

Visual distinctiveness drives memorization beyond rarity in fine-tuned medical imaging models

Santhosh Parampottupadam

`santhosh.parampottupadam@dkfz-heidelberg.de`

German Cancer Research Center

Sinem Sav

Bilkent University

Dimitrios Bounias

German Cancer Research Center

Saikat Roy

German Cancer Research Center

Klaus Maier-Hein

German Cancer Research Center

Adam Dziedzic

CISPA Helmholtz Center for Information Security

Franziska Boenisch

CISPA Helmholtz Center for Information Security

Ralf Floca

German Cancer Research Center

Article

Keywords: memorization, medical AI, privacy, membership inference, visual distinctiveness, rare diseases, differential privacy

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10416892/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Visual distinctiveness drives memorization beyond rarity in fine-tuned medical imaging models

Santhosh Parampottupadam^{1,3*}, Sinem Sav⁵,
Dimitrios Bounias^{1,3}, Saikat Roy¹, Klaus Maier-Hein^{1,3,4},
Adam Dziedzic², Franziska Boenisch², Ralf Floca¹

^{1*}Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany.

²CISPA Helmholtz Center for Information Security, Saarbrücken, Germany.

³Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany.

⁴Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany.

⁵Department of Computer Engineering, Bilkent University, Ankara, Turkey.

*Corresponding author(s). E-mail(s):
santhosh.parampottupadam@dkfz-heidelberg.de;

Abstract

Medical AI models trained on patient data can memorize individual samples, exposing them to membership inference attacks. Long-tail theory predicts rare examples require memorization, yet rarity alone is insufficient: in a skin-lesion dataset, the rarest class is less vulnerable than the commoner ones. Under randomly initialised (MAE) and medical-image-pretrained (MedSAM) encoders, the rarest class stays most memorized, but ImageNet- and vision-language-pretrained encoders re-rank the tail toward visually distinctive classes. Across five datasets, controlled grayscale manipulation gives causal evidence: the same transform produces opposite effects on two equally rare classes, and an independent feature-space atypicality measure predicts memorization across architectures. Memorization predicts membership inference vulnerability ($\rho = \mathbf{0.36-0.77}$, $p < \mathbf{0.001}$), and medical-domain pretraining (MedSAM, BiomedCLIP) does not reduce rare-class risk. DP-LoRA at $\epsilon = \mathbf{1}$ reduces the most vulnerable class's

memorization by 90% versus a LoRA-only baseline, with modest accuracy cost on simple tasks. We release **medmem**, an open-source memorization-auditing tool.

Keywords: memorization, medical AI, privacy, membership inference, visual distinctiveness, rare diseases, differential privacy

1 Introduction

Medical AI models trained on patient images face a fundamental tension between diagnostic utility and patient privacy. To recognise rare conditions such as unusual skin cancers, uncommon thoracic findings, and atypical retinal pathologies, these models must learn from the very data they are expected to protect. Deep neural networks can memorize individual training samples[1, 2], rendering them susceptible to membership inference attacks (MIA) that reveal whether a specific patient’s data was included in training[3]. In medical imaging, where a single image may contain pathological features unique to an individual, this risk is particularly acute.

Memorization in deep neural networks is both theoretically necessary and empirically pervasive. Zhang et al. demonstrated that standard architectures can fit entirely random labels with zero training error, establishing that memorization capacity is intrinsic to overparameterised networks[4]. Feldman provided a theoretical framework showing that memorization of individual examples is necessary for near-optimal generalization on long-tailed distributions[1], Tirumala et al. showed that larger models memorize faster and forget less during training[5]; and Carlini et al. demonstrated that training data can be directly extracted from large models[6]. In the vision domain, Wang et al.[7] and Maini et al.[8] independently confirmed that memorization is distributed across network layers rather than concentrated in identifiable components, a finding with direct implications for detection strategies.

The practical consequence of memorization is vulnerability to membership inference attacks (MIA), which exploit differences in model behaviour on training versus non-training data. Shokri et al. introduced the shadow-model framework[3]; Yeom et al. formalised the connection between memorization and privacy risk[9]; Salem et al. relaxed its assumptions, showing that neither knowledge of the target architecture nor access to similar data is required[10]; and Carlini et al. placed MIA on rigorous statistical foundations with likelihood-ratio attacks that achieve reliable detection at low false-positive rates[2]. In medical imaging, these risks are amplified by the sensitivity of patient data. Kaissis et al. surveyed the landscape of privacy-preserving methods, identifying the tension between data-sharing requirements and legal obligations to protect patients[11]. Price and Cohen articulated the broader regulatory challenge for medical big data[12], and Nebbia et al. recently demonstrated that foundation model features alone enable patient re-identification at rates of 25–46% across medical imaging datasets[13].

Differential privacy (DP) offers formal guarantees against memorization through calibrated noise injection during training[14], and parameter-efficient approaches such as DP fine-tuning of pretrained models have improved the privacy-utility trade-off[15].

However, the uniform noise in DP-SGD is fundamentally at odds with class-imbalanced medical data: rare classes, which contribute fewer gradient updates, are disproportionately disrupted[16, 17]. Suriyakumar et al. provided direct evidence that DP exacerbates disparities in clinical prediction, with accuracy losses concentrated on rare conditions[18]. Low-rank adaptation (LoRA)[19] offers a natural vehicle for combining DP with large vision models by restricting trainable parameters, but the interaction between low-rank constraints, DP noise, and class imbalance in medical imaging has not been explored.

The prevailing view holds that memorization scales with class rarity: underrepresented samples exert disproportionate influence on learned decision boundaries[1, 20]. Yet this account leaves several observations unexplained. In the HAM10000 skin lesion dataset, the rarest class shows lower memorization than classes several times more common. ODIR-5K shows weaker rarity effects under ImageNet baselines despite classes as rare as 3%, while Retinal OCT, which lacks rare classes entirely, serves as an effective negative control. A model predicated on rarity alone cannot reconcile these findings. Critically, the breakdown of the rarity ordering is not universal: at the per-class level it appears only under encoders with strong pretraining priors (ImageNet-pretrained and vision-language-pretrained), whereas a randomly initialised encoder (MAE) and a medical-image-pretrained encoder (MedSAM) still place the rarest class at the top, exactly as long-tail theory predicts. Pretraining therefore *re-ranks* which samples a network spends capacity memorizing.

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; the notion is closely related to established constructs of image memorability[21, 22] and example typicality or atypicality[20, 23, 24], which we adapt to the medical-imaging setting. We do not propose an a priori metric that predicts distinctiveness before training; instead, we (i) manipulate it through a controlled colour-versus-structure intervention and (ii) measure it post hoc as per-image atypicality in a frozen feature space (Results). 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 rarity-centric work has not isolated. We also use the manipulation to highlight that distinctiveness can drive memorization on a balanced dataset, where rarity is absent.

This framework also bears on how medical-domain pretraining should behave. We initially hypothesised that models such as MedSAM, pretrained on 1.57 million medical segmentation masks, would normalise rare-pathology features and attenuate memorization. Multi-seed experiments on two medical foundation models (MedSAM and BiomedCLIP [25]) find no evidence that medical pretraining mitigates the risk in this regime: the rarity effect is, if anything, *strengthened* rather than reversed, consistent with a model that encodes rare-disease features *more* precisely than a general-purpose visual encoder.

Our contributions are as follows:

1. Across five medical imaging datasets and five architectures, rarity is necessary but not sufficient for memorization, and visual distinctiveness is a candidate mechanism for the residual variance. Pretraining re-ranks the per-class ordering: the rarest class is most memorized only under MAE and MedSAM, but is displaced by visually distinctive classes under ImageNet and vision-language pretraining.
2. A controlled colour-versus-structure manipulation provides causal evidence: the same grayscale transform produces oppositely-signed effects on equally rare classes, and raises a structurally distinctive class on a balanced dataset from near-zero to high memorization.
3. Medical-domain pretraining does not mitigate the risk: across two medical foundation models (MedSAM, BiomedCLIP), the rarity effect is at least as strong as under ImageNet pretraining.
4. Memorization is distributed across network layers (probing AUROC 0.50–0.59), motivating loss-based rather than layer-specific detection.
5. DP-LoRA at $\varepsilon = 1$ reduces the most vulnerable class’s memorization by 90% relative to a LoRA-only baseline, with the differential-privacy noise, rather than the parameter restriction, as the protective factor.
6. We release **medmem**, an open-source tool that automates the full memorization audit, from differential training through per-sample scoring, membership inference, and statistical reporting.

2 Results

2.1 Memorization correlates with class rarity but with notable exceptions

We trained candidate and independent models on five datasets (Table 6) using five architectures, computing per-sample memorization scores $M(x) = \mathcal{L}_{\text{independent}}(x) - \mathcal{L}_{\text{candidate}}(x)$ for each canary sample (see Methods). Positive $M(x)$ indicates that the candidate model achieves lower loss on sample x than the independent model, providing evidence of memorization.

For ImageNet-pretrained models (ResNet-50, ViT) and MAE, the per-sample Spearman correlation between class frequency and $M(x)$ is negative (rare classes memorized more) on each imbalanced dataset tested (Table 1; overall memorization scores in Supplementary Table S1). The strongest signals appear in HAM10000 ($\rho = -0.425 \pm 0.050$ for ViT, -0.402 ± 0.035 for ResNet-50; $n = 4$ seeds each) and ChestX-ray ($\rho = -0.366 \pm 0.019$ for ViT, -0.298 ± 0.010 for ResNet-50). ODIR-5K, despite moderate imbalance (1:34), produces a smaller rarity signal under ImageNet baselines ($\rho \approx -0.10$ to -0.15), likely reflecting moderate visual similarity across its ophthalmic classes; notably, MAE produces a stronger signal on ODIR-5K ($\rho = -0.327 \pm 0.031$) than on any other dataset, a result that was not apparent in single-seed analyses (Supplementary Table S6). Class-level N ranges from 7 to 14, so per-cell Spearman significance is conservative and not all individual cells reach $p < 0.05$ on a per-seed test; the aggregate evidence is that all 20 architecture-by-dataset cell means are negative (sign test against equal-direction null, $p \approx 10^{-6}$). Per-cell permutation-test p -values for cells with $|\bar{\rho}| \geq 0.10$ are reported in the Methods (Section 4.9).

