Supplementary Information for "Visual distinctiveness drives memorization beyond rarity in fine-tuned medical imaging models"

Supplementary Information

Cross-references to Tables, Figures, Sections, and Equations without an "S" prefix (e.g. "Table 1", "Section 2.4") refer to the main text. Items prefixed with "S" are within this Supplementary Information.

Per-canary atypicality: an independent feature-space measure of distinctiveness

To check that visual distinctiveness 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 audited architecture) and correlate it with per-canary $M(x)$ (Methods, Section 4.5). For each canary x we define three scores: the intra-class centroid distance, the distance to the nearest other-class centroid, and their ratio. Table S11 reports raw and within-class-residualised Spearman correlations with $M(x)$ for all five architectures and all three scores ($N = 1{,}503$ canary images). Across the 15 architecture-by-score cells with a defined residualised test, 13 are significant at $p < 0.05$; the two exceptions are the nearest-other-class distance for ViT and MedSAM. The scale-free ratio score (used in the main-text Table 3) is the most consistent, with within-class residualised ρ between 0.082 (MAE) and 0.236 (ResNet-50). 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 (within-class residualised $\rho = +0.114$ for ResNet-50, $+0.113$ for ViT, $+0.103$ for MedSAM, $+0.152$ for BiomedCLIP; MAE $+0.025$, not significant), confirming that the signal is not an artefact of the particular reference encoder.

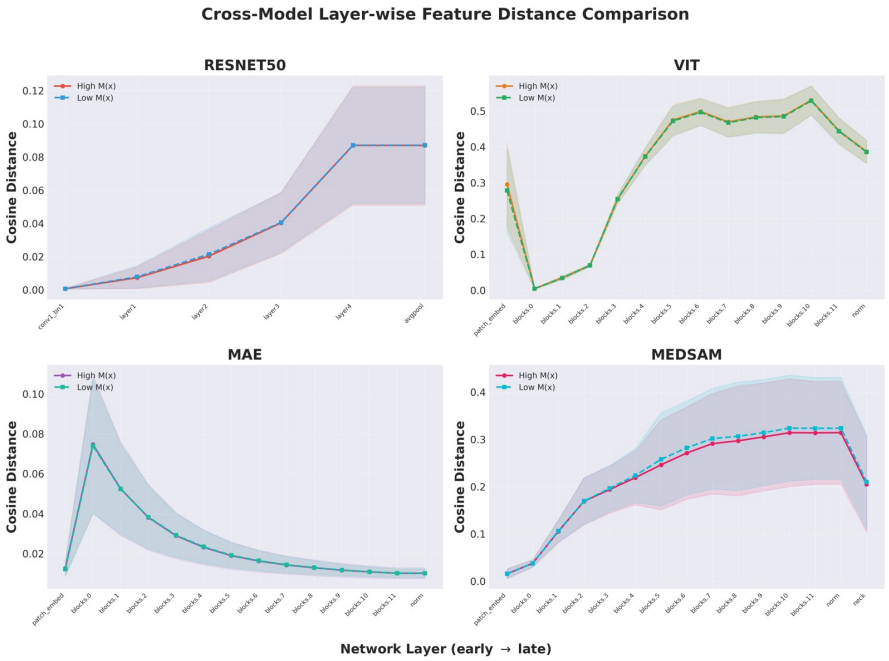

Fig. S1 Layer-wise memorization probing across four architectures (HAM10000). Cosine distance between candidate and independent model activations at each layer, split by high-$M(x)$ (red) and low-$M(x)$ (blue) samples. Overlapping confidence bands confirm that no single layer differentiates memorized from non-memorized samples, supporting the conclusion that memorization is distributed.

