BATCH SIZE IS NOT SAMPLE SIZE: NEAR-TIE AMPLIFICATION IN CONTRASTIVE MEMBERSHIP INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset inference asks whether a set of examples was used to train a model. At web scale each example carries almost no evidence, so audits average a per-example score over the set and rely on a square-root law: separability grows as the square root of the number of examples. We show that this law fails for contrastive image–text models when the score is the InfoNCE loss computed within a batch, because batch size is then not sample size: a larger batch averages more pairs but also gives every pair more negatives to beat. On a CLIP model trained on 400M pairs with a pre-training holdout, the effect size of the batch-mean loss at the trained temperature grows as $B^{0.64}$, whereas scoring the same pairs against fixed negatives restores $B^{0.50}$. The mechanism is near-tie amplification: the loss mean shift increases when a true pair barely beats its hardest negative; the gain over $\sqrt{B}$ therefore peaks at an intermediate number of negatives and then reverses. Fixed-negative measurements alone predict the in-batch curve up to batch size $65{,}536$ on three separately trained checkpoints and, in a pre-registered test, the optimal number of negatives and the detector ranking on a 40M-pair checkpoint of the same recipe. Because the signal lives in a pair's own similarity, auditors need no negatives beyond their reference set and no temperature: the expected crossing score obeys the square-root law and, when membership mainly affects poorly aligned pairs, outperforms the in-batch loss by up to 11 AUC points on 1,024 pairs and raises the true-positive rate at 1% false-positive rate by up to 38 percentage points (57% vs. 95%). Audits remain fragile: verdicts need thousands of pairs, can become uninterpretable when fewer than about a fifth of candidates are members, and can be manufactured by deleting 1.25% of the least-aligned non-member pairs. We recommend blind-model controls and pre-registered candidate sets.

1 INTRODUCTION

Membership inference asks whether a given example was used to train a model (Shokri et al., 2017; Yeom et al., 2018). For models trained on web-scale data, the evidence in any single example is weak: when members and non-members are drawn from the same distribution, common membership scores such as the training loss barely tell them apart (Duan et al., 2024; Das et al., 2025; Gomez, 2026). *Dataset inference* (Maini et al., 2021; Dziedzic et al., 2022) asks a coarser question that matters for provenance and copyright audits: was a given *set* of examples used for training? A common approach averages per-example scores so that their weak evidence accumulates (Maini et al., 2024; Puerto et al., 2025), and the standard justification is a central-limit argument: if members' scores differ from non-members' by Δ on average, with per-example standard deviation σ, the mean over m independent examples separates the two classes with effect size $d'(m) = \sqrt{m}\,\Delta/\sigma$. Taken at face value, this square-root law says any score, however weak, suffices given enough examples.

That promise rests on the assumption that an example's score does not depend on which other examples it is aggregated with. The assumption holds for a classifier's margin, a token likelihood, or an image–text pair's cosine similarity, but not for the InfoNCE loss (van den Oord et al., 2018) that trains CLIP-style dual encoders (Radford et al., 2021). InfoNCE scores each pair against the other pairs of its evaluation batch (its rivals), so a pair's loss depends on the batch's size and composition.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

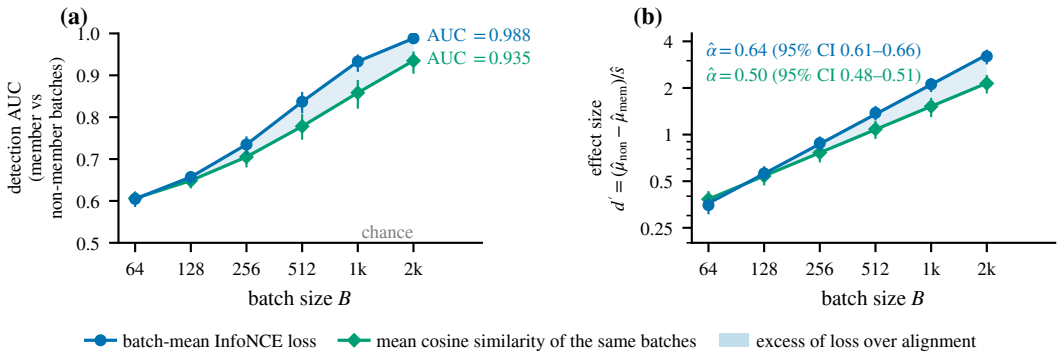


Figure 1: **With in-batch rivals, membership separability outgrows the square-root law.** Two set statistics of the same batches, formed by 50 random partitions of the member and non-member pools into disjoint batches of size B: the batch-mean InfoNCE loss at the trained logit scale $t = 100$, and the batch-mean cosine similarity of each image–text pair (*alignment*). **(a)** Detection AUC of member against non-member batches. **(b)** The same signal as effect size d' on log–log axes, where a power law is a straight line; lines are weighted least-squares fits over $B = 64$–2048, with slope $\hat{\alpha} = 0.50$ for alignment (the square-root law) and $\hat{\alpha} = 0.64$ for the loss. Error bars: 95% half-sample intervals.

Therefore, for the set-level score, *batch size is not sample size*: a larger batch averages more pairs, but also confronts every pair with more rivals.

Figure 1 shows the consequence on a CLIP model trained on 400M image–text pairs, with a member/non-member split drawn from the same distribution. A single pair's loss is uninformative (AUC 0.51). Averaged over batches, two statistics of the *same* batches diverge: the mean cosine similarity of each image–text pair, which ignores rivals, follows the square-root law ($\hat{\alpha} = 0.50$, 95% CI $[0.48, 0.51]$), while the batch-mean InfoNCE loss outgrows it ($\hat{\alpha} = 0.64$, 95% CI $[0.61, 0.66]$).

We explain this with a decomposition of set-level effect size (Section 3), $d'_{\text{set}}(m, N) \approx \sqrt{m}\,\delta(N)\,\text{DEFF}^{-1/2}$, where m is the number of pairs averaged, $\delta(N)$ is the per-pair effect size when each pair must beat N rivals, and DEFF is the design effect (Kish, 1965), the factor by which correlation among pairs of one set inflates the variance of their mean. The square-root law is the special case in which δ does not depend on N and $\text{DEFF} = 1$. Scoring every pair against a fixed *bank* of held-out pairs sets m and N independently and recovers the square-root law; in-batch scoring ties them together, $m = N + 1 = B$, so any growth of δ with N adds directly to the exponent.

We trace the membership signal to *near-ties*, pairs whose true partner is close to their strongest rival: membership raises a pair's own similarity slightly, and the loss responds far more strongly to such a change in near-ties than in pairs that win comfortably. Adding rivals at first turns more pairs into near-ties, and δ rises. Beyond a point, the pairs that carry most of the signal are already overtaken, extra rivals add variance but little signal, and δ falls. We call this *near-tie amplification* (Sections 3.1 and 4). It implies that the faster-than-$\sqrt{B}$ regime is transient, that in-batch detection has an optimal batch size, and that the same peak of $\delta(N)$ fixes the best size of a fixed bank.

Contributions. **(i) Mechanism** (Section 3): a decomposition of set-level effect size that explains when the square-root law holds and predicts the in-batch curve from fixed-bank measurements up to $B = 65{,}536$ on three checkpoints. **(ii) Where the signal lives** (Section 4): permutation nulls place the membership signal in the similarity of an image–text pair, not in its rivals; whether that shift sits in the low-alignment tail or is uniform decides the best number of rivals and which detector wins. **(iii) Detectors and audits** (Section 5): a rival-free lower-tail statistic follows the square-root law and beats the in-batch loss when the shift is tail-concentrated; we derive an audit's sample size and the confound floor below which its verdict is uninterpretable, and show that trimming 1.25% of a non-member set fabricates a positive verdict. **(iv) Prediction and replication** (Section 6): a pre-registered test on a 40M-pair checkpoint confirms the predicted bank size and detector ranking; the near-tie signatures recur on two separately trained DataComp checkpoints, and fail as predicted on a one-pass DataComp model whose shift is uniform.

Table 1: **Checkpoints.** All are ViT-B/16 dual encoders trained with InfoNCE, all but the blind model at batch size 8,192; n is the number of unique training pairs, κ the number of passes.

Checkpoint	n	Training data	κ	Evaluation pools	Role
A	400M	top 40% of CommonPool-L	32	pre-training holdout	audit target
40M, 4M rungs	40M, 4M	as A	32	A's pools	vary n alone
B	1.28×10^9	all of CommonPool-L	1.0	CommonPool-L / -XL	one-pass contrast
L	130M	LAION filter of CommonPool-L	9.8	filtered L / XL	replication
C	384M	top 30% of CommonPool-L	3.3	filtered L / XL	replication
Blind	400M	undisclosed (OpenAI CLIP)	32	saw no pool pair	split confound

2 SETUP

Threat model. An auditor wants to know whether a candidate set of T image–text pairs was used to train a dual encoder with image tower f and text tower g. The auditor queries both towers as a black box returning unit-norm embeddings, holds a reference set believed to be non-member, and has no access to weights, gradients, or training logs.