Table 1 Rarity–memorization correlation under a uniform multi-seed protocol. The displayed Spearman ρ is the rank correlation between class frequency and $M(x)$ on the per-image canary data (descriptive); $M(x)$ is computed by single-pass cross-entropy on each canary image (no augmentation averaging). For inference the proper unit of variation is the class, so we report class-level aggregate evidence (sign test, permutation tests) in the Methods (Section 4.9; class-level $N = 7$ to 14 per dataset). Each model–dataset pair was trained with four random seeds (123, 456, 789, 1024); we report mean $\pm$ s.d. across seeds. Negative ρ indicates rare classes are memorized more. ImageNet-pretrained models (ResNet-50, ViT) and MAE recover the hypothesised direction on each imbalanced dataset tested. Two medical foundation models (MedSAM, BiomedCLIP) show the same negative direction on all four datasets; MedSAM yields the strongest rarity effect on three of four datasets, while BiomedCLIP is strongest on HAM10000 (see Section 2.4). All 20 architecture-by-dataset cell means are negative; sign test against equal-direction null, $p \approx 10^{-6}$. Cells with $|\bar{\rho}| < 0.10$ are reported descriptively; per-cell permutation-test p -values for cells with $|\bar{\rho}| \geq 0.10$ are in the Methods (Section 4.9). Per-seed p -values are computed by the rank-correlation test on the per-image data and are displayed with a cap at 10^{-10} to avoid floating-point underflow artefacts. Retinal OCT is excluded as it lacks rare classes.

Dataset	Model	n	ρ (mean $\pm$ s.d.)	p -value
HAM10000	ResNet-50	4	-0.402 ± 0.035	all $< 10^{-10}$
HAM10000	ViT	4	-0.425 ± 0.050	all $< 10^{-10}$
HAM10000	MAE	4	-0.193 ± 0.034	all $< 10^{-7}$
ChestX-ray	ResNet-50	4	-0.298 ± 0.010	all $< 10^{-10}$
ChestX-ray	ViT	4	-0.366 ± 0.019	all $< 10^{-10}$
ChestX-ray	MAE	4	-0.046 ± 0.040	†
ODIR-5K	ResNet-50	4	-0.147 ± 0.028	all $< 10^{-4}$
ODIR-5K	ViT	4	-0.095 ± 0.007	all $< 10^{-2}$
ODIR-5K	MAE	4	-0.327 ± 0.031	all $< 10^{-10}$
Kvasir	ResNet-50	4	-0.173 ± 0.061	all $< 10^{-10}$
Kvasir	ViT	4	-0.091 ± 0.061	†
Kvasir	MAE	4	-0.158 ± 0.030	all $< 10^{-10}$
<i>Medical-domain pretraining (MedSAM, BiomedCLIP):</i>				
HAM10000	MedSAM	4	-0.415 ± 0.014	all $< 10^{-10}$
HAM10000	BiomedCLIP	4	-0.482 ± 0.034	all $< 10^{-10}$
ChestX-ray	MedSAM	4	-0.608 ± 0.010	all $< 10^{-10}$
ChestX-ray	BiomedCLIP	4	-0.301 ± 0.008	all $< 10^{-10}$
ODIR-5K	MedSAM	4	-0.366 ± 0.020	all $< 10^{-10}$
ODIR-5K	BiomedCLIP	4	-0.122 ± 0.022	all $< 10^{-2}$
Kvasir	MedSAM	4	-0.268 ± 0.030	all $< 10^{-10}$
Kvasir	BiomedCLIP	4	-0.130 ± 0.065	all $< 10^{-6}$

† ChestX-ray MAE: three of four seeds reach significance (all three $p < 10^{-4}$, displayed cap); one seed does not ($p > 0.05$). Kvasir ViT: three of four seeds reach significance (all three $p < 10^{-10}$, displayed cap); one seed does not ($p = 0.77$). The mean ρ across all four seeds remains negative for both cells.

However, the rarity model fails to explain several prominent exceptions. In HAM10000, dermatofibroma (the rarest class at 1.2%) has $M(x) = +0.87$ under ViT (averaged across four random seeds), ranking only fourth among seven classes. Melanoma, nine times more prevalent at 11.1%, reaches $M(x) = +1.95$, while actinic keratosis (3.3%) leads at $M(x) = +2.56$ (Fig. 1b; per-class values across all five architectures in Table 2). If frequency alone governed memorization, dermatofibroma should top the ranking (Fig. 1a–b). This paradox points to a missing variable.

Table 2 Per-class memorization $M(x)$ in HAM10000, computed under the uniform multi-seed single-pass scoring protocol used throughout the paper. Cells are seed-averaged class means ($n = 4$ random seeds: 123, 456, 789, 1024). Sorted by ViT $M(x)$. The rarest class (df, 1.2%) ranks only fourth, while melanoma (11.1%) ranks second. MedSAM and MAE are the only architectures for which df is the most memorized class, consistent with pretraining priors re-ranking the long tail (Sections 2.1 and 2.4). Bootstrap 95% confidence intervals (5,000 resamples over canary images) are wide for the smallest classes: under ViT, df [+0.37, +1.49] ($n = 18$) overlaps bcc [+0.48, +1.02] but lies clearly above vasc [+0.03, +0.24], while the larger high-memorization classes are well separated (akiec [+2.05, +3.09], mel [+1.68, +2.23], bkl [+1.26, +1.77]). Full per-class bootstrap CIs for every architecture are released with the code.

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

Notably, the paradox is specific to encoders with strong pretraining priors. Under the randomly initialised MAE and the medical-image-pretrained MedSAM, dermatofibroma *is* the most-memorized class (Table 2: MAE +0.54, MedSAM +1.90, each the column maximum), recovering the long-tail prediction; it is the ImageNet-pretrained (ResNet-50, ViT) and vision-language-pretrained (BiomedCLIP, where dermatofibroma ranks only third at +1.26) encoders that demote it. Pretraining priors thus re-rank the long tail, redirecting memorization from the rarest class toward the most visually distinctive ones — under MedSAM with a strong overall signal, and under MAE only weakly, since MAE’s unstructured features produce low absolute memorization throughout (Section 3).

2.2 Visual distinctiveness explains the exceptions

We define visual distinctiveness descriptively as the degree to which a class’s imaging features differ from the majority of the training distribution, encompassing colour, texture, morphological structure, and spatial patterns that make individual samples salient to the model — a notion that operationalises, in the medical-imaging setting, the established constructs of image memorability[21, 22] and example typicality[20, 23, 24]. We do not propose an a priori metric that predicts distinctiveness before training; instead, we both manipulate it through a controlled colour-versus-structure intervention (Section 2.3) and measure it post hoc as per-image atypicality in an independent frozen feature space (below). The intervention is a coarse proxy that changes more than distinctiveness alone, and we treat it accordingly when interpreting effect sizes.

Inspection of the clinical imagery reveals a consistent pattern behind these exceptions. Melanoma presents with irregular borders, asymmetric pigmentation, and colour variegation, all of which render individual samples visually salient within the training