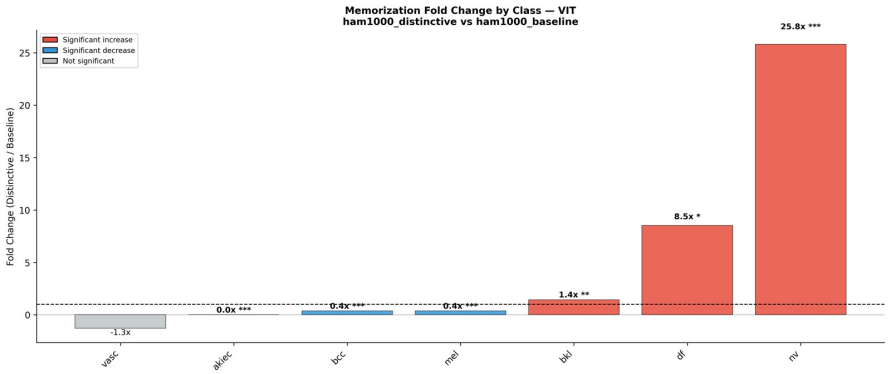

Fig. S2 Distinctiveness manipulation: baseline versus grayscale per-class memorization (HAM10000, ViT). Fold change for each class shows that dermatofibroma (df) increases 8.5× (equivalently +753%; Table 4) while actinic keratosis (akiec) collapses to near-zero. Untargeted classes also shift because grayscale alters the global training distribution: the common class nv rises from +0.04 to +0.92 ($\approx 26\times$ in fold terms, inflated by its near-zero baseline), which illustrates both the mixed-distribution confound (Section 2.3) and why fold change is unreliable for low baselines. Only df and akiec were targeted; the decisive evidence is the *opposite sign* of their changes under the identical transform, not the fold magnitudes.