Models and data. Table 1 lists the seven checkpoints used in this work. *Checkpoint A* is the ViT-B/16 CLIP-MIA-Bench (Gomez, 2026) model, which follows the OpenAI CLIP training recipe (Radford et al., 2021) on a 400M CLIPScore-filtered CommonPool-L corpus (Gadre et al., 2023). Its member and non-member pools are drawn i.i.d. from one corpus, hence exchangeable by construction (details in Appendix D). Its two smaller *rungs* share its recipe and pools and vary only the number of unique training pairs n. *Checkpoint B* is the DataComp model `CommonPool.L-s1B-b8K` (Gadre et al., 2023). It resamples shards with replacement, so only a fraction $\pi_B \approx 0.63$ of its nominal members was seen, which biases its membership estimates downward (Appendix B). *Checkpoints L and C* share B's recipe and number of samples seen but train on the LAION-filtered (130M pairs) and CLIP-score-filtered (384M) subsets of CommonPool-L, hence make 9.8 and 3.3 passes; each is evaluated on unique pairs, with its own filter applied to both pools (Appendix C). A's corpus is a subset of B's, so the two are *separately* trained but not independent, and B saw both of A's pools at rate π_B, which makes it an exposure-balanced control on A's split (Section 5). A and B differ in passes, corpus filtering, non-member construction and logit scale ($t = 100$ against 85.4); no experiment here attributes their differences to one factor. The *blind* model, OpenAI CLIP (Radford et al., 2021), is assumed to have seen none of the pool pairs.

Admissible targets. Membership claims need **(C1)** an exchangeable split, both pools drawn i.i.d. from one corpus through one pipeline, so that a model blind to both separates them only up to a measurable confound, at **(C2)** web scale, $n \gtrsim 10^8$, because the per-pair similarity gap is bounded by $O(\kappa/n)$ (Gomez, 2026); **(C3)** the pass count κ is a covariate, not a gate. Of the public CLIP-family checkpoints with a constructible split, A, L and C satisfy all three (survey: Appendix A, Table 5): a filtered checkpoint meets C1 once its filter is applied to both pools. B meets C1 but makes one pass, so it enters as a contrast, not a replication.

Scores. Write $c_{ij} = \langle f(x_i), g(y_j) \rangle$ for image i and caption j. For a batch of B pairs from one class, the symmetric per-pair InfoNCE loss at logit scale t is

$$\ell_i(B,t) \;=\; -\frac{1}{2}\left[\log \frac{e^{tc_{ii}}}{\sum_{j=1}^{B} e^{tc_{ij}}} + \log \frac{e^{tc_{ii}}}{\sum_{j=1}^{B} e^{tc_{ji}}}\right], \tag{1}$$

and the batch statistic is its mean $\bar{\ell}_B$. The *in-batch collision rate* is the fraction of pairs beaten by another pair of the batch, averaged over the two retrieval directions, $\bar{C}_B = \frac{1}{B}\sum_i \frac{1}{2}\big(\mathbf{1}[\max_{j\neq i} c_{ij} > c_{ii}] + \mathbf{1}[\max_{j\neq i} c_{ji} > c_{ii}]\big)$. A *fixed bank* $\mathcal{K}$ is a set of N_0 held-out pairs (x'_k, y'_k), shared by all candidate pairs and independent of the candidate set. The *bank crossing score* of pair i is $s_i = \frac{1}{2}\big(\mathbf{1}[\max_k \langle f(x_i), g(y'_k)\rangle > c_{ii}] + \mathbf{1}[\max_k \langle f(x'_k), g(y_i)\rangle > c_{ii}]\big)$, that is, one minus Recall@1 averaged over the two retrieval directions. Unless stated otherwise the bank is *neutral*, drawn in equal numbers from the held-out member and non-member pools. The set statistic is $\bar{S}_T = \frac{1}{T}\sum_i s_i$. Finally, the *mean alignment* $\bar{A} = \frac{1}{m}\sum_i c_{ii}$ is the set-level form of the cosine-similarity attack of Ko et al. (2023). All scores are oriented so that larger values indicate non-membership.

Separability and protocol. For a statistic computed on disjoint member and non-member sets of size m, separability is the effect size $d' = (\hat{\mu}_{\text{non}} - \hat{\mu}_{\text{mem}})/\hat{s}_{\text{pooled}}$ across sets; we also report AUC, and under equal-variance Gaussian set statistics $\text{AUC} = \Phi(d'/\sqrt{2})$ (Appendix E). The scaling exponent α is the slope of $\log d'$ against $\log B$. We write $\hat{\alpha}$ for a weighted least-squares fit over $B \in \{64, \ldots, 2048\}$ (weights $1/\text{Var}_{\text{boot}} \log d'$) and $\alpha(B)$ for the local slope between adjacent batch sizes. Each class pool is 20 shards of the released data (16 for C); the remaining shards are held out as calibration data (non-member) and bank sources, so nothing evaluated is also a reference. Sets are formed by randomly partitioning each evaluation pool into disjoint batches at every B; point estimates average 50 independent partitions. Confidence intervals are 95% intervals from 500 half-sample replicates, each drawing half of each pool without replacement, re-partitioning, and recomputing everything, with paired differences formed within replicates; pairs are the resampling unit (details: Appendix D). The logit scale t is varied at evaluation time.

3 Batch size is not sample size

3.1 A decomposition of set-level effect size

Consider a per-pair score u_i computed against a set R_i of N rivals per retrieval direction, and a set statistic $\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i$ over m pairs of one class. We assume, and test in Section 4, that **(A1)** rivals carry no class information: the rival similarities $\{c_{ij}\}_{j \in R_i}$ have the same distribution for member and non-member anchors; and **(A2)** rivals are exchangeable: given c_{ii}, an anchor's rival similarities are i.i.d. draws from a common distribution.

Proposition 1 (Decomposition)**.** *Under (A1)–(A2), the class-conditional mean and variance of u_i depend only on the class and on N. Let $\delta(N) = (\mu_{\text{non}}(N) - \mu_{\text{mem}}(N))/\sigma(N)$ be the per-pair effect size and $\bar{\rho}$ the average correlation between the scores of distinct pairs in a set. If both classes share σ and $\bar{\rho}$, then*

$$d'_{\text{set}}(m, N) = \sqrt{m}\,\delta(N)/\sqrt{\text{DEFF}}, \qquad \text{DEFF} = 1 + (m-1)\bar{\rho}. \tag{2}$$

The proof is a short variance computation (Appendix F). For a *fixed bank*, $m = T$ and $N = N_0$ do not interact and pairs are scored independently given the bank ($\bar{\rho} = 0$), so $d'_{\text{set}} = \sqrt{T}\,\delta(N_0)$ exactly. For an *in-batch* statistic, $m = B$ and $N = B - 1$, so batch size moves both factors:

$$\alpha(B) = \frac{1}{2} + \frac{\mathrm{d}\log\delta}{\mathrm{d}\log N}\bigg|_{N=B-1} - \frac{1}{2}\,\frac{\mathrm{d}\log \text{DEFF}}{\mathrm{d}\log B}. \tag{3}$$

The square-root law is the special case of a flat δ and a constant design effect.

Proposition 1 needs a class shift, not a membership shift: its inputs are the distribution of c_{ii}, the rival law F and a small shift in c_{ii} between two labeled groups of anchors, so the mechanism claims of this section hold for any encoder scored with in-batch rivals. Where real training puts the shift is a property of a checkpoint and its split (Sections 4–6).

Near-tie model: why $\delta(N)$ rises and then falls. Take the crossing indicator in one retrieval direction, with rival similarities i.i.d. with distribution function F and density f. An anchor with positive similarity c is beaten by at least one of N rivals with probability $p_N(c) = 1 - F(c)^N$. If the members' law of c is the non-members' shifted by a small $\Delta c > 0$, then to first order

$$\mu_{\text{non}}(N) - \mu_{\text{mem}}(N) \approx \Delta c\, \mathbb{E}_c\big[N f(c) F(c)^{N-1}\big], \qquad \sigma^2(N) = \bar{p}_N(1 - \bar{p}_N), \tag{4}$$

with $\bar{p}_N = \mathbb{E}_c[p_N(c)]$ the mean crossing rate. The gap weights each anchor by $N f(c) F(c)^{N-1}$, the density of its strongest rival at c: the shift changes the outcome only for *near-ties*. While crossings are rare the gap grows as N and the standard deviation as $\sqrt{N}$, so $\delta \propto \sqrt{N}$ and $\alpha(B) \to 1$; once most pairs are beaten, extra rivals add variance but no signal, and δ peaks and falls. For identical anchors the peak sits at a crossing rate near 0.8; heterogeneity in c_{ii} moves it much lower, because a growing N saturates poorly aligned anchors before well-aligned ones come into contention. Evaluated with the empirical F and c_{ii}, the model reproduces the measured $\delta(N)$ to within 0.003, including its rise, peak and decline (full predictions vs. measurements in Appendix F, Table 15).

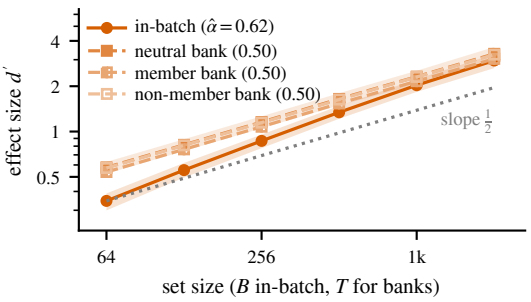


Figure 2: **Holding the rivals fixed restores the square-root law.** Checkpoint A: effect size of the crossing indicator against in-batch rivals (solid) and against neutral, member-only, and non-member-only fixed banks of $N_0 = 4096$ pairs (dashed); legend: fitted slopes over set sizes 64–2048; bands: 95% half-sample CIs; dotted: slope $\frac{1}{2}$. The fixed-bank loss control at every logit scale is in Appendix H, Figures 8 and 9.