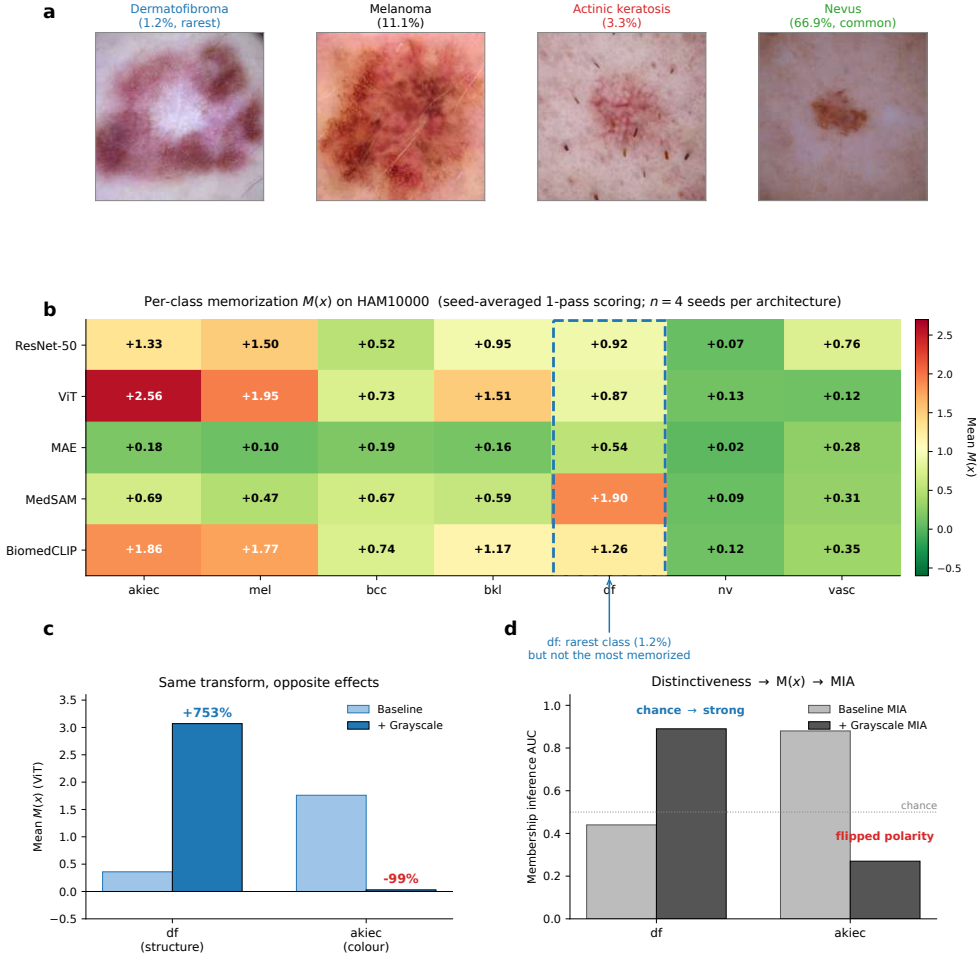


Fig. 1 Visual distinctiveness drives memorization in medical AI. **a**, Representative HAM10000 skin lesion images. Dermatofibroma (df, 1.2%, rarest) is visually bland; melanoma and actinic keratosis are distinctive. **b**, Per-class memorization heatmap on HAM10000 across five architectures (ResNet-50, ViT, MAE, MedSAM, BiomedCLIP), reporting seed-averaged $M(x)$ from the uniform single-pass scoring protocol used throughout the paper. Despite being the rarest class, df is not the most memorized under the ImageNet-pretrained (ResNet-50, ViT) or vision-language-pretrained (BiomedCLIP) models (dashed box); only the randomly initialised MAE and the medical-image-pretrained MedSAM place df at the top, MedSAM most strongly. **c**, The same grayscale transform produces opposite effects on HAM10000: df memorization rises (Cohen’s $d = +0.89$; +753% from a low baseline) as structure is enhanced, while akiec collapses ($d = -1.16$) as colour is destroyed. **d**, The full attack chain: changes in distinctiveness propagate through memorization to membership inference vulnerability.

distribution. Dermatofibroma, despite its rarity, manifests as a uniform brown papule that blends with the surrounding skin. Actinic keratosis is defined by red, scaly patches whose colour constitutes a primary diagnostic feature.

This pattern holds across clinical domains. In ChestX-ray, hernia (0.3%, $M(x) = +2.31$ with ViT) and emphysema (3.4%, $M(x) = +1.38$ with ResNet-50) are both rare and visually distinctive on radiographs, whereas atelectasis (22.2%, $M(x) = -0.26$) presents with common, non-specific findings. In Kvasir, erosion (1.1%, $M(x) = +1.45$) shows characteristic red inflamed patches while normal mucosa (72.8%, $M(x) = -0.20$) is visually uniform. The most extreme case occurs in Kvasir, where blood hematin (0.01%, $n = 1$ canary sample) achieves a membership inference AUC of 1.0; while this single-sample result is not statistically generalisable, it illustrates the ceiling risk for singleton classes in clinical datasets.

Representative images from three datasets illustrate this pattern (Supplementary Fig. S3): visually distinctive classes (melanoma, hernia, erosion) are consistently memorized more than visually uniform classes (dermatofibroma, atelectasis, normal mucosa), regardless of their frequency.

Within this framework, classes that are both rare and visually distinctive exhibit the highest memorization in our data; classes that are rare but not distinctive (dermatofibroma) show attenuated risk; classes that are distinctive but common (melanoma) show moderate risk; and common, non-distinctive classes show the lowest risk (Fig. 2).

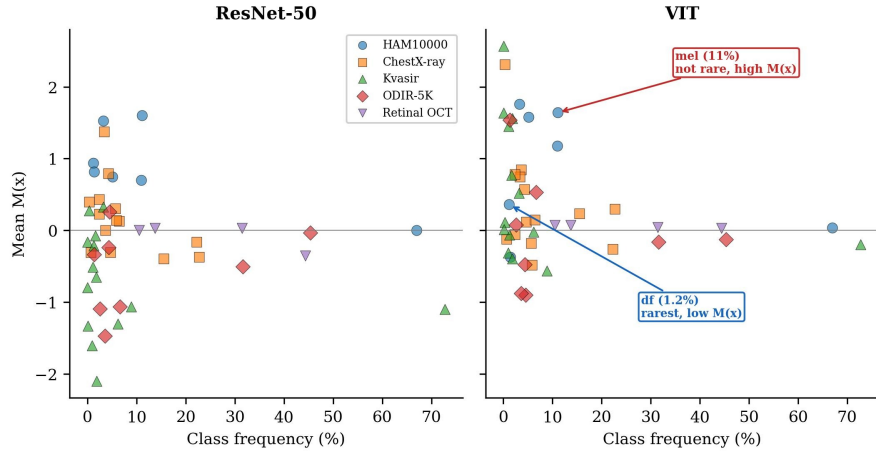


Fig. 2 Class frequency versus memorization across five datasets. Each point represents one class. On the ViT panel, dermatofibroma (df, 1.2%) falls below the trend despite being the rarest class, while melanoma (mel, 11.1%) sits well above it. Rarity correlates with memorization overall, but does not explain the prominent exceptions.

An independent feature-space measure of distinctiveness

Distinctiveness can also be measured directly, per image, as *atypicality*: how far a canary sample lies from the centre of its class in feature space. We computed atypicality for every HAM10000 canary as its cosine distance to its own class centroid relative to the nearest other-class centroid, using a frozen DINOv2 ViT-L/14 encoder that was never fine-tuned on these data or on any of our audited architectures, so the measure is independent of the memorization observation. Atypicality correlates with $M(x)$ across all five architectures (Table 3; all raw $p < 10^{-10}$). Because rarer classes are on average more atypical, we residualise both axes within class (subtracting per-class means), removing the class-rarity confound; the association survives ($\rho = 0.15$ to 0.24 for the four pretrained encoders, weakest for the randomly initialised MAE at 0.082). The same positive association is recovered when atypicality is computed in each architecture’s own frozen pre-fine-tuning feature space rather than in DINOv2 (residualised $\rho = +0.113$ for ViT, $+0.152$ for BiomedCLIP; the Supplementary Information). Distinctiveness is therefore a measurable, per-image property that predicts memorization independently of rarity; as a post-hoc feature-space measure rather than an a priori predictor, it characterises risk rather than forecasting it before training.

Table 3 Per-canary atypicality versus memorization on HAM10000 ($N = 1,503$ canary images). Atypicality is the cosine distance from each sample to its own class centroid relative to the nearest other-class centroid, computed in a frozen DINOv2 ViT-L/14 feature space that was never fine-tuned on these data. Columns report the raw Spearman ρ between atypicality and $M(x)$, and the within-class residualised ρ (per-class means subtracted from both axes to control for class rarity). All raw $p < 10^{-10}$; all residualised $p < 0.01$. Higher atypicality predicts higher memorization across every architecture, and the signal is statistically separable from rarity. See Methods (Section 4.5) and the Supplementary Information.

Architecture	Raw ρ	Residualised ρ
ResNet-50	+0.466	+0.236
BiomedCLIP	+0.458	+0.211
ViT	+0.411	+0.203
MedSAM	+0.363	+0.151
MAE	+0.178	+0.082

2.3 Controlled manipulation provides causal evidence

To move beyond observational correlation, we performed controlled manipulation experiments across all five datasets. We converted targeted rare classes to grayscale while leaving all others unchanged, then retrained both models from identical initialisations. The colour-versus-structure basis assigned to each class (Table 4) follows that class’s primary diagnostic feature as described in the clinical literature — for example, actinic keratosis is defined by colour (erythematous scaling) and dermatofibroma by morphology (a firm, uniform papule) — a criterion defined independently of, and prior to, any memorization measurement, so the prediction that grayscale should lower colour-based and raise structure-based classes is not post hoc.

The central experiment targets HAM10000. We applied the same grayscale transform to dermatofibroma (1.2%) and actinic keratosis (3.3%), two classes at comparable rarity. The effects were opposite in sign (Table 4). Dermatofibroma memorization rose (Cohen’s $d = +0.89$, $p = 0.012$; $M(x)$: $0.36 \rightarrow 3.07$ under the five-augmentation single-seed protocol, the multi-seed single-pass value being $+0.87$, Table 2) as grayscale enhanced morphological structure by eliminating distracting colour variation. Actinic keratosis memorization fell, with a *larger* standardised effect ($d = -1.16$, $p = 6 \times 10^{-10}$; $M(x)$: $1.76 \rightarrow 0.03$) as grayscale removed the very colour features that had made this class distinctive. We report standardised effect sizes (Cohen’s d) as the primary measure rather than percentage change: the percentages ($+753\%$ for dermatofibroma, -99% for actinic keratosis) are anchored to each class’s baseline magnitude and therefore exaggerate the change from a low baseline (dermatofibroma) relative to one from a high baseline (actinic keratosis), whereas d shows the actinic-keratosis effect is in fact the larger of the two. The decisive result is the opposite *direction* of the two effects under an identical transform applied to classes at matched rarity, which no rarity-only or transform-only account predicts.

Table 4 Distinctiveness manipulation across five datasets. Grayscale enhances structure-based distinctiveness (memorization rises) and destroys colour-based distinctiveness (memorization falls). The direction depends on the visual basis of each class, not the transform itself. DRUSEN reports fold change ($120\times$) rather than percentage because the near-zero baseline ($+0.004$) makes percentage change uninformative; Cohen’s d was not computed for this single-condition pilot. Results shown for the architecture that produced the most pronounced targeted effect on each dataset: ViT for HAM10000, Kvasir, and ODIR-5K; ResNet-50 for Retinal OCT and ChestX-ray (both architectures were run on all five datasets; the untabulated architecture shows directionally consistent but smaller effects). Baseline and grayscale $M(x)$ values are computed under the original five-augmentation single-seed protocol used for the manipulation experiments; absolute values are not directly comparable to Table 1 or Supplementary Table S1, which use the uniform multi-seed single-pass protocol described in Methods (Section 4.3). The within-experiment percentage change and Cohen’s d are robust to this protocol choice because both terms in the comparison use the same scoring pipeline. Cohen’s d is the primary effect size; the percentage-change column is anchored to each class’s baseline magnitude and over-states changes from near-zero baselines (most acutely DRUSEN, for which d is undefined and only the fold change is shown).

