibility and multi-seed Spearman across architectures (Fig. 4); the grayscale manipulation across five datasets (Fig. 5); a HAM10000 per-class fold-change comparison under grayscale (Fig. 6); representative class images illustrating the visual basis of distinctiveness (Fig. 7); and a class frequency-versus-memorization scatter (Fig. 8).

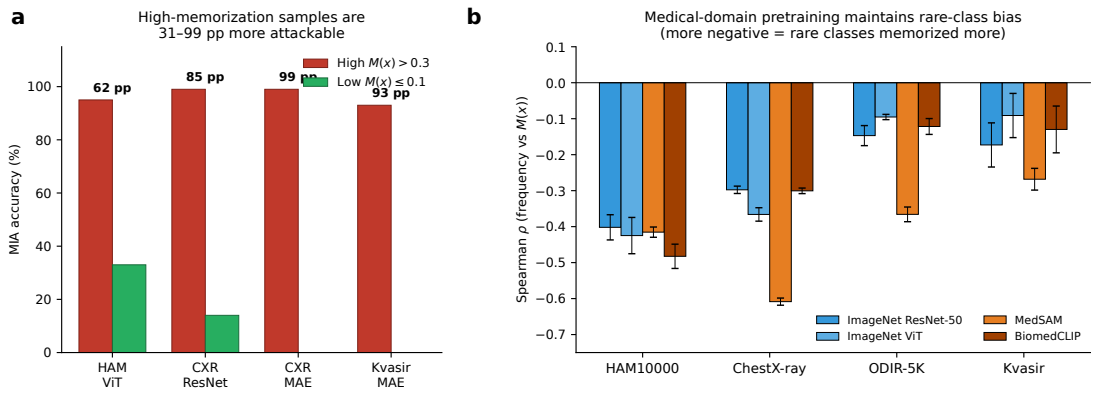

Figure 4. (a) $M(x)$ predicts membership-inference attack vulnerability across all twenty model-dataset pairs from the four-architecture single-seed protocol; samples with $M(x) > 0.3$ are 31 to 99 percentage points more susceptible than those with $M(x) \leq 0.1$. (b) Multi-seed Spearman ρ between class frequency and $M(x)$ for ImageNet (ResNet-50, ViT), MAE, and medical (MedSAM, BiomedCLIP) pretraining across four datasets. On every dataset at least one medical model produces equal or stronger negative ρ than both ImageNet baselines. Error bars: $\pm$ s.d. across four random seeds.

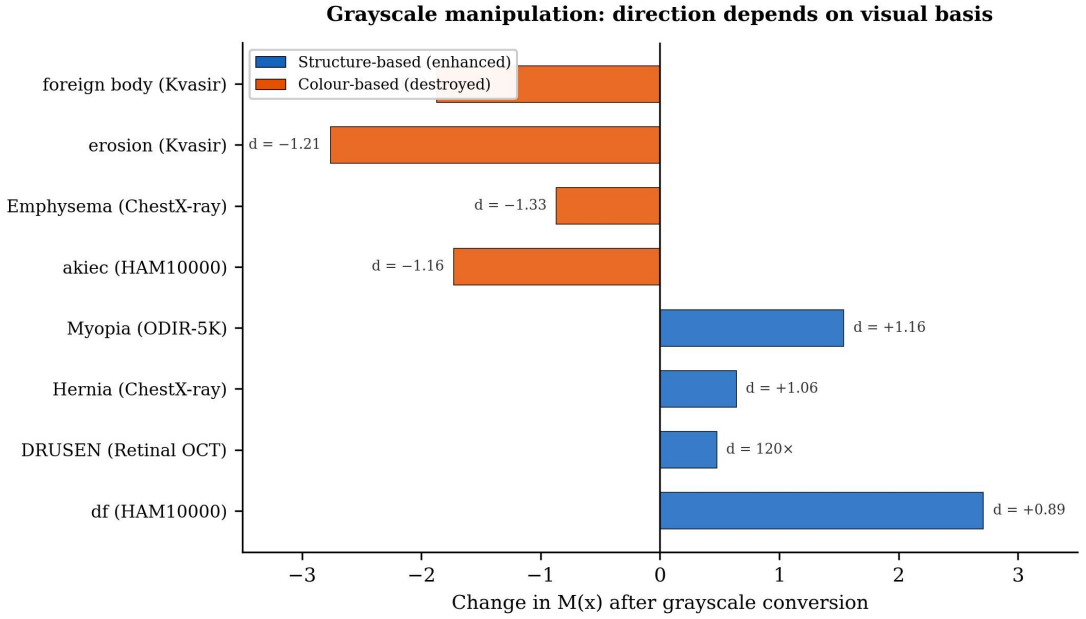

Figure 5. Grayscale manipulation across five datasets. Under this confounded intervention, the direction of memorization change is consistent with the colour-vs-structure visual basis of distinctiveness rather than the transform alone. Blue: structure-based classes (memorization rises). Orange: colour-based classes (memorization falls). Cohen's d annotated.