3.2 Fixed rivals restore the square-root law

Two controls on checkpoint A test the prediction of Proposition 1 that separating the rivals from the averaging set restores $\alpha = \frac{1}{2}$ for any score (Figure 2). *Same statistic, different rivals:* the crossing indicator has $\hat{\alpha} = 0.62$ (95% CI $[0.59, 0.65]$) against in-batch rivals and $\hat{\alpha} = 0.50$ ($[0.49, 0.52]$) against a neutral fixed bank of $N_0 = 4096$ pairs, and likewise against member-only and non-member-only banks. The paired difference, 0.12 ($[0.08, 0.15]$), isolates the effect of the rival set. On unique pairs of checkpoints L and C the paired difference is 0.23 ($[0.20, 0.26]$) and 0.17 ($[0.08, 0.24]$), with fixed-bank slopes of 0.50 and 0.50 (Appendix C). *Same loss, fixed rivals:* scoring each pair's InfoNCE loss against a frozen class-matched bank and averaging gives slopes of 0.498–0.501 over three bank sizes and four logit scales, every CI containing $\frac{1}{2}$, whereas the in-batch loss slope rises with the logit scale, from 0.42 ($[0.40, 0.44]$) at $t = 20$ to 0.64 ($[0.61, 0.66]$) at $t = 100$. The loss is a soft crossing count that sharpens with t: its fixed-bank detectability slope changes sign between the two scales and, through Equation 3, reproduces the rise, overshooting the measured interval slightly at $t = 100$ (predicted 0.68; Appendix H); the design-effect term is negligible (Appendix G).

3.3 Rival-bank sweeps predict the in-batch curve

Equation 3 predicts the in-batch local slope from fixed-bank measurements alone. Across neutral banks of 16 to 65,536 held-out pairs, $\delta(N_0)$ rises, peaks at $N_0^\star = 4{,}096$ (95% CI 1,879–7,510), where the mean crossing rate is $\bar{p}^\star = 0.10$ (CI 0.07–0.13), and falls (Figure 3a); the optimum targets near-ties, so it sits far below the $\frac{1}{2}$ that maximizes one Bernoulli trial's information. The prediction agrees with the measured local slopes, each estimator compared with its own prediction

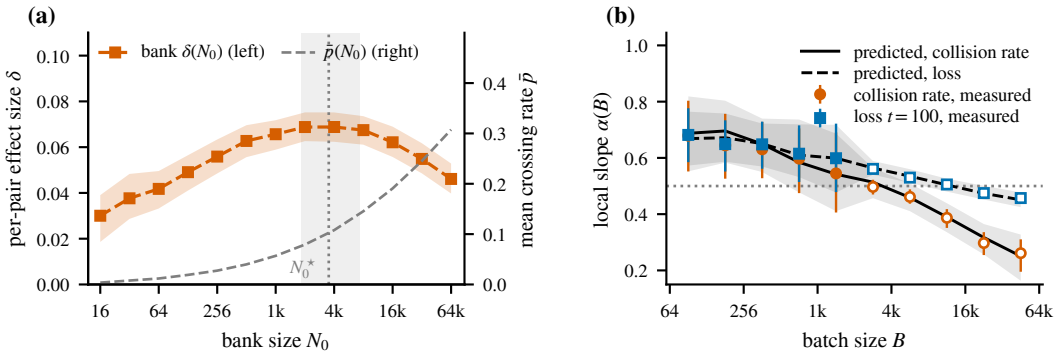


Figure 3: **Fixed-bank measurements predict the in-batch scaling curve.** Checkpoint A. **(a)** Per-pair effect size $\delta(N_0)$ of the crossing score against neutral banks (left axis; 95% CI band) and mean crossing rate $\bar{p}$ (right axis, dashed); dotted: $N_0^\star$, 95% CI. **(b)** Local slopes $\alpha(B)$ of the in-batch collision rate (circles) and loss at $t = 100$ (squares), 95% CIs; filled: direct estimator ($B \leq 2048$), open: per-pair estimator (Appendix J). Black: predictions from (a) through Equation 3 (collision solid, loss dashed; 95% CI band), with the design-effect term for the direct estimator and without it for the per-pair estimator, which omits it by construction; dotted: $\frac{1}{2}$.

(Figure 3b): over the ten intervals the mean absolute deviation is 0.01 for the collision rate and 0.01 for the loss at $t = 100$.

All ten measured intervals contain the prediction for both statistics, and a nominal χ^2 test against it does not reject ($\chi^2 = 3.7$ for the collision rate and 1.5 for the loss on ten intervals, treated as independent although adjacent local slopes share an endpoint), while the flat null $\alpha(B) = \frac{1}{2}$ is rejected decisively ($\chi^2 = 238.1$ and 154.2). It also predicts the local slopes of checkpoints L and C on unique pairs (mean absolute deviation 0.01 and 0.02 for the collision rate, 0.02 and 0.02 for the loss), where the flat null is again rejected ($\chi^2 = 576.6$ and 45.7), and those of checkpoint B (0.04 and 0.02), where it is not, because B's $\delta(N)$ is nearly flat (Appendices J and C). Per-pair efficiency $d'/\sqrt{B}$ therefore peaks where $\alpha(B)$ crosses $\frac{1}{2}$, which by Equation 3 is the peak of δ: at $B \approx 2{,}784$ for the collision rate, inside the interval of $N_0^\star$, and at $B \approx 13{,}127$ for the loss (Appendix J): the largest available batch is not the best one.

4 What carries the membership signal

A marginal shift, not relational structure. We test (A1) and (A2) on the real similarity matrix with permutation nulls that preserve every marginal distribution (Table 2). The *diagonal shuffle* permutes positive similarities among a batch's anchors, breaking the link between an anchor and its own rivals; the *negative swap* scores member anchors against non-member rivals and vice versa, which would reverse any signal carried by the negatives. Both reproduce the real slopes and most of the effect size (Appendix K). The negative swap's 95% intervals contain 1 on all three checkpoints at the trained logit scale: there, the negatives carry no class information. The diagonal shuffle keeps 0.93 of the loss effect size on A and 0.84 ($[0.79, 0.89]$) and 0.91 on L and C, so (A2) holds only approximately: a smaller part of the signal depends on how a pair's positive sits among its own rivals. The rest is the class shift in positive similarity, $\Delta c = 2.2 \times 10^{-3}$ on A, or 0.048 standard deviations of c_{ii}.

Table 2: **Permutation nulls** on the real similarity matrix: InfoNCE loss at each checkpoint's trained logit scale (L and C on unique pairs). Each null preserves all marginals; entries are effect sizes at $B = 2048$ relative to the real matrix, paired within replicates, with 95% half-sample intervals. A null removing the signal would give ratios near 0. Slopes and the collision rate: Appendix K.

		d'/d'_{real}, InfoNCE loss		
Condition	What it destroys	A	L	C
Diagonal shuffle	anchor–rival link	0.93 $[0.87, 0.99]$	0.84 $[0.79, 0.89]$	0.91 $[0.82, 1.00]$
Negative swap	class of the rivals	1.04 $[0.97, 1.10]$	1.01 $[0.95, 1.06]$	1.00 $[0.91, 1.10]$

The signal lives at near-ties. The quantile shift function $\Delta(q) = Q_{\text{mem}}(q) - Q_{\text{non}}(q)$ of c_{ii} (Figure 4a) is $2.91\times$ the mean shift in the lowest quintile of alignment on checkpoint A and $0.22\times$ ($[0.01, 0.41]$) over the top fifth; the difference between the low and high bands, 2.49 ($[2.10, 3.00]$ in units of the mean shift), excludes zero. On checkpoints L and C the same contrast is 2.32 ($[2.05, 2.63]$) and 2.34 ($[1.65, 3.17]$); on checkpoint B it is 0.16 ($[-0.23, 0.52]$): B's shift is uniform. The loss reads the same tail: the top decile of per-pair losses in each batch retains 97–99% of the full-batch effect size at every B on checkpoint A and 91% and 90% on L and C at $B = 2048$, against 53% on checkpoint B, and widening the crossing criterion to near-misses dilutes it (Appendix M).

Why counting crossings can beat averaging. Under (A1)–(A2) the crossing score reads c_{ii} through $\mathbb{E}[s_i \mid c_{ii}] = 1 - F(c_{ii})^N$, a soft indicator of the quantiles of c_{ii} near the crossing rate, so under a uniform shift it does not beat the mean: with A's *empirical* c_{ii} and rival law the best it reaches is $0.87\times$ mean alignment. On checkpoint A the shift is not uniform, and a lower-tail statistic is a matched filter: the bank crossing score achieves $1.52\times$ the effect size of mean alignment at every batch size (Figure 4b), and the near-tie model driven by the measured $\Delta(q)$ reproduces that ratio (1.53) and its optimal crossing rate (0.08). On L and C it reaches $1.37\times$ and $1.28\times$ mean alignment at $T = 2048$, and the model gives 1.38 and 1.33. On checkpoint B the shift is flat and the ranking reverses (0.81 at $B = 2048$), as the model predicts. The advantage survives single-class banks and a permutation of c_{ii} across candidates (Appendices I and N).