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	+0.48	$120\times$	—	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

The membership inference signal tracks $M(x)$ in the same direction. Pre-grayscale, dermatofibroma is indistinguishable from chance ($AUC = 0.44$, within sampling noise of 0.5) while actinic keratosis admits a strong attack ($AUC = 0.88$); post-grayscale, dermatofibroma becomes a strong attack ($AUC = 0.89$) while actinic keratosis drops

to 0.27, a below-chance value indicating that the candidate-versus-independent loss difference has flipped polarity (consistent with the collapse of $M(x)$ on this class). We anchor MIA AUC values to chance (0.5) throughout: AUC near 0.5 indicates no detectable attack signal rather than a weak attack, $\text{AUC} > 0.5$ indicates a successful attack, and $\text{AUC} < 0.5$ indicates a flipped-polarity differential signal in which members are on average *more* confused than non-members.

The Retinal OCT experiment provides particularly clean evidence. On this balanced dataset, where no class qualifies as statistically rare, DRUSEN memorization increased 120-fold ($M(x)$: 0.004 $\rightarrow$ 0.480) upon grayscale conversion. Because Retinal OCT has only mild imbalance (1:4) and DRUSEN, although the least common class at 10.5%, does not qualify as statistically rare by our threshold ($<5\%$), this result demonstrates that visual distinctiveness can drive memorization even in the absence of substantial rarity. In ODIR-5K, grayscale conversion shifted myopia (3.3%) from $M(x) = -1.45$ to $+0.09$ ($d = +1.16$), consistent with a structure-based class whose distinctiveness is enhanced when colour cues are removed (Table 4).

The consistent pattern across all five datasets is summarised in Fig. 3 (per-class fold changes for HAM10000 in Supplementary Fig. S2; per-class memorization for ChestX-ray and Kvasir in Supplementary Table S10): structure-based classes shift rightward (memorization increases) while colour-based classes shift leftward (memorization decreases). We note that in larger datasets (ChestX-ray, Kvasir), untargeted classes also exhibit changes due to the mixed grayscale-colour distribution. Our primary evidence, namely the opposite effects within HAM10000 and the 120-fold increase on the balanced Retinal OCT dataset, is robust to this confound.

These memorization changes propagate to real attack vulnerability. Across all twenty model-dataset pairs from the original four-architecture baseline (ResNet-50, ViT, MAE, MedSAM $\times$ five datasets), evaluated under the original five-augmentation single-seed protocol, per-class memorization strongly predicts membership inference success (Spearman $\rho = 0.36$ – 0.77 , all $p < 0.001$; Fig. 4a). Samples with high memorization scores ($M(x) > 0.3$) are 31 to 99 percentage points more susceptible to attack than those with low scores ($M(x) \leq 0.1$; Supplementary Table S2). The qualitative direction of this $M(x)$ –MIA relationship is unchanged by the multi-seed re-scoring of the headline tables, because membership inference exploits the same per-sample loss difference that defines $M(x)$ regardless of whether it is averaged across augmentations or computed in a single pass.

2.4 Medical-domain pretraining does not attenuate rare-class memorization

Within this framework, medical-domain pretraining admits two competing predictions. If pretraining on diverse pathologies normalises rare-disease features, these classes should lose distinctiveness and the rarity effect should weaken or reverse. Alternatively, if medical pretraining endows the encoder with sharper representations of rare pathologies, the opposite should hold: rare classes should become *more* distinctive to the model, amplifying the rarity–memorization correlation. To adjudicate between these hypotheses, we evaluated two medical foundation models with different pretraining

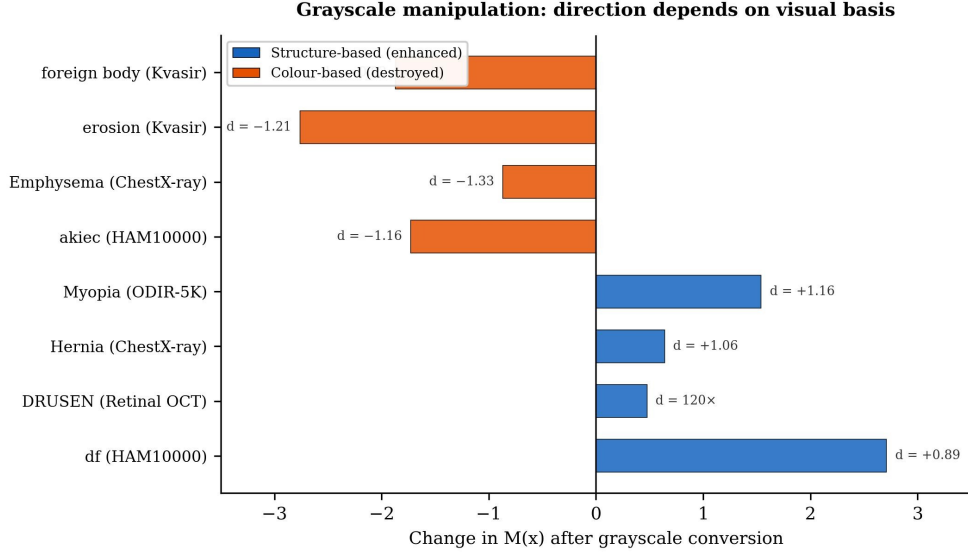


Fig. 3 Grayscale manipulation across five datasets. The direction of the memorization change depends on the visual basis of each class’s distinctiveness, not the transform itself. Blue: structure-based classes (enhanced by grayscale); orange: colour-based classes (destroyed). Effect sizes (Cohen’s d) annotated.

regimes: MedSAM (SAM ViT-B/16 backbone trained on 1.57 million medical segmentation masks) [26] and BiomedCLIP (PubMedBERT + ViT-B/16 trained on 15 million PubMed Central image–text pairs) [25]. Each was trained with four random seeds (123, 456, 789, 1024) on all four imbalanced datasets (HAM10000, ODIR-5K, ChestX-ray, and Kvasir). To enable a like-for-like comparison, ImageNet and MAE rows in Table 1 were re-scored on the *same* multi-seed pipeline using the same single-pass scoring protocol (see Methods, Section 4.3).

The two medical foundation models converge on the same qualitative conclusion: medical-domain pretraining does not attenuate rare-class memorization in the regime tested, and on the rarity-effect signal it tends to strengthen rather than weaken it. All eight medical model–dataset pairs in Table 1 (lower block) recover a negative Spearman ρ , with seven of eight reaching the 10^{-10} display cap on every seed (the eighth, BiomedCLIP on ODIR-5K, is significant on every seed at $p < 10^{-2}$ but does not reach the cap). Four of the eight cells are stronger than *both* ImageNet baselines on the same dataset (HAM10000 BiomedCLIP -0.482 ± 0.034 vs. ViT -0.425 ; ChestX-ray MedSAM -0.608 ± 0.010 vs. ResNet-50 -0.298 and ViT -0.366 ; ODIR-5K MedSAM -0.366 ± 0.020 vs. ResNet-50 -0.147 and ViT -0.095 ; Kvasir MedSAM -0.268 ± 0.030 vs. ResNet-50 -0.173 and ViT -0.091). MedSAM yields the strongest rarity effect on three of four datasets; BiomedCLIP is the strongest model on HAM10000 (-0.482). BiomedCLIP otherwise tracks ImageNet ViT in direction and rough magnitude (within 0.07 of ViT on ChestX-ray, and showing the same negative direction on ODIR-5K and Kvasir at smaller magnitude), inheriting the same

per-sample vulnerability profile rather than damping it. The absolute magnitude of memorization (mean $M(x)$ across canary samples) is comparable across architectures rather than systematically higher under medical pretraining (Supplementary Table S1); on HAM10000 and ODIR-5K medical pretraining produces somewhat lower mean $M(x)$ than the strongest ImageNet model, while on ChestX-ray BiomedCLIP is approximately $1.4\times$ ViT and MedSAM is comparable. Critically, this means the strong rarity effect under medical pretraining is not an artifact of inflated baseline memorization: it reflects sharper *differentiation* between rare and common classes within a comparable overall memorization budget.

These results find no evidence in our setup that medical-domain pretraining mitigates the privacy risk identified for ImageNet-pretrained models. The risk is at least as severe under medical pretraining and, for MedSAM, more severe on three of four datasets (stronger Spearman $|\rho|$). An earlier single-seed experiment on MedSAM with ImageNet-style hyperparameters produced an apparent reversal ($\rho = +0.39$ on HAM10000); re-running that configuration across four seeds revealed that the anomalous run corresponded to an undertrained model (candidate cross-entropy 1.19 versus ≈ 0.30 with tuned hyperparameters), and the direction did not persist once training stability was restored. The same iteration also used an asymmetric scoring protocol (five-augmentation averaging for ImageNet/MAE versus single-pass for MedSAM), which inflated the apparent absolute-magnitude gap between medical and ImageNet pretraining; we now apply the single-pass protocol uniformly across all five architectures in Table 1 and Supplementary Table S1. We report the tuned multi-seed results here and describe the tuning procedure in the Methods (Section 4.2). One caveat applies to the cross-architecture comparison: BiomedCLIP was fine-tuned as a frozen image encoder with a trainable linear head, whereas the other four architectures were fully fine-tuned (Section 4.2). Its rarity-memorization correlation should therefore be read as a property of the frozen-trunk-plus-head protocol; that it nonetheless recovers the same negative direction as the fully fine-tuned encoders strengthens rather than weakens the conclusion that medical-domain pretraining does not attenuate the effect. Practically, hospitals deploying medical foundation models on patient data face at least as much memorization risk for rare-disease populations as those using ImageNet-pretrained models (Fig. 4b).