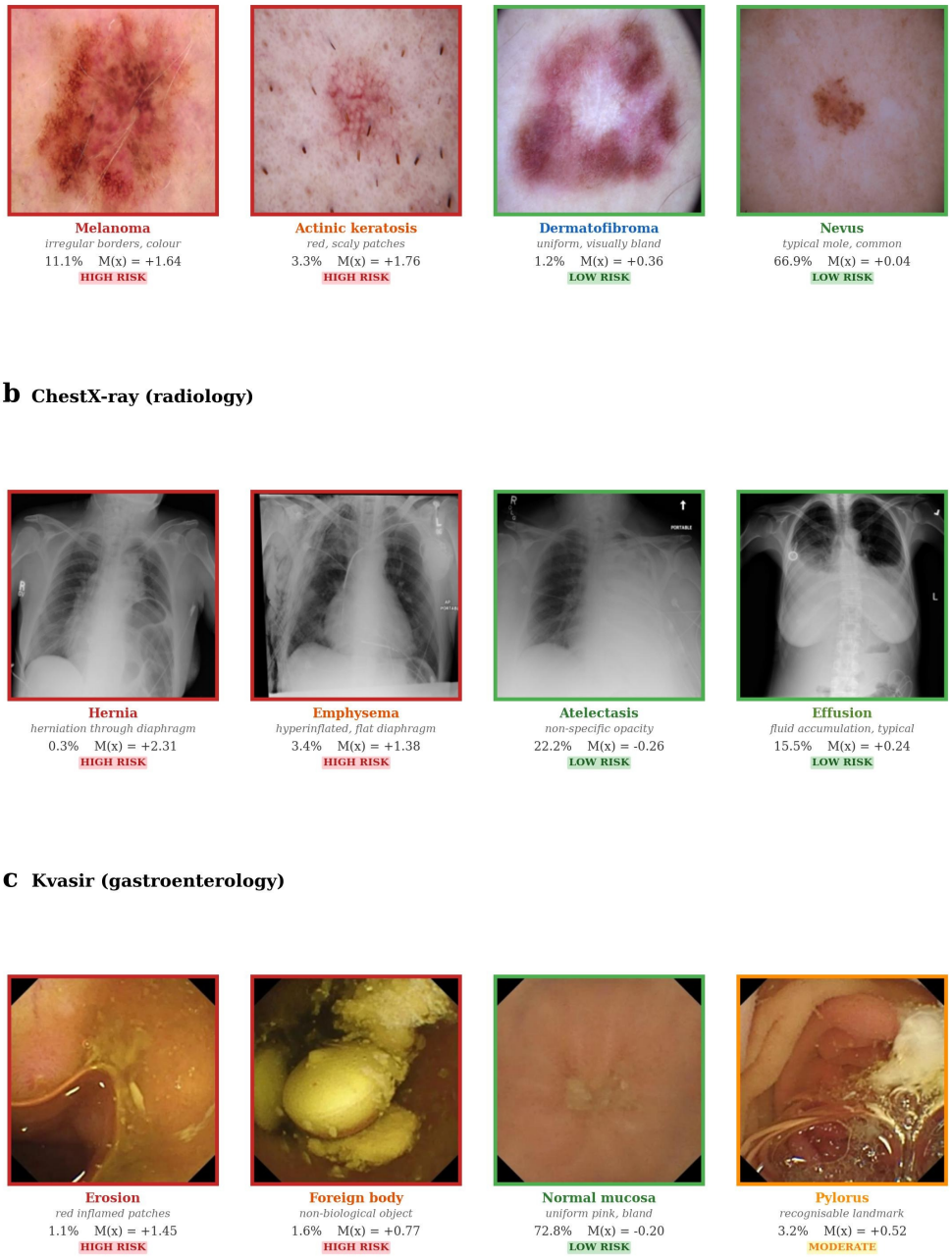

Fig. S3 Visual distinctiveness varies across clinical domains. Representative images from three datasets, with class frequency, mean $M(x)$, and memorization risk indicated. **a,** HAM10000: melanoma and actinic keratosis are visually distinctive (irregular borders, colour) and highly memorized; dermatofibroma is the rarest class but visually bland and at low risk. **b,** ChestX-ray: hernia and emphysema show distinctive radiographic findings; atelectasis and effusion are common and non-specific. **c,** Kvasir: erosion and foreign body are colour-distinctive; normal mucosa is visually uniform.

Table S1 Overall memorization scores under the uniform multi-seed single-pass scoring protocol (Section 4.3). Mean $M(x)$ across canary samples, reported as mean ± s.d. across four random seeds (123, 456, 789, 1024). All architectures (ResNet-50, ViT, MAE, MedSAM, BiomedCLIP) are scored on the same pipeline so that absolute values are directly comparable across rows. MedSAM and BiomedCLIP were not evaluated on Retinal OCT, which lacks rare classes and serves only as a negative control. Retinal OCT values are from the original single-seed run (seed 42) and are the only exception to the uniform multi-seed protocol.

Dataset	Model	n	Mean $M(x)$ (mean ± s.d.)
HAM10000	ResNet-50	4	+0.411 ± 0.015
	ViT	4	+0.601 ± 0.062
	MAE	4	+0.067 ± 0.003
	MedSAM	4	+0.262 ± 0.004
	BiomedCLIP	4	+0.523 ± 0.013
ChestX-ray	ResNet-50	4	+1.818 ± 0.003
	ViT	4	+1.830 ± 0.071
	MAE	4	+0.097 ± 0.038
	MedSAM	4	+1.613 ± 0.025
	BiomedCLIP	4	+2.484 ± 0.024
ODIR-5K	ResNet-50	4	+1.175 ± 0.018
	ViT	4	+1.706 ± 0.222
	MAE	4	+0.286 ± 0.067
	MedSAM	4	+1.239 ± 0.029
	BiomedCLIP	4	+1.149 ± 0.027
Kvasir	ResNet-50	4	+0.033 ± 0.001
	ViT	4	+0.029 ± 0.002
	MAE	4	+0.094 ± 0.004
	MedSAM	4	+0.068 ± 0.007
	BiomedCLIP	4	+0.031 ± 0.006
Retinal OCT	ResNet-50	1	−0.142 *(single seed)*
	ViT	1	+0.046 *(single seed)*
	MAE	1	+0.022 *(single seed)*