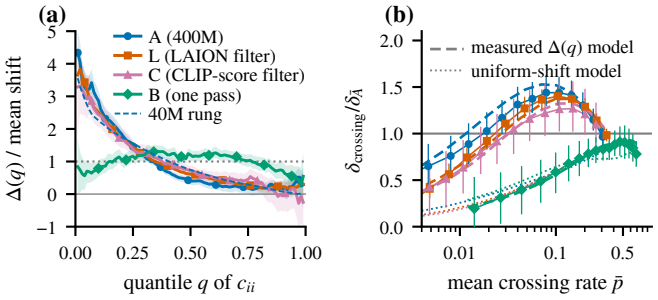


Figure 4: **The membership shift sits in the low-alignment tail on A, L and C and is uniform on B** (95% pair-bootstrap intervals). **(a)** Quantile shift function $\Delta(q)$ of c_{ii} in units of each checkpoint's mean shift. **(b)** Per-pair effect size of the bank crossing score relative to mean alignment's, with the near-tie model

Score	Rivals	AUC A	AUC L	AUC C
Mean alignment $\bar{A}$ (Ko et al., 2023)	none	0.858	0.952	0.749
Rank score against the reference	none	0.808	0.938	0.738
Lower-tail mean, $q = 0.2$	none	0.977	0.993	0.830
Expected crossing ψ_N, $N = 4096$	none	0.971	0.997	0.854
Crossing score $\bar{S}_T$	fixed	0.942	0.989	0.811
InfoNCE loss (trained t)	fixed	0.964	0.986	0.793
InfoNCE loss (trained t)	in-batch	0.933	0.951	0.747
Collision rate $\bar{C}_B$	in-batch	0.927	0.977	0.766

Table 3: **Detector comparison** ($T = 1024$). AUC of member versus non-member sets; all scores use the same sets and, per checkpoint, one neutral bank of $N_0 = 4096$ pairs; Q_{ref} comes from held-out non-members. For A, the rank, lower-tail, ψ_N and collision rows come from a 5-partition run (Appendix D). TPR at 1% FPR and $T = 256$: Appendix N, Table 22.

5 A FAMILY OF SQUARE-ROOT-LAW DETECTORS AND THEIR USE IN AUDITS

The decomposition suggests a family of detectors that obey the square-root law by construction, need only embeddings, and win over the in-batch loss when the shift is tail-concentrated, as it is on checkpoints A, L and C: score every candidate pair on its own, against a fixed bank sized near the peak of δ or a quantile of held-out non-members, and average. With per-pair scores s_i of T candidates and $\tilde{s}_r$ of R references believed to be non-member (for bank scores, against the same bank), the test statistic and its one-sided p-value are

$$z = \frac{\bar{s} - \bar{\tilde{s}}}{\sqrt{\hat{\sigma}_s^2/T + \hat{\sigma}_{\tilde{s}}^2/R}}, \qquad p = 1 - \Phi(z). \tag{5}$$

Separability. On every checkpoint each fixed-rival statistic beats its in-batch counterpart (Table 3), and the strongest scores use no rivals and no logit scale at all: the lower-tail mean of alignment, $-\min(c_{ii}, Q_{\text{ref}}(0.2))$, and the expected crossing score $\psi_N(c_{ii}) = 1 - F(c_{ii})^N$ lead on A, L and C, ahead of the fixed-bank InfoNCE loss, the only one of the three that needs a logit scale. The bank crossing score, the Monte-Carlo version of ψ_N, follows the square-root law exactly ($\hat{\alpha} = 0.50$) with per-pair effect size $\delta = 0.069$; averaging out the bank draw is worth $1.16\times$. The ranking is unchanged over $q \in [0.05, 0.2]$, so q is not tuned. Which quantiles of c_{ii} a score reads matters more than its form: on checkpoint A every score weighting the lowest fifth beats mean alignment, and on checkpoint B, whose shift is uniform, the order reverses: scores that read the whole distribution lead (rank score, AUC 0.834 at $T = 2048$), the lower-tail mean falls to 0.706, and no score separates significantly better than mean alignment (fifteen scores on all checkpoints: Appendix N, Tabs 2627).

Operating point. Detection is best near $N_0^\star = 4{,}096$, where about 10% of pairs are beaten (Section 3.3), but the optimum sits where the shift is, which an auditor does not know in advance: on checkpoint B a bank sized for $\bar{p} = 0.1$ keeps only 0.59 of mean alignment against 0.91 at its own optimum. We therefore recommend reporting the test at two operating points, $\bar{p} \approx 0.1$ and $\bar{p} \approx \frac{1}{2}$ (for the rival-free scores, quantiles $q \approx 0.2$ and $q \approx 0.5$), with a Bonferroni correction.

Sample size and calibration. A fixed bank makes pairs independent, so a candidate set with member fraction π gives $d'_{\text{set}} = \pi\delta\sqrt{T}$, and a one-sided test at level α with power $1 - \beta$ needs $T = (1 + T/R)\,((z_{1-\alpha} + z_{1-\beta})/\pi\delta)^2$ pairs against a reference set of R pairs. With $\delta = 0.069$, $\alpha = 10^{-3}$ and 80% power that is about 3,200 pairs against a large reference set, 6,500 with $R = T$, 13,000 if half the candidates are members and 3.2×10^5 if one in ten is. These figures are conservative: the

lower-tail mean ($\delta = 0.086$) needs about 2,100 pairs for the same power, and accepting 50% power lowers the first figure to about 2,000. A positive result identifies $\pi\delta$, not π; the formula is general, and every figure here is checkpoint A's. Drawing candidates and references at random from one non-member pool, the test rejects at the nominal rate at every set size and its power matches the formula. With a fixed halving of the pool instead, the small offset between the halves inflates the nominal 1% type-I error to 3.3% for the crossing score and 9.1% for the in-batch loss at $T = R = 16{,}384$: the reference set, not the statistic, is what breaks an audit (Appendix N, Table 25).

How much of a verdict is the data split? A model that saw neither pool should not separate them. The OpenAI model separates checkpoint A's pools at signed $d' = +0.36$ (95% CI $[+0.08, +0.66]$), and checkpoint B, which trained on the corpus both of A's pools come from and is exposure-balanced on them, at $+0.70$ ($[+0.42, +0.99]$) through alignment alone (its rival term is $+0.03$ $[-0.26, +0.32]$; Table 4); we take the larger figure as the confound for A's split, which the membership signal exceeds by a factor of 4.5. The two models see a shift on A's pools of 0.13 and 0.40 of A's own, tail-concentrated in the control, so lower-tail detectors inherit the confound about as the mean does (Appendix B). Per pair, a blind separation of d' at set size T is a confound of $\epsilon = d'/\sqrt{T} = 0.0155$ standard deviations, so with the crossing score's $\delta = 0.069$ a verdict is interpretable only above $\pi_{\min} = \epsilon/\delta = 0.22$ (0.32 at the upper end of its interval); the $\pi = 0.1$ scenario above lies below that floor, which is checkpoint A's: another audit must re-measure ϵ and δ. On the filter-matched splits of L and C the blind model is within noise of zero ($d' = -0.07$ and -0.07). At the upper end of the blind interval the floor is 0.08 on L and 0.23 on C, whose weaker signal ($\delta = 0.037$) raises it although its split is as clean as L's.

Adversarial assembly. Selecting on alignment is observationally identical to membership: removing the lowest-aligned 1.25% of a wholly non-member candidate set buys a verdict equivalent to $\pi \approx \frac{1}{2}$ at $T = 10^4$ with only the embedding access the audit requires (a Gaussian calculation; only the conversion to a member fraction uses checkpoint A's δ). External audits therefore need candidate and reference sets fixed by provenance, with any cleaning rule, and registered before any model output is inspected, plus a blind-model control (Appendix O).

6 Generality and a pre-registered test

Two filter-matched replications. On unique pairs of checkpoints L and C every signature of Sections 3 and 4 recurs (Table 4): in-batch rivals raise the slopes of the loss and the crossing indicator above their fixed-bank value of $\frac{1}{2}$, the fixed-bank sweep predicts the in-batch curve for both statistics, the shift sits in the low-alignment tail, the negative swap leaves the signal intact, and the blind model sits at zero. The 2026 downloads of their pools contain placeholder images for dead links: 5.1% and 3.4% of pairs have a duplicated image or caption, and on all pairs these depress every loss-based statistic while leaving the bounded crossing and threshold scores nearly intact (Appendix C).

A one-pass contrast. On checkpoint B the near-tie signatures fail as predicted for its uniform shift: the crossing-indicator difference, 0.04 $[-0.02, 0.10]$, includes zero, and the near-tie model bounds any crossing score at about $0.80\times$ mean alignment, while the loss keeps a small excess (0.07 [0.03, 0.10]). Its estimates are biased down: only $\pi_B \approx 0.63$ of its nominal members were seen, and its split leans against membership in the rival term (blind $d' = -0.61$; Appendix B).