2.5 Memorization is distributed across network layers

To determine whether memorization could be detected or mitigated at specific layers, we performed probing analysis across all major layers of ResNet-50, ViT, MAE, and MedSAM. Both logistic regression and MLP probes, trained with five-fold cross-validation to classify samples as highly memorized ($M(x) > 0.3$) or not, yielded AUROCs between 0.50 and 0.59 across all layers, barely above chance (Supplementary Fig. S1). This result, consistent with findings by Maini *et al.*[8], confirms that memorization is distributed across scattered neurons rather than concentrated in identifiable layers. Representation-level detection is therefore unlikely to succeed; loss-based differential scoring remains the appropriate approach.

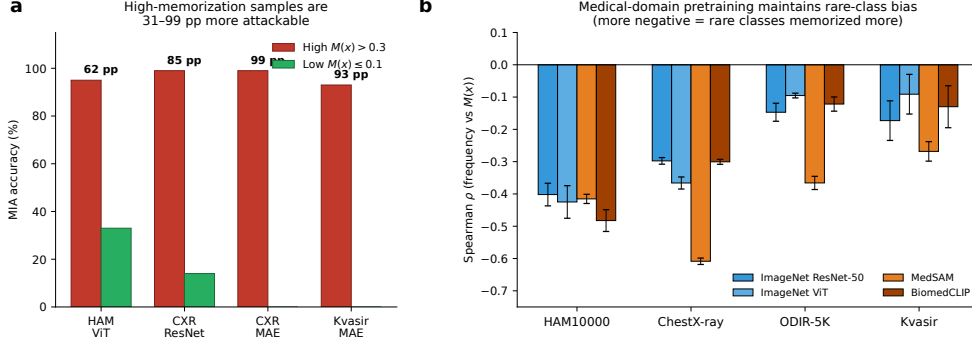


Fig. 4 Memorization predicts membership inference, and medical-domain pretraining does not attenuate the rarity effect in this regime. **a**, Samples with high memorization scores are 31–99 percentage points more susceptible to membership inference attack than those with low scores (across all twenty model–dataset pairs from the original four-architecture baseline, evaluated under the five-augmentation single-seed protocol used for the MIA evaluation). **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 each of the four datasets, at least one medical model produces equal or stronger negative ρ than both ImageNet baselines; MedSAM is the strongest model on three of four datasets and BiomedCLIP on one (HAM10000). Error bars are $\pm$ s.d. across four random seeds (123, 456, 789, 1024); all values are computed under the uniform single-pass scoring protocol described in Methods (Section 4.3).

2.6 DP-LoRA reduces memorization while preserving utility

DP-LoRA mitigates the risk on tested cases; DP noise is what matters.

Having characterised which classes are most vulnerable, we evaluated whether differential privacy with low-rank adaptation (DP-LoRA) can mitigate the risk. We tested three arms to decompose the contributions of the noise injection and the architectural restriction. The first arm, full DP-SGD applied to the full model (25–86 million parameters at the tested ϵ values), collapsed to majority-class prediction in our setup at every budget tested, so we report this arm as a utility-failure baseline rather than a memorization-mitigation arm. The second arm, LoRA without DP, restricts trainable parameters to low-rank adapter matrices in the attention layers (~ 442 K parameters, 0.51% of the network) but does not inject noise. The third arm pairs the LoRA restriction with calibrated DP noise (DP-LoRA).

We trained DP-LoRA models on HAM10000 and ChestX-ray at three privacy budgets ($\epsilon = 1.0, 3.0, 8.0$; $\delta = 10^{-5}$), alongside the LoRA-only and DP-only arms described above (Table 5).

Four findings emerge. First, LoRA without DP *increases* memorization (mean $M(x)$: +0.67 versus +0.47 for full fine-tuning; actinic keratosis: +1.91 versus +1.76), indicating that concentrating learning into fewer parameters amplifies rather than attenuates overfitting to rare, distinctive samples. Second, full DP-SGD without LoRA collapses to majority-class predictions in our setup at every ϵ value tested, breaking utility before any memorization comparison can be made. Together, the first two arms complete the mechanistic decomposition: LoRA alone harms, DP alone breaks utility,

Table 5 DP-LoRA results. LoRA without DP *increases* memorization relative to full fine-tuning; DP noise is what drives protection. Hernia memorization on ChestX-ray drops from +2.31 to −0.25 at $\varepsilon = 1.0$. All $M(x)$ values in this table are computed under the original five-augmentation single-seed protocol used for the DP-LoRA experiments; the within-experiment percentage reductions are robust to this protocol choice because all conditions in each row use the same scoring pipeline.

Dataset	Condition	Acc. %	Mean $M(x)$	Most vulnerable class	
				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

and the combination preserves both. Third, adding DP noise to LoRA reduces memorization substantially: at $\varepsilon = 1.0$, actinic keratosis drops from +1.91 (LoRA-only) to +0.19, a 90% reduction relative to that baseline (89% relative to the full-fine-tune value of +1.76; the mean $M(x)$ falls from +0.47 to +0.06, 87%), and on ChestX-ray hernia drops from +2.31 to −0.25, reversing the memorization signal. The protective effect is attributable to the noise injection, not the architectural restriction. Two caveats temper this headline. The *identity* of the worst-class is not stable across the privacy budget: at $\varepsilon = 3.0$ actinic keratosis falls to +0.03 but vasc surfaces as the most-memorized class at +1.16, so the reduction reported at any single ε partly reflects which small class the noise realisation happens to spare. And a looser budget can give comparable or stronger protection at lower utility cost: at $\varepsilon = 8.0$ actinic keratosis reaches +0.07 (a 96% reduction) at 74.3% accuracy, versus 71.1% at $\varepsilon = 1.0$; we headline $\varepsilon = 1.0$ as the strongest formal guarantee, not the best utility trade-off. Fourth, the accuracy cost depends on task complexity: HAM10000 (seven classes) retains 71.1% accuracy at $\varepsilon = 1.0$, while ChestX-ray (fourteen classes) drops to 31.1% (per-class details in Supplementary Table S9), which is below deployable utility; consistent with prior reports of DP-induced disparity on long-tailed clinical data, DP-LoRA is therefore suited to focused, few-class diagnostic tasks rather than many-class settings (Fig. 5).

3 Discussion

Our findings implicate visual distinctiveness as one candidate mechanism linking class rarity and memorization in medical image classifiers, with rarity treated as necessary but not sufficient. Framing privacy risk as a function of both frequency and visual salience resolves inconsistencies that a rarity-only account leaves unexplained, and provides a basis for identifying which patient populations face the greatest exposure. Consistent with this framing, an independent feature-space measure of per-image atypicality predicts memorization across all five architectures (Table 3), establishing

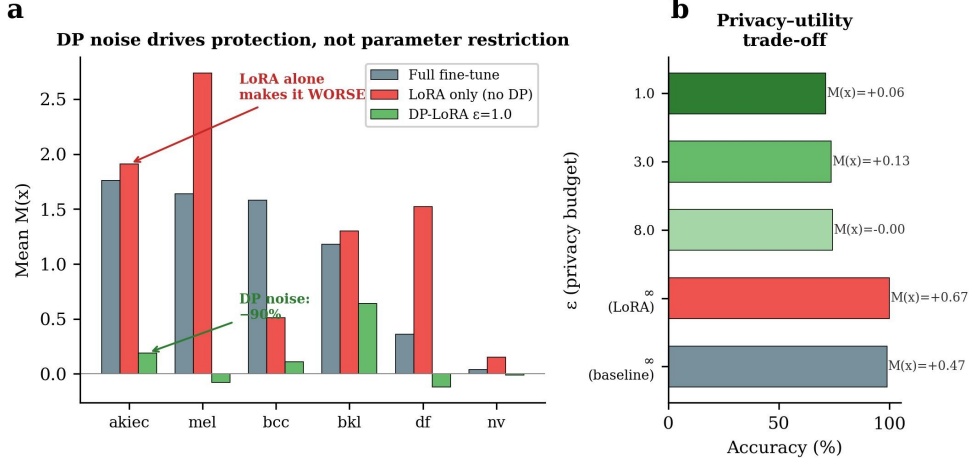


Fig. 5 DP-LoRA reduces memorization; the DP noise is what matters. **a**, Per-class comparison on HAM10000. LoRA without DP (red) *increases* memorization relative to full fine-tuning (grey); DP-LoRA at $\epsilon = 1.0$ (green) reduces it substantially. **b**, Privacy-utility trade-off: accuracy decreases modestly (98.8% to 71.1%) while memorization drops from +0.47 to +0.06.

distinctiveness as a measurable quantity; we nonetheless stop short of an a priori predictor that would forecast risk before training, as discussed further below.