Table S2 Per-class membership inference attack AUC for selected datasets and architectures, from the original four-architecture single-seed evaluation (seed 42, five-augmentation scoring). The $M(x)$ column therefore uses a different absolute scale from Table 1 and Supplementary Table S1, which use the uniform multi-seed single-pass protocol; within-table comparisons remain valid. Best-performing attack method reported. * Singleton classes (blood hematin, ampulla of Vater; $n = 1$ canary each) yield a degenerate AUC = 1.0 that is not statistically generalisable; these are shown unemphasised and excluded from any summary statistic. Anchoring AUC to chance (0.5): 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 in which the candidate-versus-independent 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

Table S3 Per-class membership inference attack AUC for BiomedCLIP and ImageNet ViT on the four imbalanced datasets, computed from the multi-seed candidate vs independent loss differences and pooled across seeds. Members are canary samples; non-members are held-out test samples. AUC bootstrap 95% CIs are computed with 1,000 resamples within each class. Per-dataset summary (mean class AUC; fraction of classes with AUC > 0.55): ChestX-ray BiomedCLIP 0.975 (14/14), ViT 0.973 (14/14); ODIR-5K BiomedCLIP 0.876 (8/8), ViT 0.879 (8/8); HAM10000 BiomedCLIP 0.749 (7/7), ViT 0.720 (7/7); Kvasir BiomedCLIP 0.550 (4/12), ViT 0.549 (5/12). On every dataset, BiomedCLIP per-class AUCs track ImageNet ViT within sampling noise, indicating that BiomedCLIP inherits the same per-sample vulnerability profile rather than damping it. The ChestX-ray pattern is striking: every class is attackable with AUC ≥ 0.93 under both encoders, with class-to-class coefficient of variation of (AUC $- 0.5$) below 0.06.

Dataset	Class	BiomedCLIP AUC	ViT AUC
ChestX-ray (uniformly high under both encoders):			
ChestX-ray	Pneumonia	1.000	1.000
ChestX-ray	Pleural Thickening	0.999	0.999
ChestX-ray	Consolidation	0.997	0.998
ChestX-ray	Fibrosis	0.994	0.995
ChestX-ray	Hernia	0.969	0.967
ChestX-ray	Cardiomegaly	0.931	0.930
ODIR-5K:			
ODIR-5K	AMD	0.928	0.903
ODIR-5K	Other	0.888	0.881
ODIR-5K	Cataract	0.674	0.765
HAM10000:			
HAM10000	akiec	0.820	0.755
HAM10000	mel	0.808	0.791
HAM10000	df	0.789	0.748
HAM10000	nv	0.569	0.572
Kvasir (mostly chance under both encoders):			
Kvasir	erosion	0.691	0.657
Kvasir	foreign body	0.565	0.573
Kvasir	normal clean mucosa	0.504	0.508

Selected classes shown for compactness; full per-class table with bootstrap CIs is in `biomedclip_per_class_mia.csv`. Score is $-\text{loss}_{\text{candidate}}$; AUC is computed with the candidate model's own per-sample loss as the discriminator on canary (member) versus held-out test (non-member) samples. This is a loss-based MIA without shadow-model calibration (i.e. not LiRA); per-class AUC values are therefore an upper bound on attack strength relative to a calibrated reference, but the BiomedCLIP-versus-ViT comparison is unaffected because both encoders use the identical pipeline. AUC values around 0.5 indicate no detectable attack signal rather than a weak attack.

Table S4 Mann–Whitney U tests comparing memorization between rare (<5%) and common (≥5%) classes, from the original single-seed evaluation (seed 42, five-augmentation scoring). Only the six model–dataset pairs with the clearest rare/common separation are shown; the full multi-seed Spearman analysis in Table 1 covers all twenty model–dataset pairs. Positive difference indicates rare classes are memorized more (supports hypothesis). All tests two-sided; p-values below 10^{-10} are reported as $< 10^{-10}$.