A pre-registered test. Before measuring any ranking quantity on A's 40M rung, which shares A's recipe and pools with ten times fewer pairs, we predicted them from its label-free inputs and A's shift shape, and committed each prediction with an agreement criterion (Appendix P). The bank optimum (predicted 3,989, measured 6,109), its crossing rate, the in-batch loss slope and the shift contrast pass, and the eight detectors of Table 3 come out in the predicted order (Kendall's $\tau_K = 0.93$ and 0.86); the near-tie gain misses (1.32 against 1.60) through a residual shape difference, and the 4M rung, outside the small-shift regime, fails where the register said it would.

Table 4: **Core measurements on four checkpoints** (L and C on unique pairs). Canonical estimator and fit window (Section 2); differences within half-sample replicates, 95% CIs; detector: bank crossing score, AUC at $T = 2048$; blind rows: OpenAI CLIP, which saw no pool, signed $d' > 0$ in the membership direction, at $T = 2048$. Control row: checkpoint B on A's pools; having seen L's and C's members only, B is no control there. Rungs: Appendix P.

Quantity	A	L	C	B
$\hat{\alpha}$, InfoNCE loss, in-batch	0.64	0.83	0.73	0.57
$\hat{\alpha}$, crossing indicator, in-batch	0.62	0.73	0.67	0.54
$\hat{\alpha}$, crossing indicator, fixed bank	0.50	0.50	0.50	0.50
In-batch − bank (loss)	0.14 [0.11, 0.16]	0.33 [0.28, 0.38]	0.23 [0.11, 0.30]	0.07 [0.03, 0.10]
In-batch − bank (crossing)	0.12 [0.08, 0.15]	0.23 [0.20, 0.26]	0.17 [0.08, 0.24]	0.04 [−0.02, 0.10]
Detector AUC	0.986	0.999	0.889	0.796
Negative swap, d'/d'_{real}	1.04	1.01	1.00	0.95
Blind, full detector, signed d'	+0.36 [+0.08, +0.66]	−0.07 [−0.39, +0.23]	−0.07 [−0.38, +0.27]	+0.06 [−0.22, +0.34]
Blind, rival term only, signed d'	+0.05 [−0.24, +0.35]	−0.29 [−0.57, −0.01]	−0.38 [−0.68, −0.10]	−0.61 [−0.89, −0.35]
Control: B on A's pools	+0.70 [+0.42, +0.99]	—	—	—

7 Related work

Membership inference on foundation models. Loss- and confidence-based attacks (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022; Bertran et al., 2023) lose most of their power at web scale (Duan et al., 2024). Contrastive-encoder attacks threshold the similarity of augmented views (Liu et al., 2021; He & Zhang, 2021) or of an image–text pair (Ko et al., 2023); captioning and generative models have their own (Hu et al., 2022; Li et al., 2024), and CLIP's memorization has been probed through captions (Wang et al., 2025; Jayaraman et al., 2024) and identities (Hintersdorf et al., 2024). Under randomized splits, per-example similarity attacks on CLIP fall to near chance (Gomez, 2026); post-hoc splits are often dominated by distribution shift (Duan et al., 2024), so model-free baselines match (Meeus et al., 2025) or beat (Das et al., 2025) published attacks.

Aggregating weak signals over sets. Dataset inference tests a suspect set with aggregated per-example scores, for classifiers (Maini et al., 2021), encoders (Dziedzic et al., 2022) and language models (Maini et al., 2024); related threat models ask whether a user's or a document's data was used (Kandpal et al., 2024; Meeus et al., 2024), how evidence accumulates over documents (Puerto et al., 2025), or audit differential privacy with canaries (Steinke et al., 2023). All aggregate a per-example statistic that does not depend on what else is aggregated, and so inherit the square-root law; we study statistics for which that premise fails. Near-tie concentration parallels pretraining-data detectors that focus on a language model's least likely tokens (Shi et al., 2024; Zhang et al., 2025).

Batch size and negatives in contrastive learning. InfoNCE (van den Oord et al., 2018) bounds mutual information by the log of the number of candidates (Poole et al., 2019), one reason contrastive pretraining benefits from large batches (Chen et al., 2020; Radford et al., 2021; Cherti et al., 2023); the batch-size dependence of its gradient is a bias that decomposable or global objectives remove (Chen et al., 2022; Yuan et al., 2022), and the sigmoid loss drops the batch-wise normalization (Zhai et al., 2023). Alignment and uniformity (Wang & Isola, 2020) describe the geometry contrastive training produces; membership lives in the alignment term here.

8 Limitations and conclusion

Limitations. Every checkpoint here is a ViT dual encoder trained with InfoNCE at batch size 8,192, so whether the best number of rivals tracks the training batch size is untested. The tail-concentrated checkpoints make several passes over filtered corpora and B one pass over an unfiltered one, so which sets the shape of the shift is open. A sigmoid-trained encoder (Zhai et al., 2023) keeps its rivals under in-batch InfoNCE or Recall@1, but no admissible split exists for one (Appendix A). No detector is uniformly best: lower-tail statistics win on a tail-concentrated shift, mean alignment on a uniform one. Audits need thousands of pairs, and a null result cannot distinguish non-membership from training without overfitting. **Conclusion.** The auditor's batch size sets both how many examples a test averages and how hard each comparison is; separating the two explains when set-level leakage outpaces averaging, how many rivals to use, and which detector wins.

ETHICS STATEMENT

This work studies how much a contrastive model's embeddings reveal about whether a set of image–text pairs was used in training. The capability is dual-use: it can help rights holders, auditors, and regulators substantiate claims about training data, and it can help third parties probe a model's training set. Three properties limit the second use. The detector is uninformative about individual pairs (AUC about 0.52 for a single pair scored against a fixed bank) and needs blocks of thousands of pairs, so it does not expose individuals. We quantify how confounded reference sets and adversarial trimming produce false verdicts, and we state the protocol an external audit requires. All experiments use publicly released models and datasets, and no new personal data were collected. Model owners can reduce exposure by deduplicating training data, limiting the number of passes over it, or not exposing embeddings.

AI USE STATEMENT

We disclose our use of generative AI tools following the ICLR 2027 policy. AI tools served as assistants in discussions that helped develop the conceptual framework of our work, refine hypotheses, give feedback on the design of experiments, and interpret results, and they helped implement the analysis code and figure scripts. They also helped draft and edit parts of the text, suggest its structure, and identify relevant literature. They were not used to generate synthetic data, to provide critical ingredients for or write proofs, or to clean or reformat datasets; translation and qualitative data analysis do not apply. The research question, the initial finding that in-batch separability grows faster than $\sqrt{B}$, the use of fixed banks and count crossings, the initial base code for dataset inference, the choice of models and splits, and the decision of which experiments to run were the authors'. Every AI contribution was checked by the authors: we re-derived the mathematical statements, regenerated every reported number, table and figure from the results files with the analysis code and compared them against the text, and checked every reference against its publisher record. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix D specify the models, data splits, batch construction, estimators, fit windows, and the half-sample interval protocol; Appendices F, G, and I give all derivations. All checkpoints and datasets are public. The supplementary material contains anonymized code and the results' files.

REFERENCES

Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie Morgenstern, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression. In *Advances in Neural Information Processing Systems 36*, pp. 314–330. Neural Information Processing Systems Foundation, Inc., 2023.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33860–33875. Curran Associates, Inc., 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 118–125. IEEE, 2025.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? arXiv:2402.07841, 2024. Conference on Language Modeling (COLM) 2024.

Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, and Nicolas Papernot. Dataset inference for self-supervised models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 12058–12070. Curran Associates, Inc., 2022.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27092–27112. Curran Associates, Inc., 2023. Datasets and Benchmarks Track.

Lluis Gomez. Rethinking membership inference attacks for CLIP. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):21305–21313, 2026.

Xinlei He and Yang Zhang. Quantifying and mitigating privacy risks of contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 845–863. ACM, 2021.

Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Does CLIP know my face? *Journal of Artificial Intelligence Research*, 80: 1033–1062, 2024.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M^4I: Multi-modal models membership inference. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1867–1882. Curran Associates, Inc., 2022.

Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision–language models. In *Advances in Neural Information Processing Systems 37*, pp. 50722–50749. Neural Information Processing Systems Foundation, Inc., 2024.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, and Zheng Xu. User inference attacks on large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18238–18265, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965. Reprinted 1995 in the Wiley Classics Library, ISBN 978-0-471-10949-5.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4848–4858. IEEE, 2023.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98645–98674. Curran Associates, Inc., 2024.

Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. EncoderMI: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2081–2095. ACM, 2021.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. arXiv:2104.10706, 2021. International Conference on Learning Representations (ICLR) 2021.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 124069–124092. Curran Associates, Inc., 2024.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2369–2385, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. SoK: Membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401. IEEE, 2025.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019.

Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4165–4182, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400m: Open dataset of CLIP-filtered 400 million image-text pairs. arXiv:2111.02114, 2021.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, pp. 51826–51843, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49268–49280. Curran Associates, Inc., 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Wenhao Wang, Adam Dziedzic, Grace C. Kim, Michael Backes, and Franziska Boenisch. Captured by captions: On memorization and its mitigation in CLIP models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. arXiv:2309.16671, 2024. International Conference on Learning Representations (ICLR) 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25760–25782. PMLR, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-K%++: Improved baseline for pre-training data detection from large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, pp. 64845–64862, 2025.