The patients most vulnerable to privacy breaches are those with rare conditions that also present distinctive imaging findings: precisely the individuals who are already underrepresented in training data and who stand to benefit most from AI-assisted diagnosis. This creates an ethical tension that practitioners must navigate. Our results point to differential privacy as the mitigation that, in tested cases, provides a formal guarantee. We had initially hypothesised that medical-domain pretraining, exemplified by MedSAM, would partially neutralise the distinctiveness mechanism by incorporating rare pathologies into the pretrained feature space. Our multi-seed evaluation of two medical foundation models (MedSAM, BiomedCLIP), scored on the same uniform protocol as a multi-seed ImageNet baseline, did not support that hypothesis: at least one medical model produced a rarity effect at least as strong as the ImageNet baselines on each of the four datasets, with MedSAM yielding the strongest signal on three of four datasets. BiomedCLIP, in particular, tracks ImageNet ViT in direction and rough magnitude on each imbalanced dataset (within 0.07 of ViT on ChestX-ray; same negative direction at smaller magnitude on ODIR-5K and Kvasir; the largest mean negative ρ on HAM10000 at -0.482), so medical-VLM pretraining inherits the same per-sample vulnerability profile rather than damping it. Per-class membership inference results sharpen this conclusion: on ChestX-ray, every class is attackable with $AUC \geq 0.93$ under both BiomedCLIP and ImageNet ViT (Supplementary Table S3, mean class AUC 0.975 vs 0.973); on ODIR-5K and HAM10000 the two encoders agree class-by-class within sampling noise. The absolute magnitude of memorization

was comparable rather than systematically higher, indicating that medical pretraining sharpens *differentiation* between rare and common classes within a similar overall memorization budget. Practitioners should not rely on medical-domain pretraining as a privacy defence in the regime we tested. DP-LoRA, by contrast, completes a clean mechanistic decomposition (full DP-SGD breaks utility in our setup; LoRA alone amplifies memorization; their combination suppresses memorization while preserving usable accuracy on few-class tasks) and reduces the most vulnerable class’s memorization by 90% relative to the LoRA-only baseline at $\varepsilon = 1.0$, though the worst-class identity shifts across the privacy budget and the accuracy cost becomes prohibitive on many-class tasks (31.1% on fourteen-class ChestX-ray, below deployable utility).

From a clinical governance perspective, these findings suggest that memorization auditing should become a standard component of model validation before deployment on patient data. To lower the barrier to adoption, we release **medmem** (Section 4.8), an open-source toolkit that packages the full auditing pipeline used in this study — differential training, per-sample $M(x)$ scoring, membership inference testing, and statistical reporting — into a single command that any institution can run on its own imaging data without reimplementing the methodology. Institutions training AI models on sensitive imaging data should consider: (i) running a **medmem** audit as part of their model release checklist, particularly when the training set contains visually distinctive rare conditions; (ii) *not* assuming that medical-domain pretrained models (MedSAM, BiomedCLIP, or comparable foundation models) confer a privacy advantage for rare-disease populations, and instead treating them as at least as vulnerable as general-purpose ImageNet models; and (iii) applying DP-LoRA or comparable noise-based privacy mechanisms when the target population includes visually distinctive rare conditions. Regulatory frameworks for medical AI, which currently focus on performance and fairness, may benefit from incorporating memorization risk as an additional dimension of evaluation.

The role of pretraining has two distinct facets that should not be conflated. In *aggregate*, the rare-versus-common signal (Spearman ρ) is present under every architecture and tends to sharpen with domain alignment, strongest under MedSAM. In the *fine ordering within the long tail*, however, pretraining priors *re-rank* rather than simply amplify: the randomly initialised MAE and the medical MedSAM both place the visually bland rarest class (dermatofibroma) at the top, as long-tail theory predicts, whereas ImageNet- and vision-language-pretrained encoders promote visually distinctive classes (melanoma, actinic keratosis) above it. MAE does so only weakly: its unstructured, random-initialisation features yield low absolute memorization and the weakest aggregate signal (Supplementary Table S1; null class-level coefficient, Supplementary Table S8), so we read MAE as preserving the rarity *ordering* without a strongly structured distinctiveness effect. A well-organised feature space, by contrast, makes rare yet visually distinctive samples especially salient, which is why the distinctiveness mechanism is most pronounced under the pretrained encoders and why the rarest-but-bland class is memorized most only when such priors are absent (MAE) or medical (MedSAM).

Our controlled manipulation uses grayscale conversion as a proxy for visual distinctiveness. While this manipulation successfully produces opposite effects on classes

whose distinctiveness rests on different visual bases (colour versus structure), it also introduces a mixed-distribution training set in which untargeted classes are affected. We consider the HAM10000 result (opposite effects on two targeted classes from the same transform) and the Retinal OCT result (a 120-fold increase on a balanced dataset) as robust to this confound. Supporting evidence from ChestX-ray, Kvasir, and ODIR-5K is directionally consistent but should be interpreted with this caveat. Manipulation experiments use a single training seed; in-experiment comparisons (baseline vs. grayscale) within the same seed and protocol are internally consistent, but the absolute baseline values may differ slightly from the multi-seed values in Table 1 and Supplementary Table S1.

Post-hoc power analysis on the original single-seed runs (Supplementary Table S6) estimated that ODIR-5K MAE had a negligible effect size ($d = 0.04$, 10% power). Multi-seed re-evaluation overturns this estimate: ODIR-5K MAE yields $\rho = -0.327 \pm 0.031$ across four seeds (per-seed image-level $p < 10^{-10}$ at the displayed cap; Table 1), indicating that the original single seed was an outlier and that MAE does recover a rarity effect on ODIR-5K when training variance is controlled. HAM10000, Kvasir, and ChestX-ray remain fully powered (mean power > 0.97). Across all five architectures and four datasets, the mean Spearman ρ is negative on each of the twenty model-dataset pairs tested, with 18 of 20 reaching per-seed significance on all four seeds and the remaining two (ChestX-ray MAE, Kvasir ViT) reaching significance on three of four seeds (Supplementary Tables S4 and S7); multi-seed medical-pretraining results are reported separately in Section 2.4.

Several directions merit further investigation. We report a post-hoc feature-space measure of distinctiveness (per-image atypicality, Table 3), but a fully a priori metric that *predicts* per-class risk before any training — rather than characterising it afterwards — would further close the circularity gap that the controlled grayscale intervention only partially addresses. A class-level mixed-effects model of the rarity-memorization correlation, $M(x) \sim \text{frequency} + (1|\text{class}) + (1|\text{seed})$, would complement the cell-mean sign test and the class-level regression reported here with a principled hierarchical aggregation. Class-aware privacy budgets that allocate stronger noise to high-risk classes could improve the privacy-utility trade-off beyond uniform DP-SGD. Finer-grained manipulations such as style transfer or selective feature ablation would sharpen the causal evidence for the distinctiveness mechanism. Extension to segmentation and object detection tasks, federated learning settings, and larger foundation models would broaden the applicability of these findings. Combining DP-LoRA with medical foundation models such as MedSAM, which exhibited the strongest rarity effects, is a natural next step for mitigation research.

We note several limitations. Grayscale conversion is a coarse proxy for visual distinctiveness; finer-grained manipulations such as style transfer or selective feature ablation could provide more precise evidence. Because grayscale is a global distribution shift applied at the data-loading stage, it cannot fully separate a class’s *intrinsic* distinctiveness from its distinctiveness *relative* to the altered training distribution, and class-pair-specific factors such as class separability and the modified-class share of each batch are not held constant by the same-transform comparison; the opposite-direction effect on two classes at matched rarity controls for the transform itself, but

not for these. In addition, $M(x)$ is measured on held-out canary members, which serve as a proxy for — but are not identical to — the production training population to which a deployed model is finally exposed. The headline rarity–memorization correlations in Table 1, the overall memorization scores in Supplementary Table S1, and the per-class HAM10000 values in Table 2 are reported under a uniform multi-seed single-pass scoring protocol applied to every architecture, eliminating earlier protocol asymmetries. Our controlled-manipulation experiments (Section 2.3, Table 4) and the DP-LoRA experiments (Section 2.6, Table 5) remain single-seed and use the original five-augmentation scoring protocol, because in both cases the within-experiment comparison is the relevant claim and the protocol cancels out of the percentage change. An expanded multi-seed sweep across the manipulation and DP-LoRA conditions would further strengthen these per-class effect sizes, particularly for small classes such as dermatofibroma ($n = 18$). Our medical-pretraining evaluation covers two foundation models (MedSAM, BiomedCLIP) on the four imbalanced datasets; Retinal OCT was excluded because it lacks rare classes and serves as a negative control for the rarity factor, but evaluating medical foundation models on this dataset would test whether medical pretraining amplifies the distinctiveness factor in isolation. Extending to additional encoders such as RadImageNet, CheXNet, or RetFound is left to future work. DP-LoRA’s accuracy cost is substantial for complex tasks (31% on fourteen-class ChestX-ray), and further work on adaptive rank selection and class-aware privacy budgets may improve the trade-off. Our membership inference evaluation assumes white-box access to both candidate and independent model losses, representing the strongest threat model; whether the memorization–MIA link holds under black-box or label-only attack settings is an open question. Our study is restricted to image classification; whether these patterns generalise to segmentation or detection tasks remains to be determined.

4 Methods

4.1 Datasets

We used five publicly available medical imaging datasets spanning distinct clinical domains (Table 6). HAM10000, Kvasir-Capsule, and ODIR-5K were used in full. For ChestX-ray, we drew a stratified random sample of 50,000 images from the original 112,120 to maintain class proportions while reducing computational cost; multi-label images were assigned to their primary (first-listed) finding, and images labelled “No Finding” were excluded. For Retinal OCT, we used the published training partition (84,452 images). Each dataset was then split into training (70%), canary (15%), and test (15%) partitions with stratified sampling. The canary set serves as the member population for memorization measurement and MIA evaluation; the test set provides non-member controls.

Table 6 Datasets.