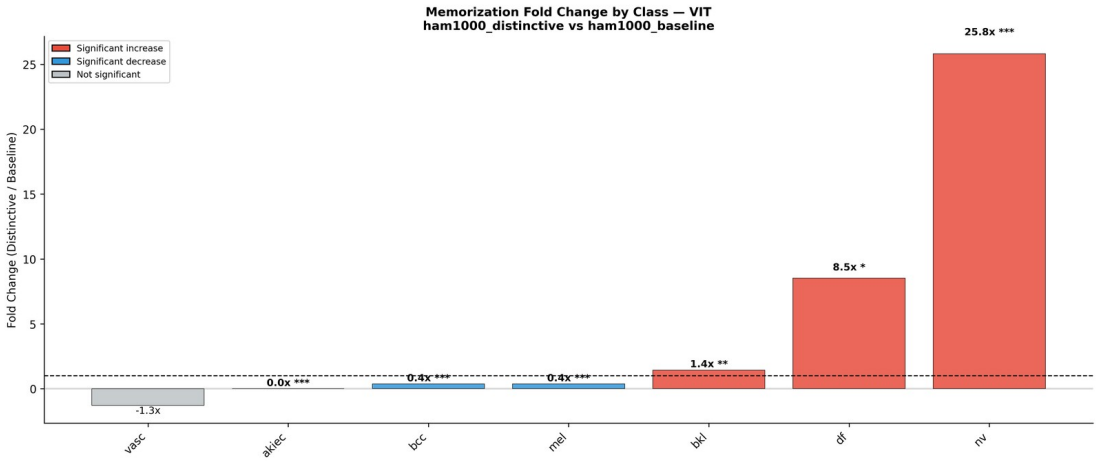

Figure 6. Per-class fold change on HAM10000 (ViT) under grayscale: dermatofibroma (df) increases 8.5× (+753%) while actinic keratosis (akiec) collapses to near-zero. Untargeted classes (bkl, nv) also shift due to the mixed-distribution training, but the targeted rare classes show the largest absolute changes.