A Admissible targets: conditions and public-checkpoint survey

This appendix gives the quantitative basis for the three conditions of Section 2 and applies them to every public CLIP-family checkpoint we considered (Table 5). Throughout, κ is the number of passes a checkpoint makes over its training pool (training samples seen divided by pool size) and n is the number of unique pairs in that pool; neither is the batch size B or the rival count N of the main text.

C1: an exchangeable split. Condition C1 separates membership inference from out-of-distribution detection. Gomez (2026) measure the size of the violation under the standard cross-dataset protocol, with LAION-400M pairs as members and CC3M, CC12M and MSCOCO pairs as non-members: a bag-of-words classifier over captions, which never queries the model, reaches AUC 98.4, and Monte-Carlo dropout, an out-of-distribution detector, reaches 98.3–99.5; on a split that satisfies C1 the same attacks fall to 50.1–52.6. In the units of this paper a split artifact enters as a per-pair confound ϵ, and a verdict is interpretable only above the member fraction $\pi_{\min} = \epsilon/\delta$ (Section 5); a cross-dataset split puts ϵ above δ and leaves no interpretable region. Checkpoint B

satisfies it: DataComp's file-level nesting places its non-member files outside CommonPool-L (Appendix B), and a blind model separates its pools only through a small lean of the rival term against membership. A filtered checkpoint violates C1 against unfiltered non-members, whose alignment is lower than its members' by construction; applying its filter to both pools restores C1, which is how checkpoints L and C are evaluated (Appendix C).

C2: web scale. Condition C2 places the target in the regime a provenance audit faces. Gomez (2026) bound the per-pair similarity gap between members and non-members by $O(\kappa/n)$ and confirm the law across the three CLIP-MIA-Bench checkpoints, trained on 4M, 40M and 400M pairs with every other hyperparameter fixed: the fitted log–log slope is 0.99 ± 0.02 with $R^2 = 0.997$, and the per-pair AUC of the cosine-similarity attack falls from 94.1 to 62.5 to 51.1. The per-pair effect sizes this paper reports on checkpoint A (Section 4) are those of the 400M model; on the 4M model the same statistic separates most pairs individually, outside the small-shift regime in which Proposition 1 and the near-tie model are used. A checkpoint trained on 10^6–10^7 pairs is therefore not a weaker instance of the target regime but a different one, tens to hundreds of times larger in its per-pair gap, and results on it do not transfer to the audit setting.

C3: passes, as a covariate. Condition C3 is weaker than C1 and C2 and is applied differently. The $O(\kappa/n)$ law is linear and smooth in κ: it predicts a proportionally smaller gap for a one-pass checkpoint, not a qualitative boundary, so a one-pass checkpoint is placed on the κ/n axis and compared there rather than excluded (Figure 5). Measured, the law does not control the gap across the two checkpoints of this paper: with $\kappa/n = 8.0 \times 10^{-8}$ for A and 7.8×10^{-10} for B it predicts a raw mean-alignment gap $102.4\times$ smaller on B, whereas the measured gaps are $\Delta c = 2.2 \times 10^{-3}$ on A and 1.6×10^{-3} $[1.3 \times 10^{-3}, 2.0 \times 10^{-3}]$ on B, a ratio of 1.37 $[0.98, 1.66]$ that contains one and excludes both the pass ratio of 32 and the law's prediction; after exposure correction by π_B the ratio is 0.86 $[0.62, 1.04]$ and the standardized ratio is 1.00 $[0.72, 1.22]$. The two checkpoints differ in four ways at once, so this is a statement about the pair rather than about the law, and it fixes how C3 is used: passes are a covariate to report, not a gate, and the A↔B comparison is a contrast of profile and never of amount. Checkpoint B, at $\kappa = 1.0$ with shards resampled, is the case in point: a fraction $\pi_B = 0.63$ of its nominal members was seen at least once, and Appendix B propagates that fraction through every set-level effect. A stronger reading of C3 is a hypothesis this paper raises and does not establish: that a single gradient visit per pair changes the *shape* of the shift rather than its size, that is, whether it concentrates in the low-alignment tail. The A↔B contrast (low-minus-high contrast 2.49 against 0.16; Section 4) is consistent with it, as are checkpoints L and C, which make 9.8and 3.3passes and share A's shape; but A and B differ in four ways at once: passes (32 against 1.0), corpus (the CLIPScore-filtered top 40% of CommonPool-L against all of it), non-member construction (a pre-training holdout against a disjoint superset), and trained logit scale ($t = 100$ against 85.4). No public checkpoint isolates passes: every unfiltered CommonPool baseline makes one pass, and the filtered baselines that reach about nine passes have filter-defined splits (Table 5). We therefore report κ as a covariate and leave the shape hypothesis open.

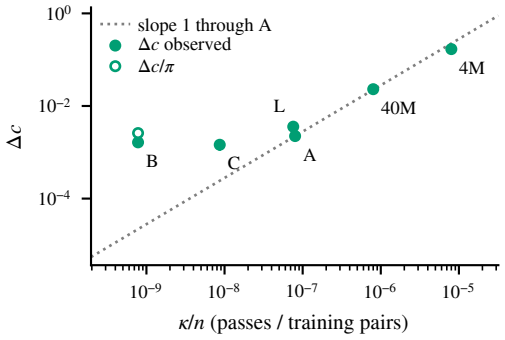


Figure 5: **Six checkpoints on one n axis.** Class gap Δc against passes per training pair κ/n for A, its 40M and 4M rungs, B, L and C (L and C on unique pairs), log–log, 95% half-sample intervals; dotted grey: slope 1 through A; open circle: B's gap corrected for exposure, $\Delta c/\pi_B$. The rungs and L lie on or near the line; B and C lie well above it.

One recipe, three pass counts. Checkpoints B, C and L share one recipe, batch size and number of samples seen, so across them κ/n grows about a hundredfold, from 7.8×10^{-10} through 8.7×10^{-9} to 7.6×10^{-8} (Figure 5). If the gap were proportional to κ/n within this family, L's would be about a hundred times B's. It is about twice as large (3.6×10^{-3} against 1.6×10^{-3}, or 2.6×10^{-3} after

B's exposure correction), and C's, 1.5×10^{-3}, matches B's within the intervals. The comparison is not a clean test of the law: each checkpoint is evaluated on the pairs that pass its own filter, and filtering changes the distribution of c_{ii} on which the gap is measured; in units of that distribution's spread the gaps are 0.030, 0.029 and 0.072. What it does show is a contrast between two ways of raising κ/n: shrinking the corpus at random with passes fixed, as along A's ladder, raises the gap in proportion, whereas shrinking it by filtering with the training budget fixed, as from B to L, raises it far less.

How the conditions bind. C1 and C2 are exclusion criteria for the membership claims only. Where real memorization puts the shift (Sections 4 and 6) requires a genuine split at scale, and the audit quantities of Section 5, the sample sizes, the confound floor $\pi_{\min}$ and the trimming result, are properties of each split at its scale: checkpoint A's in Section 5, with L's and C's confound floors beside it. The mechanism of Section 3 takes three inputs, a distribution of positive similarities c_{ii}, a rival law F and a class shift, and nothing in Proposition 1 or the near-tie model requires the shift to be a membership shift, so those claims are not constrained by any of the three conditions.

Survey. Table 5 lists every public CLIP-family checkpoint we considered. Of those with a constructible member/non-member split, three satisfy all three conditions: checkpoint A, and checkpoints L and C once each is evaluated on a filter-matched split; every other checkpoint fails a condition the table names. Four entries deserve a sentence. The DataComp filtered baselines (Gadre et al., 2023) are the only public checkpoints that make several passes at scale. Against unfiltered non-members their members are higher-alignment by construction and C1 fails; for the LAION and CLIP-score baselines, whose filters are functions of the released metadata, a filter-matched split restores it (Appendix C). `DataComp.L-s1B-b8K` adds an image-based filter that we did not reproduce. `DataComp.XL-s13B-b90K`, the only public route to a different training batch size ($90{,}112$) at scale, trained on a filtered subset of the largest pool, so every pair that passes its filter is a member and no filter-matched split exists. MetaCLIP (Xu et al., 2024) is the near miss: its balancing step subsamples over-represented metadata entries at random, so an exchangeable split between selected and rejected pairs exists in principle, but the rejected pairs were never downloaded, and the release consists of the curation code and the metadata distribution rather than the excluded data. The two-model design of Jayaraman et al. (2024), which trains separate CLIP models on disjoint i.i.d. halves of one corpus, is the cleanest existing instance of the split construction C1 asks for, at about 40M pairs; its checkpoints are not public, so it does not appear in the table. SigLIP (Zhai et al., 2023) is trained on WebLI, which is not public, so no membership labels exist for it; Section 8 states what the sigmoid objective does and does not change under Proposition 1. What would extend the survey is stated by the conditions themselves: a checkpoint trained at a different batch size on a pretraining holdout of an exchangeable corpus of at least 10^8 pairs, which no public release provides and which is outside the compute of this work.