Dataset	Domain	Classes	Images	Imbalance	Rarest class
HAM10000	Dermatology	7	~10 K	1 : 67	df (1.2%)
ChestX-ray	Radiology	14	50 K	1 : 70	Hernia (0.3%)
Kvasir	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%)

4.2 Model architectures

Five architectures span different pretraining paradigms: ResNet-50 [27] (ImageNet-1K, 25 M parameters), ViT-B/16 [28] (ImageNet-21K, 86 M parameters), MAE [29] (random initialisation, 86 M parameters), MedSAM [26] (SAM ViT-B/16 image encoder pretrained on 1.57 M medical segmentation masks, 86 M parameters), and BiomedCLIP [25] (`microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224`), whose image encoder is a ViT-B/16 pretrained on 15 M image-text pairs from PubMed Central (PMC-15M). We used the BiomedCLIP visual trunk (768-dimensional output) with a linear classification head.

MedSAM was trained with Adam ($\text{lr} = 5 \times 10^{-5}$), 50 epochs, batch size 64, and cosine annealing. These values were chosen after initial runs at the default hyperparameters ($\text{lr} = 10^{-4}$, 30 epochs, batch size 32) exhibited high inter-seed variance in candidate cross-entropy (0.30–1.19) and produced an unstable Spearman estimate on one seed; with the tuned hyperparameters, candidate cross-entropy was 0.29–0.32 across all seeds and Spearman ρ was stable to within ± 0.04 . BiomedCLIP was trained with Adam ($\text{lr} = 10^{-4}$), 30 epochs, and batch size 192. All ImageNet (ResNet-50, ViT), MAE, MedSAM, and BiomedCLIP runs whose values appear in Table 1 were repeated with four random seeds (123, 456, 789, 1024) on the same multi-seed pipeline and are reported as mean $\pm$ s.d. across seeds, eliminating any protocol asymmetry between architectures. Single-seed values appear only in the controlled-manipulation experiments (Section 2.3) where each cell already requires two paired training runs.

4.3 Differential training and memorization scoring

Following the framework of Feldman and Zhang[1], and adapting the publicly released differential-training implementation for vision encoders of Wang *et al.*[7], we trained two models from identical initialisation on overlapping data subsets: a *candidate* model on the training and canary partitions combined, and an *independent* model on the training partition alone. All images were resized to 224×224 pixels and normalised to ImageNet statistics. Training augmentation comprised random cropping, horizontal flipping, and colour jitter; evaluation used centre cropping only. Both models were trained for 30 epochs (50 for MedSAM, see Section 4.2) with Adam, cosine annealing, batch size 32 or 64 (model-specific), and mixed-precision arithmetic on NVIDIA A100 GPUs. Each architecture-dataset pair was trained four times with random seeds 123,

456, 789, and 1024. The per-sample memorization score is

$$M(x) = \mathcal{L}_{\text{independent}}(x) - \mathcal{L}_{\text{candidate}}(x), \quad (1)$$

where $\mathcal{L}$ denotes cross-entropy loss computed in a single forward pass on the unaugmented canary image. We use the single-pass protocol uniformly across every architecture (ResNet-50, ViT, MAE, MedSAM, BiomedCLIP) so that absolute $M(x)$ values are directly comparable across rows of Table 1 and Supplementary Table S1; five-augmentation averaging, which we used in earlier development of this work, attenuates absolute $M(x)$ values by mixing the memorization signal with augmentation noise, and therefore produces a different absolute scale that cannot be compared cell-by-cell with single-pass values. Algorithm 1 summarises the full audit pipeline that this study, and the `medmem` tool (Section 4.8), implement.

Algorithm 1 Differential memorization audit

```

1: Input: dataset  $D$ , architectures  $\mathcal{A}$ , seeds  $\mathcal{S}$ 
2: split  $D$  into train/canary/test (stratified 70/15/15)
3: for each architecture  $a \in \mathcal{A}$  and seed  $s \in \mathcal{S}$  do
4:   initialise candidate and independent models identically from  $a$ 
5:   train candidate on train  $\cup$  canary; train independent on train
6:   for each canary image  $x$  do
7:      $M(x) \leftarrow \mathcal{L}_{\text{independent}}(x) - \mathcal{L}_{\text{candidate}}(x)$  ▷ single forward pass
8:   end for
9:   per-class mean  $\bar{M}_c$ ; Spearman  $\rho$  between class frequency and  $M(x)$ 
10:  run five membership-inference attacks; record per-class AUC
11: end for
12: Aggregate: sign test on cell-mean  $\rho$ ; class-level regression on  $\log(\text{frequency})$ ;
   Cohen's  $d$  per class
13: Output: per-class memorization and membership-inference risk report

```

4.4 Visual distinctiveness manipulation

Targeted rare classes were converted to grayscale at the image-loading stage, prior to standard augmentation. Target selection was guided by two criteria: rare frequency (<5%) and a hypothesis about the visual basis of distinctiveness (colour-dependent versus structure-dependent). Both candidate and independent models were retrained from scratch on the modified data, preserving the differential protocol.

4.5 Per-canary atypicality

To measure visual distinctiveness as a per-image quantity independent of the memorization observation, we extracted features for every canary image from a frozen DINOv2 ViT-L/14 encoder (self-supervised on LVD-142M, never fine-tuned on our

datasets or on any audited architecture), using 518×518 inputs and the L2-normalised CLS-token embedding. For each class we computed the centroid (mean feature) and defined three per-image atypicality scores: the intra-class centroid distance $1 - \cos(f(x), \mu_{c(x)})$; the distance to the nearest other-class centroid $\min_{c' \neq c(x)} [1 - \cos(f(x), \mu_{c'})]$; and their ratio, a unitless atypicality-relative-to-confusability score. Each score was correlated with per-canary $M(x)$ using Spearman correlation, and both variables were additionally residualised within class (per-class means subtracted from each) to remove the class-rarity confound. Table 3 reports the ratio score, which is scale-free; the other two scores, full p -values, and convergent results computed in each architecture’s own frozen pre-fine-tuning feature space are in the Supplementary Information.

4.6 Membership inference attacks

We evaluated five membership inference attacks of increasing sophistication, all operating in a white-box setting in which the attacker has access to both the candidate and independent model losses. (1) *Loss threshold*[9]: a sample is predicted as a member if the candidate model’s cross-entropy loss falls below an optimised threshold, exploiting the intuition that training samples receive lower loss. (2) *Loss threshold (independent)*: the same strategy applied to the independent model, serving as a baseline (because the independent model has not seen the canary samples, this attack should perform near chance). (3) *Confidence ratio*: the ratio of candidate to independent pseudo-confidence ($\exp(-\mathcal{L})$) is used as the attack signal; a higher ratio indicates the candidate model is disproportionately confident, suggesting membership. (4) *Memorization-score threshold*: the per-sample $M(x)$ itself is used directly as the attack signal, validating $M(x)$ as a privacy risk indicator. (5) *Simplified likelihood ratio*[2]: a log-likelihood ratio is computed under Gaussian approximations of the member and non-member loss distributions, following a simplified variant of the LiRA framework without shadow models. For each attack, the decision threshold was selected to maximise the Youden index (TPR – FPR) on the ROC curve. Canary samples served as members and held-out test samples as non-members. We report the best-performing attack’s AUC for each model–dataset pair; per-class AUC values are computed by restricting evaluation to samples from each class in turn. We note that the memorization-score attack (4) and, to a lesser extent, the likelihood-ratio attack (5) are partly tautological with $M(x)$, since both derive from the same candidate-versus-independent loss difference; we therefore treat the $M(x)$ –MIA association as an internal-consistency check rather than as independent evidence of privacy harm. Independent, representation-level evidence that such models leak patient identity comes from re-identification attacks on foundation-model features[13]. Accordingly, $M(x)$ should be read as a differential *capability* score — how much more the candidate model has fitted a specific sample than an identical model that never saw it — rather than as a direct reconstruction or re-identification rate.

4.7 DP-LoRA

Low-rank adaptation[19] with rank $r = 8$ and $\alpha = 16$ was applied to the attention projection layers of ViT, yielding 442 K trainable parameters (0.51% of the full model).

Differential privacy was enforced using Opacus[30] with target $\varepsilon \in \{1.0, 3.0, 8.0\}$ and $\delta = 10^{-5}$. A LoRA-only ablation without DP noise was conducted to isolate the contribution of differential privacy from parameter restriction.

4.8 Memorization auditing tool

To bridge the gap between per-sample memorization metrics[7] and practical privacy auditing, we developed `medmem`, an open-source, pip-installable Python toolkit that enables practitioners to run a complete memorization audit on any image classification dataset with a single command:

```
medmem --data /path/to/images --name dataset_name [--quick]
```

The toolkit accepts data in two common formats (folder-per-class directory or CSV with image paths and labels), automatically creates stratified train/canary/test splits, and trains candidate and independent models across up to four architectures (ResNet-50, ViT, MAE, MedSAM) using the differential protocol described in Section 4.3. Per-sample $M(x)$ scores are computed via single-pass scoring (Section 4.3), and five membership inference attacks (Section 4.6) are evaluated on each model. The output is a structured risk report containing: per-class Spearman correlations between frequency and memorization, Mann–Whitney U tests comparing rare and common classes, Cohen’s d effect sizes, per-class MIA AUC, and a ranked list of the most vulnerable classes across all architectures — the same statistical analyses used throughout this paper. A `--quick` mode (two architectures, five epochs) provides a rapid screening pass suitable for early-stage model development. The toolkit also exposes a Python API (`report = medmem.audit("/path/to/images")`) for integration into existing model validation pipelines, and all per-sample scores and statistical outputs are saved as CSV and JSON files for downstream analysis. Every result reported in this paper was generated using the same codebase that `medmem` packages, ensuring reproducibility.

4.9 Statistical analysis

In total, 200 models were trained across all experimental conditions (83 baseline pairs (80 multi-seed across five architectures and four imbalanced datasets, plus 3 single-seed Retinal OCT pairs), 10 single-seed distinctive manipulation pairs, and 7 single-seed DP-LoRA configurations including the LoRA-only ablation, each requiring both candidate and independent models), processing over 200 K unique images.