Dataset	Model	Rare mean	Common mean	Diff.	p-value
Supports hypothesis (rare > common):					
HAM10000	ResNet-50	+1.24	+0.31	+0.92	$< 10^{-10}$
HAM10000	ViT	+0.97	+0.44	+0.52	0.012
HAM10000	MAE	+0.31	+0.07	+0.24	3.2×10^{-7}
Kvasir	ViT	+0.52	−0.22	+0.74	$< 10^{-10}$
ChestX-ray	ResNet-50	+0.37	−0.19	+0.56	$< 10^{-10}$
ChestX-ray	ViT	+0.51	+0.02	+0.48	$< 10^{-10}$

Mann–Whitney U tests for multi-seed MedSAM and BiomedCLIP are deferred to Section 2.4 of the main text; both medical foundation models produce rare > common memorization on each of the four datasets tested. Multi-seed effect-size estimates with bootstrap confidence intervals across all 20 architecture-by-dataset cells are reported in Supplementary Table S5.

Table S5 Multi-seed Cliff's δ (rare versus common classes) with bootstrap 95% confidence intervals, pooled across seeds. Class-level N ranges from 7 to 14 per dataset; seeds within each cell range from 4 to 9. Bootstrap is sub-sample stratified to 500 per side per iteration with 5,000 resamples; cell mean δ is taken across seed runs and pooled CI is the mean of per-seed bootstrap CIs. Positive δ indicates rare classes are memorized more (supports the rarity hypothesis); CIs that exclude zero indicate the direction is robust to resampling. 17 of 20 cells reach a CI excluding zero; the three exceptions are all on ODIR-5K (BiomedCLIP, ViT, ResNet-50), consistent with that dataset's known weak rarity signal under ImageNet pretraining.

Dataset	Architecture	$\bar{\delta}$	95% CI low	95% CI high	n_{seeds}
ChestX-ray	ResNet-50	+0.380	+0.352	+0.485	5
ChestX-ray	ViT	+0.417	+0.366	+0.494	5
ChestX-ray	MAE	+0.149	+0.058	+0.200	5
ChestX-ray	MedSAM	+0.538	+0.493	+0.586	5
ChestX-ray	BiomedCLIP	+0.355	+0.309	+0.443	4
HAM10000	ResNet-50	+0.576	+0.434	+0.679	5
HAM10000	ViT	+0.525	+0.378	+0.638	5
HAM10000	MAE	+0.352	+0.173	+0.498	5
HAM10000	MedSAM	+0.455	+0.300	+0.571	9
HAM10000	BiomedCLIP	+0.685	+0.564	+0.757	4
ODIR-5K	ResNet-50	+0.062	−0.040	+0.175	5
ODIR-5K	ViT	−0.026	−0.132	+0.093	5
ODIR-5K	MAE	+0.463	+0.372	+0.560	5
ODIR-5K	MedSAM	+0.291	+0.173	+0.401	5
ODIR-5K	BiomedCLIP	−0.018	−0.138	+0.108	4
Kvasir	ResNet-50	+0.237	+0.172	+0.305	5
Kvasir	ViT	+0.189	+0.136	+0.266	5
Kvasir	MAE	+0.241	+0.202	+0.341	5
Kvasir	MedSAM	+0.347	+0.290	+0.424	5
Kvasir	BiomedCLIP	+0.175	+0.111	+0.256	4

Table S6 Post-hoc statistical power analysis. Power for detecting rare versus common memorization differences. Retinal OCT is omitted (no rare classes <5%). Values are from single-seed (seed 42) runs used in the original power analysis; multi-seed re-evaluation (Table 1, n=4) reveals that ODIR-5K MAE, originally estimated as underpowered ($d = 0.04$, 10% power), in fact yields a robust multi-seed $\rho = -0.327 \pm 0.031$ (per-seed image-level $p < 10^{-10}$ at the displayed cap), indicating that the original single-seed estimate was an outlier. MedSAM and BiomedCLIP multi-seed Cohen's d values are summarised in Section 2.4 of the main text.