Other objectives (the sigmoid corollary). The sigmoid loss (Zhai et al., 2023) separates the training objective from the auditor's statistic. As the statistic, $-\log \sigma(t\, c_{ii} - b)$ has no normalization over candidates: it is rival-free, $\delta(N)$ is flat by definition, and Proposition 1 gives $\alpha = \frac{1}{2}$ exactly, the fixed-bank result of Section 3.2 with a different loss. Scored with in-batch InfoNCE or Recall@1, as a retrieval harness does, a sigmoid-trained encoder keeps its rivals and the excess slope should persist; sigmoid training may move $\Delta(q)$, not the N-dependence of $\delta(N)$. We evaluate no such checkpoint as a target because WebLI is not public, so no admissible split exists (Section 2).

B CHECKPOINT B: SPLIT CONSTRUCTION AND MEMBERSHIP DILUTION

Split. Checkpoint B is the unfiltered large-scale DataComp baseline (Gadre et al., 2023), trained on CommonPool-L. Members are drawn from CommonPool-L and non-members from CommonPool-XL; both are downloaded with the DataComp tooling as shards of $10{,}000$ image–caption pairs with their metadata. The member pool is CommonPool-L shards 0–19 and the non-member pool is CommonPool-XL shards 0–19, that is, $200{,}000$ pairs per class, on which every statistic is computed; shards 20–24 of each pool ($50{,}000$ pairs per class) are held out and serve only as bank sources and as the calibration reference. Nothing evaluated is ever also a reference. The pools are used as downloaded, without deduplication: the two pools share no pair identifier and $2{,}995$ of their $500{,}000$ image hashes, and about 8% of the pairs in each pool share an image hash with another pair of the

Table 5: **Public CLIP-family checkpoints against the three admissibility conditions** of Section 2. Unique pairs, passes κ and training batch size B_{train} are as stated in each release or its paper (Gomez, 2026; Gadre et al., 2023; Radford et al., 2021; Schuhmann et al., 2021; Cherti et al., 2023; Xu et al., 2024; Zhai et al., 2023); the checkpoint-A and checkpoint-B values are those of Section 2. Legend: ✓ satisfied; $\sim$ approximately satisfied, quantified in the text; $\times$ violated; n/a not applicable; — not stated in the release or not needed for the verdict. In the C3 column $\times$ marks a single-pass checkpoint, which the $O(\kappa/n)$ law places at a proportionally smaller gap rather than excludes.

Checkpoint	Arch.	Unique pairs	κ	B_{train}	C1	C2	C3	Verdict
CLIP-MIA-Bench 400M (checkpoint A)	ViT-B/16	$\sim$400M	32	8,192	✓	✓	✓	Admissible; the main target used.
CLIP-MIA-Bench 40M	ViT-B/16	$\sim$40M	32	8,192	✓	$\times$	✓	Fails C2 by $10\times$; shares A's member and non-member shards, so a rung on a κ/n ladder.
CLIP-MIA-Bench 4M	ViT-B/16	$\sim$4M	32	8,192	✓	$\times$	✓	Fails C2 by $100\times$; per-pair AUC 94.1, outside the small-shift regime of Section 3.
CommonPool.L-s1B-b8K (checkpoint B)	ViT-B/16	1.28B	1.0	8,192	✓	✓	$\times$	Non-members from CommonPool-XL files outside CommonPool-L; one pass, $\pi_B = 0.63$. Used as the one-pass contrast.
CommonPool.L.laion-s1B-b8K (checkpoint L)	ViT-B/16	130M	9.8	8,192	✓	✓	✓	Admissible on a filter-matched split (Appendix C); a replication.
CommonPool.L.clip-s1B-b8K (checkpoint C)	ViT-B/16	384M	3.3	8,192	✓	✓	✓	Admissible on a filter-matched split (Appendix C); a replication.
CommonPool.M-s128M-b4K	ViT-B/32	128M	1.0	4,096	✓	$\sim$	$\times$	Different architecture and B_{train}, but one pass and n at the C2 boundary; a second one-pass point, not a new axis.
CommonPool.S-s13M-b4K	ViT-B/32	12.8M	1.0	4,096	✓	$\times$	$\times$	Fails C2 and makes one pass.
DataComp.L-s1B-b8K	ViT-B/16	$\sim$140M	$\approx$9	8,192	$\times$	$\sim$	✓	A filter-matched split needs its image-based filter, not reproduced here.
DataComp.XL-s13B-b90K	ViT-L/14	$\sim$1.4B	$\approx$9	90,112	$\times$	✓	✓	"Trained on a filtered subset of the largest pool, so every pair passing its filter is a member: no filter-matched split exists.
CommonPool.XL-s13B-b90K	ViT-L/14	12.8B	1.0	90,112	n/a	✓	$\times$	Trained on all of CommonPool-XL, so no non-members exist; a candidate exposure-balanced blind control only.
OpenAI CLIP	ViT-B/16	$\sim$400M	32	32,768	$\times$	✓	✓	WIT-400M is private; no non-member set is constructible.
OpenCLIP LAION-400M	ViT-B/16	$\sim$400M	32	32,768	$\times$	✓	✓	Trained on all of LAION-400M; no held-out subset exists.
MetaCLIP-400M	ViT-B/16	$\sim$400M	32	—	$\times$	✓	✓	Balancing subsamples at random, so a split exists in principle, but the rejected pairs were never downloaded.
SigLIP (WebLI)	ViT-B/16	private	—	$\sim$32k	$\times$	✓	—	WebLI is not public; no membership labels exist. Not a target (Section 8).

same pool; Appendix O repeats every measurement on unique pairs. DataComp nests its pools at the file level: the 2,510 parquet files of CommonPool-L are byte-identical to files of CommonPool-XL, a fraction 0.105 of its 23,874, so non-members drawn from the other files of CommonPool-XL lie outside the training pool. Our non-member shards come from such files, which we verified by file hash and, pair by pair, by identifier (Appendix C).

Covariate audit. Table 6 compares the DataComp metadata of 30,000 pairs from each pool. The largest standardized difference is $|d| = 0.013$ (original width); every non-zero d has the same sign, with the non-member pool marginally larger on every covariate, including the two precomputed CLIP alignment scores, the closest model-free proxy for the statistic under study. Any leakage through these covariates therefore biases the measured membership effect downward. The audit cannot by itself certify exchangeability at the set sizes used, since a per-pair shift of $|d| = 0.013$ passed fully into a score would reach $d' \approx 0.59$ at $T = 2048$; the blind model of Section 5 does that job.

Table 6: **Covariate audit of checkpoint B's pools**: DataComp metadata of 30,000 pairs per pool; $d = (\text{member} - \text{non-member})/\sqrt{(s^2_{\text{mem}} + s^2_{\text{non}})/2}$. The release strips faces, so the face count is zero in both pools.

Covariate	Member pool, mean (SD)	Non-member pool, mean (SD)	Cohen's d
Original image width (px)	423.7 (423.5)	429.2 (435.8)	-0.013
DataComp CLIP ViT-L/14 score	0.209 (0.064)	0.209 (0.064)	-0.012
Caption length (words)	6.8 (8.7)	6.9 (8.2)	-0.011
Caption length (characters)	46.7 (60.5)	47.2 (55.3)	-0.008
Original image height (px)	361.4 (476.0)	364.1 (375.6)	-0.006
DataComp CLIP ViT-B/32 score	0.254 (0.051)	0.254 (0.052)	-0.004
Detected faces	0.0 (0.0)	0.0 (0.0)	+0.000

Membership under one pass with resampled shards. Checkpoint B made 1.0 passes over its pool against 32 for checkpoint A (Section 2). It saw $\kappa = 1.0$ times its pool size in training samples, and the DataComp training script draws shards with replacement rather than iterating over the pool once. The number of times a given pair is seen is therefore approximately Poisson with mean κ, and a pair is seen at least once with probability $\pi_B = 1 - e^{-\kappa} = 0.63$; a fraction $e^{-\kappa}$ of the nominal members was never seen. Because the fixed-bank statistic is linear in the member fraction (Appendix O), every set-level effect on checkpoint B is diluted by π_B, and the per-pair effect among pairs actually seen is δ/π_B. Checkpoint A made 32 passes over a fixed corpus, so its corresponding fraction is one. The pass count is one of four factors that separate the checkpoints (Section 2), and it is not by itself an explanation of their different shift profiles: the similarity gap of a per-example attack shrinks with fewer passes (Gomez, 2026), and that dependence is smooth in the pass count, so it predicts a smaller raw gap on checkpoint B, not a change in where the shift sits; this paper isolates none of the four factors, and checkpoints L and C, which share B's recipe, vary the filter and the pass count together (Appendix C).

Class gap in crossing rate. Table 7 gives the raw class gap in the symmetric crossing rate, $\Delta p(B)$, over 50 partitions. Checkpoint A has a broad maximum: the gap at $B = 8192$ exceeds the gap at 4096, the largest value, 22.3×10^{-3}, is at $B = 16{,}384$, and the gaps from 8,192 to 32,768 lie within two units of it. Checkpoint B has no resolvable peak: its gap varies by less than one unit from 512 to 16,384, comparable to its spread over partitions, and the bootstrap places its maximum anywhere in that range. Because both models were trained with $B = 8{,}192$, these profiles cannot test whether the peak tracks the training batch size.