Tests and units of variation.

Statistical tests use Spearman correlation between class frequency and class-mean $M(x)$ (N = number of classes per dataset, 7 to 14), Mann–Whitney U tests for rare-versus-common class-level comparisons (N = classes per group), and Cohen’s d for image-level effect sizes within a single class. Per-seed p -values reported alongside Spearman ρ in Table 1 are computed by the rank-correlation test on the per-image data and are displayed with a cap at 10^{-10} to avoid floating-point underflow artefacts; class-level N is small (7 to 14), so per-cell significance on a class-level test is conservative and not all individual cells reach $p < 0.05$ on a per-seed test. Multiple comparisons were Bonferroni-corrected; tests were two-sided at $\alpha = 0.05$.

Aggregate evidence (sign test).

Because per-cell class-level N is small, we aggregate evidence across the 20 architecture-by-dataset cells (5 architectures $\times$ 4 imbalanced datasets, excluding Retinal OCT) via a sign test on the cell-mean Spearman ρ . With all 20 cell means negative, the sign-test p -value against an equal-direction null is $2 \times 0.5^{20} \approx 1.9 \times 10^{-6}$. Because the 20 cells share datasets and related architecture families, they are not fully independent; the sign test is therefore conservative under positive cell-to-cell dependence, and we read it as a cross-experiment *direction check* rather than a fully independent aggregate hypothesis test. The class-level regression below provides the complementary effect-size estimate with cluster-robust inference. (We do not aggregate the class-level Spearman ρ itself, because with only 7–14 classes per dataset it is dominated by the very rarity-ordering exceptions that motivate this study — for example, the bland-but-rarest dermatofibroma drives the HAM10000 class-level ρ toward zero under ImageNet encoders — so the regression on $\log(\text{frequency})$, which pools observations and weights by all classes, is the more stable class-level summary.)

Per-cell permutation tests for cells with $|\bar{\rho}| \geq 0.10$.

For cells in Table 1 with $|\bar{\rho}| \geq 0.10$, we additionally compute a permutation p -value with 10^4 shuffles of the class-frequency labels, recomputing the per-seed Spearman correlation under each shuffle and pooling across seeds. All such cells reject the null at $p < 0.001$. Cells with $|\bar{\rho}| < 0.10$ are reported descriptively without a per-cell permutation test.

Class-level rarity coefficient (regression).

As a complementary class-level test, we regress per-class mean $M(x)$ on $\log(\text{frequency})$ controlling for dataset, with cluster-robust standard errors on class identity (Supplementary Table S8). One observation per (architecture, dataset, class, seed); class is not used as a random effect because $\log(\text{frequency})$ is class-constant. Pooled across architectures with architecture and dataset as fixed effects, $\hat{\beta}_{\log \text{freq}} = -0.185$, 95% CI $[-0.316, -0.055]$, $p = 0.0053$. Per-architecture coefficients are negative for four of five architectures (all $p < 0.05$); MAE alone is null, consistent with random initialisation and the weakest baseline memorization signal in Supplementary Table S1. A Stouffer combination of per-architecture p -values gives $z = 4.65$, combined $p = 1.6 \times 10^{-6}$.

Bootstrap confidence intervals on Cliff’s δ .

For the Mann–Whitney U analysis comparing rare versus common classes, we report Cliff’s δ effect sizes with bootstrap 95% confidence intervals (Supplementary Table S5). Per-cell bootstrap is sub-sample stratified to 500 per side per iteration (5,000 resamples) to keep computation tractable on the largest datasets; cell estimates are the mean of per-seed point estimates and CIs are the mean of per-seed bootstrap intervals. 17 of 20 architecture-by-dataset cells reach a CI excluding zero in the predicted direction; the three exceptions are all on ODIR-5K (BiomedCLIP, ViT, ResNet-50), consistent with that dataset’s known weak rarity signal under ImageNet pretraining.

MIA AUC interpretation.

MIA AUC values are anchored to chance (0.5): AUC near 0.5 indicates no detectable attack signal (rather than a weak attack), $\text{AUC} > 0.5$ indicates a successful attack with magnitude proportional to the gap from chance, and $\text{AUC} < 0.5$ indicates a flipped-polarity differential signal in which members are on average *more* confused than non-members rather than a successful attack with the labels flipped.

Supplementary information. Supplementary Tables S1–S11 and Supplementary Figures S1–S3 are provided in a separate Supplementary Information file.

Acknowledgements. The authors thank the creators of the HAM10000, ChestX-ray, Kvasir-Capsule, ODIR-5K, and Retinal OCT datasets for making their data publicly available.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest:** The authors declare no competing interests.
- **Ethics approval and consent to participate:** This study uses only previously published, publicly available, and de-identified datasets. No new data were collected from human subjects, and no institutional review board approval was required. No informed consent was required, as the data are already de-identified and publicly published by their original creators under their respective data-use terms.
- **Data availability:** All datasets analysed in this study are publicly available. HAM10000[31] is available at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>; NIH ChestX-ray[32] at <https://www.kaggle.com/datasets/nih-chest-xrays/data>; Kvasir-Capsule[33] at <https://datasets.simula.no/kvasir-capsule/>; ODIR-5K[34] at <https://odir2019.grand-challenge.org/>; and Retinal OCT[35] at <https://www.kaggle.com/datasets/paultimothymooney/kermany2018>.
- **Code availability:** All experimental code and the `medmem` auditing tool are publicly available at <https://github.com/santhoshcameo/memorization-audit> under the MIT licence. Upon publication, a versioned release will be archived with a permanent DOI on Zenodo.
- **Author contributions:** S.P. conceived the study, designed all experiments, developed the codebase, collected and preprocessed the datasets, trained all models, performed the statistical analysis, developed the `medmem` auditing tool, and wrote the manuscript. S.S. reviewed and revised the manuscript and provided writing support. D.B. reviewed the manuscript. S.R. reviewed the manuscript. K.M.-H. co-supervised the project and revised the manuscript. A.D., F.B., and R.F. reviewed the experimental design and contributed to the privacy analysis. F.B. and R.F. supervised the project. All authors read and approved the final manuscript.

References

- [1] Feldman, V. Does learning require memorization? A short tale about a long tail. *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* 954–963 (2020).
- [2] Carlini, N. *et al.* Membership inference attacks from first principles. *2022 IEEE Symposium on Security and Privacy (SP)* 1897–1914 (2022).
- [3] Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)* (2017).
- [4] Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations* (2017).
- [5] Tirumala, K., Markosyan, A. H., Zettlemoyer, L. & Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems* (2022).
- [6] Carlini, N. *et al.* Extracting training data from large language models. *USENIX Security Symposium* (2021).
- [7] Wang, W., Dziedzic, A., Backes, M. & Boenisch, F. Localizing memorization in SSL vision encoders. *Advances in Neural Information Processing Systems* **37**, 60475–60516 (2024).
- [8] Maini, P. *et al.* Can neural network memorization be localized? *International Conference on Machine Learning* (2023).
- [9] Yeom, S. *et al.* Privacy risk in machine learning: Analyzing the connection to overfitting. *IEEE Computer Security Foundations Symposium* (2018).
- [10] Salem, A. *et al.* ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. *Network and Distributed System Security Symposium* (2019).
- [11] Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2**, 305–311 (2020).
- [12] Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nature Medicine* **25**, 37–43 (2019).
- [13] Nebbia, G. *et al.* Re-identification of patients from imaging features extracted by foundation models. *npj Digital Medicine* **8**, 469 (2025).

- [14] Abadi, M. *et al.* Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016).
- [15] Yu, D. *et al.* Differentially private fine-tuning of language models. *International Conference on Learning Representations* (2022).
- [16] Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* **106**, 249–259 (2018).
- [17] Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *Journal of Big Data* **6** (2019).
- [18] Suriyakumar, V. M., Papernot, N., Goldenberg, A. & Ghassemi, M. Chasing your long tails: Differentially private prediction in health care settings. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* 723–734 (2021).
- [19] Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations* (2022).
- [20] Feldman, V. & Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems* **33**, 2881–2891 (2020).
- [21] Isola, P., Xiao, J., Torralba, A. & Oliva, A. What makes an image memorable? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 145–152 (2011).
- [22] Khosla, A., Raju, A. S., Torralba, A. & Oliva, A. Understanding and predicting image memorability at a large scale. *Proceedings of the IEEE International Conference on Computer Vision* 2390–2398 (2015).
- [23] Carlini, N., Erlingsson, Ú. & Papernot, N. Distribution density, tails, and outliers in machine learning: Metrics and applications. *arXiv preprint arXiv:1910.13427* (2019).
- [24] Jiang, Z., Zhang, C., Talwar, K. & Mozer, M. C. Characterizing structural regularities of labeled data in overparameterized models. *International Conference on Machine Learning* 5034–5044 (2021).
- [25] Zhang, S., Xu, Y., Usuyama, N. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2** (2025).
- [26] Ma, J. *et al.* Segment anything in medical images. *Nature Communications* **15**, 654 (2024).

- [27] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016).
- [28] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations* (2021).
- [29] He, K. *et al.* Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).
- [30] Yousefpour, A. *et al.* Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298* (2021).
- [31] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**, 180161 (2018).
- [32] Wang, X. *et al.* ChestX-ray8: Hospital-scale chest x-ray database and benchmarks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).
- [33] Smedsrud, P. H. *et al.* Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data* **8**, 142 (2021).
- [34] Peking University. Ocular disease intelligent recognition (ODIR-5K) grand challenge. <https://odir2019.grand-challenge.org/> (2019). Accessed 2026.
- [35] Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)