Dataset	Model	n_{rare}	n_{common}	Cohen's d	Power
HAM10000	ResNet-50	88	1,415	0.79 (medium)	1.000
HAM10000	ViT	88	1,415	0.39 (small)	0.942
HAM10000	MAE	88	1,415	0.73 (medium)	1.000
HAM10000	MedSAM	88	1,415	0.47 (small)	0.990
ChestX-ray	ResNet-50	1,622	5,878	1.00 (large)	1.000
ChestX-ray	ViT	1,622	5,878	0.90 (large)	1.000
Kvasir	ResNet-50	869	6,216	0.30 (small)	1.000
Kvasir	ViT	869	6,216	0.82 (large)	1.000
ODIR-5K	ResNet-50	320	1,598	0.15 (negl.)	0.659
ODIR-5K	ViT	320	1,598	0.32 (small)	0.999
ODIR-5K	MAE	320	1,598	0.04 (negl.)	0.099
ODIR-5K	MedSAM	320	1,598	0.45 (small)	1.000

Table S7 Summary of rarity–memorization direction by model, updated with n=4 multi-seed results (Table 1). "Supports" indicates that the mean Spearman ρ across four seeds is negative ($p < 0.05$ on at least three of four seeds at the per-seed image-level test). The multi-seed protocol resolves two cases that appeared non-significant in single-seed analyses: ODIR-5K MAE ($\rho = -0.327 \pm 0.031$, per-seed $p < 10^{-10}$ at the displayed cap) and ODIR-5K ResNet-50 ($\rho = -0.147 \pm 0.028$, per-seed $p < 10^{-4}$).

Model	Supports	Against	Interpretation
ResNet-50	4/4	0	Strong support
ViT	4/4	0	Strong support†
MAE	4/4	0	Strong support‡
MedSAM	4/4	0	Strong support (multi-seed, n=4)
BiomedCLIP	4/4	0	Strong support (multi-seed, n=4)

† Kvasir ViT: mean ρ negative on all four seeds but one seed ($p = 0.77$) does not reach significance individually. ‡ ChestX-ray MAE: mean ρ negative on three of four seeds; one seed ($\rho = +0.01$, $p = 0.34$) is not significant. Overall effect is small ($|\rho| = 0.046$).

Table S8 Class-level rarity coefficients from a regression of per-class mean $M(x)$ on log(frequency), controlling for dataset (fixed effect), with cluster-robust standard errors on class identity. The unit of variation is the class ($N = 47$ class instances across the five datasets, each with up to 9 multi-seed observations). A negative coefficient means rare classes are memorized more. The pooled fit additionally controls for architecture as a fixed effect and uses cluster-robust SE on (dataset, class). Pooled $\hat{\beta}_{\log \text{freq}} = -0.185$, 95% CI $[-0.316, -0.055]$, $p = 0.0053$. Stouffer combination of per-architecture p-values: $z = 4.65$, $p = 1.6 \times 10^{-6}$.

Architecture	$\hat{\beta}_{\log \text{freq}}$	95% CI	p-value	n_{obs}
ResNet-50	−0.201	[−0.356, −0.046]	0.011	219
ViT	−0.295	[−0.503, −0.087]	0.0054	219
MAE	−0.031	[−0.079, +0.016]	0.20	219
MedSAM	−0.152	[−0.258, −0.047]	0.0045	247
BiomedCLIP	−0.266	[−0.487, −0.045]	0.018	172
Pooled (all architectures)	−0.185	[−0.316, −0.055]	0.0053	1,076

Per-architecture coefficients are from OLS of per-class-mean $M(x)$ on log(freq) with dataset as a fixed effect; observations are class-mean $M(x)$ values, one per (architecture, dataset, class, seed). Cluster-robust standard errors on class identity absorb within-class non-independence across seeds. Class is not used as a random effect because log(frequency) is class-constant; a class random intercept would absorb the predictor. MAE's null result is consistent with random initialisation lacking a structured feature space, mirroring its weakest baseline memorization signal in Table S1.