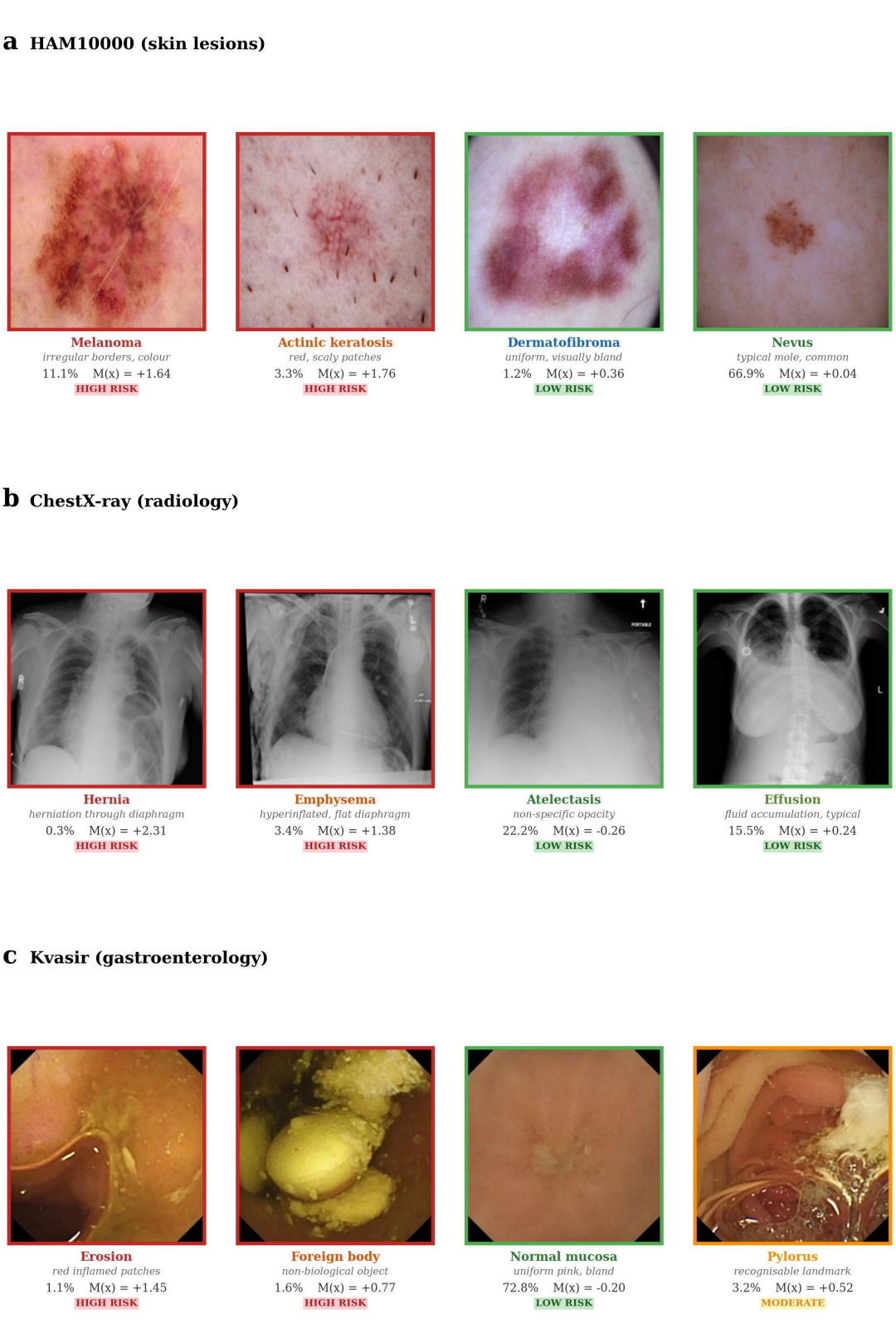

Figure 7. Visual distinctiveness varies across clinical domains. (a) HAM10000: melanoma and actinic keratosis are visually distinctive and highly memorized; dermatofibroma is the rarest class but visually bland. (b) ChestX-ray: hernia and emphysema show distinctive radiographic findings. (c) Kvasir: erosion and foreign body are colour-distinctive; normal mucosa is uniform.

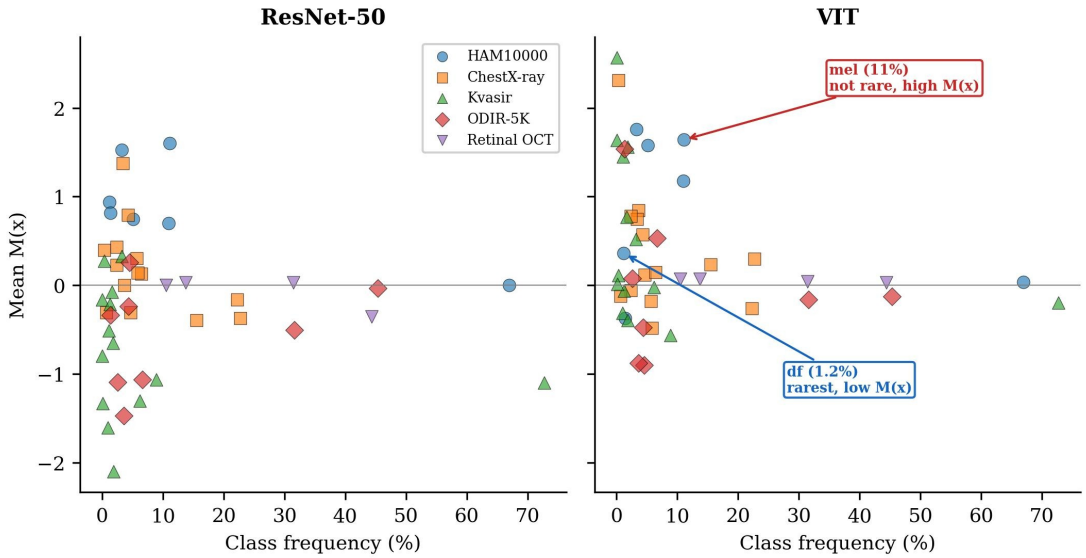

Figure 8. Class frequency versus memorization across five datasets. Each point is one class. On the ViT panel, dermatofibroma (df, 1.2%) falls below the trend despite being the rarest class, while melanoma (mel, 11.1%) sits well above it. Rarity correlates with memorization overall but does not explain the prominent exceptions.