Table S9 Per-class memorization under DP-LoRA for ChestX-ray (ViT), from the single-seed DP-LoRA experiments (five-augmentation scoring; see Table 5 caption for protocol details). Hernia, the rarest and most memorized class, drops from +2.31 to −0.25 at $\varepsilon = 1.0$ (111% reduction). Accuracy cost is substantial for this fourteen-class task. Three classes (Pneumonia, Pleural Thickening, Mass) are omitted for brevity; their baseline $M(x)$ values fall within the range shown.

Class	Freq. %	Baseline	$\varepsilon = 8.0$	$\varepsilon = 1.0$
Hernia	0.3	+2.31	+0.46	−0.25
Edema	3.6	+0.84	+0.12	+0.08
Fibrosis	2.4	+0.78	+0.27	+0.14
Emphysema	3.4	+0.75	+0.31	+0.11
Pneumothorax	4.3	+0.57	+0.18	+0.11
Infiltration	22.7	+0.30	−0.04	−0.11
Effusion	15.5	+0.24	+0.06	+0.02
Consolidation	6.4	+0.15	+0.09	−0.00
Cardiomegaly	4.7	+0.12	−0.03	−0.13
Atelectasis	22.2	−0.26	−0.05	+0.12
Nodule	5.8	−0.48	−0.01	−0.04
Overall accuracy		61.0%	34.3%	31.1%

Table S10 Per-class memorization for ChestX-ray and Kvasir (ViT), from the original single-seed evaluation (seed 42, five-augmentation scoring). Absolute $M(x)$ values use a different scale from the multi-seed single-pass protocol in Tables 1 and S1; the per-class ranking is robust to protocol choice. Classes sorted by mean $M(x)$. The most and least memorized classes in each dataset differ by over two orders of magnitude, underscoring class-specific vulnerability.

Dataset	Class	Freq. %	n	Mean $M(x)$
ChestX-ray (ViT), top five and bottom two:				
	Hernia	0.3	24	+2.31
	Edema	3.6	271	+0.84
	Fibrosis	2.4	178	+0.78
	Emphysema	3.4	253	+0.75
	Pneumothorax	4.3	320	+0.57
		...		
	Atelectasis	22.2	1667	−0.26
	Nodule	5.8	437	−0.48
Kvasir (ViT), top five and bottom two:				
	blood hematin	0.03	1	+2.57
	ampulla of Vater	0.02	1	+1.64
	erosion	1.1	76	+1.45
	foreign body	1.6	116	+0.77
	pylorus	3.2	229	+0.52
		...		
	normal clean mucosa	72.8	5151	−0.20
	ileocecal valve	8.9	629	−0.56

Table S11 Per-canary atypicality versus $M(x)$ on HAM10000 ($N = 1{,}503$), across three feature-space scores computed in a frozen DINOv2 ViT-L/14 backbone. Raw and within-class-residualised Spearman ρ (per-class means subtracted from both axes). All raw $p < 10^{-6}$. Residualised cells are significant at $p < 0.05$ except where marked † (nearest-other-class distance for ViT and MedSAM). The ratio score is reported in the main-text Table 3.

Score	Architecture	Raw ρ	Residualised ρ
Intra-class centroid distance:			
	ResNet-50	+0.404	+0.179
	BiomedCLIP	+0.409	+0.156
	ViT	+0.352	+0.131
	MedSAM	+0.344	+0.098
	MAE	+0.220	+0.119
Nearest-other-class centroid distance:			
	ResNet-50	+0.149	+0.062
	BiomedCLIP	+0.168	+0.055
	ViT	+0.124	+0.021†
	MedSAM	+0.159	+0.026†
	MAE	+0.136	+0.086
Ratio (atypicality relative to confusability):			
	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