Checkpoint B under the near-tie model. A bank of $N_0 = 4096$ operates at $\bar{p} = 0.46$ on checkpoint B, where the uniform-shift model bounds the crossing score at about $0.80\times$ mean alignment (measured ratio 0.81); shrinking the bank until one pair in ten is beaten lowers the ratio to 0.59, as the model requires. The loss's excess slope on B is small, as its nearly flat fixed-bank $\delta(N)$ predicts, and diluted by $\pi_B \approx 0.63$. On A's pools a second public model (LAION-400M) gives $+0.46$ for the full detector, between the OpenAI and checkpoint-B estimates; the checkpoint-B control sees A's split with the shift's tail-concentrated shape (contrast 0.89 $[0.24, 1.61]$), so lower-tail detectors

Table 7: **Class gap in the symmetric crossing rate**, $\Delta p(B) \times 10^3$, mean (SD) over 50 batch partitions, both checkpoints.

B	512	1,024	2,048	4,096	8,192	16,384	32,768	65,536
Checkpoint A	9.9 (0.3)	13.0 (0.4)	16.1 (0.4)	18.9 (0.5)	21.2 (0.6)	22.3 (0.6)	21.4 (0.8)	19.5 (1.2)
Checkpoint B	10.8 (0.6)	11.4 (0.6)	11.6 (0.6)	12.1 (0.6)	12.0 (0.6)	11.3 (0.5)	10.3 (0.5)	9.0 (0.6)

inherit the confound at the rate the mean does, and its rival term is $+0.03$, within noise of zero, against $+0.05$ for the OpenAI model.

An exposure-balanced blind control. A blind model measures how much of a detector's verdict is the data split rather than membership, and it is only as good as its exposure to the two pools. The OpenAI model saw neither pool, but it also saw neither corpus. Checkpoint B trained on CommonPool-L, from which both of checkpoint A's pools are drawn, with shards resampled, so it saw the two pools at the same rate and is exposure-balanced on them. Table 8 scores every model on checkpoint A's pools with the same detector. Checkpoint B separates A's pools about twice as far as the OpenAI model does, through alignment alone: its rival term is within noise of zero, as it is for the OpenAI model. We take the larger figure as the confound for A's split. Its shift has the same tail-concentrated shape as the membership signal, with a low-minus-high contrast of 0.89 $[0.24, 1.61]$ in units of its own mean shift against 2.49 $[2.10, 3.00]$ for checkpoint A itself, so a lower-tail detector inherits the confound at about the rate the mean does rather than escaping it. The OpenAI model's contrast on the same pools, 2.02, has an interval that includes zero, so its shape is not resolved. A second public model, trained on LAION-400M, agrees that A's split carries a confound: it separates the pools at $+0.46$ $[+0.16, +0.74]$, between the other two estimates. Unlike them it reads the split partly through its rival term ($+0.36$ $[+0.08, +0.67]$, against $+0.05$ and $+0.03$), which is why the exposure-balanced control, whose rival term is within noise of zero, is the one we quote.

Table 8: **Blind models on checkpoint A's pools.** Signed effect size at $T = 2048$ with 95% half-sample intervals, positive in the membership direction. "OpenAI" and "LAION-400M" are public checkpoints that saw neither pool and neither corpus; "Checkpoint B" is the CommonPool-L model of Appendix B, which saw both pools at the same rate. Checkpoint A is shown for scale.

Score, signed d' at $T = 2048$	Checkpoint A	OpenAI	LAION-400M	Checkpoint B
Full crossing detector	+3.14 [2.77, 3.47]	+0.36 [0.08, 0.66]	+0.46 [0.16, 0.74]	+0.70 [0.42, 0.99]
Rival term $s_{\max}$ alone	+0.29 [-0.03, 0.57]	+0.05 [-0.24, 0.35]	+0.36 [0.08, 0.67]	+0.03 [-0.26, 0.32]
Mean alignment	+2.14 [1.84, 2.43]	+0.32 [0.02, 0.62]	+0.44 [0.16, 0.72]	+0.90 [0.63, 1.19]
InfoNCE loss, in-batch	+3.20 [2.84, 3.54]	+0.43 [0.12, 0.68]	+0.48 [0.18, 0.76]	+0.84 [0.54, 1.12]

Checkpoint B on the filtered splits. Checkpoint B trained on CommonPool-L, so it saw the members of L's and C's splits at rate π_B and none of their non-members: on those splits it is a diluted member model, not a blind control, which is why Table 4 gives no control for L and C. Its signal there is small (full detector $+0.23$ $[-0.08, +0.52]$ on L's split and $+0.48$ $[+0.16, +0.79]$ on C's, all pairs). On L's split its shift contrast, -0.40 $[-2.71, 0.92]$, lies below L's own, 2.28 $[1.98, 2.59]$, on the same pairs, so the tail concentration is not a property of the filtered pairs themselves; on C's split its contrast, 0.40 $[-1.11, 1.57]$, is not resolved. Whether the pass count or the training filter produces the tail shape remains open (Section 8).

C FILTER-MATCHED SPLITS FOR CHECKPOINTS L AND C

Checkpoints and filters. Checkpoints L and C are the DataComp models `CommonPool.L.laion-s1B-b8K` and `CommonPool.L.clip-s1B-b8K` (Gadre et al., 2023): ViT-B/16 dual encoders trained with checkpoint B's recipe, batch size $8{,}192$ and 1.28×10^9 samples seen, but on the subsets of CommonPool-L that pass the LAION filter (130M pairs) and the CLIP-score filter (384M pairs). They therefore make $\kappa = 9.8$ and 3.3 passes; since shards are resampled with replacement (Appendix B), a fraction $1 - e^{-\kappa}$ of each training set was seen

at least once, >0.999 for L and 0.964 for C. Their trained logit scales are $t = 96.6$ and 92.5. The LAION filter keeps pairs whose ViT-B/32 CLIP similarity is at least 0.28 and whose caption `gcld3` identifies as English; the CLIP-score filter keeps pairs whose ViT-L/14 CLIP similarity lies in the top 30% of the pool. Both are functions of fields in the released metadata, so we apply DataComp's own filtering code to members and non-members alike. On our held-out pools the pass rates are 10.5% and 29.9%, consistent with the reported subset sizes. The CLIP-score threshold is a quantile of the whole pool; we estimate it from 3.17M metadata rows and raise it by a small guard band (0.2429 against the estimated 0.2426) for both classes, so that no retained member can lie below the true threshold.

Split construction. A filtered checkpoint's members are the pairs of CommonPool-L that pass its filter, so a split against unfiltered non-members violates C1: its members are better aligned by construction (Appendix A). We therefore apply the same filter to both pools. Members are drawn from CommonPool-L files, non-members from CommonPool-XL files outside CommonPool-L. DataComp nests its pools at the file level: all $2{,}510$ parquet files of CommonPool-L are byte-identical to files of CommonPool-XL, a fraction 0.105 of its $23{,}874$. Our non-member files are none of them, by file hash, and pair by pair none of their $4{,}389{,}381$ identifiers occurs among CommonPool-L's 1.28×10^9. The same non-member files serve checkpoint B, whose split is therefore also free of members (Appendix B). The evaluation pools hold $200{,}000$ pairs per class for L and $160{,}000$ for C, with held-out pools of $50{,}000$/$70{,}000$ and $20{,}000$/$30{,}000$ member/non-member pairs for banks and references. Download success is the same in both classes (63% and 61% in the two download runs), the largest standardized covariate difference between the classes is $|d| = 0.0061$ for L and 0.0095 for C (against 0.013 on B's pools, Table 6), and the blind model separates neither split (Table 4).

Placeholder images and unique pairs. Downloaded in 2026, many dead image links no longer fail but return a placeholder image from the hosting service, and one such image recurs across thousands of pairs of both classes. Together with ordinary duplicates, Appendix O's rule (an image hash or exact caption occurring more than once within a class, with the hash taken over the downloaded bytes) flags 5.1% of the pairs of L's split and 3.4% of C's, in equal proportions in both classes. These pairs are poorly aligned under every model, including the blind one, and dominate the spread of the per-pair loss (Table 9). We report L and C on the remaining unique pairs. The rule never looks at the model under audit and is applied identically to candidates, references and banks, so, unlike the trimming of Section 5, it cannot manufacture a verdict. It is the rule Appendix O already applied to A and B, whose results it leaves unchanged; we adopted it as the primary analysis for L and C after finding the placeholders, and Table 9 reports all pairs beside it.

All pairs against unique pairs. On all pairs the placeholders double to triple the per-pair spread of the loss and weaken every unbounded statistic: most the in-batch loss, the top decile of the loss and the lower-tail mean, less the mean alignment. The bounded scores, the bank crossing score and the threshold count, move by at most 0.04 in AUC, and the shape of the shift does not move (Table 9). Duplicates also inflate the in-batch amplification, because identical pairs create exact ties among batch-mates; on unique pairs it is smaller but still above checkpoint A's. The same tail explains why, on all pairs, a bounded score can beat the mean even under a uniform shift: the near-tie model's uniform-shift bound is 1.12 on L and 1.06 on C on all pairs, against 0.90 and 0.91 on unique pairs.

Results on unique pairs. Table 10 collects the quantities the main text cites for L and C without tabulating them. The bank optimum and its crossing rate match checkpoint A's, and the fixed-bank sweep predicts the in-batch local slopes of both statistics, with the flat null rejected, on both checkpoints.

D Reproducibility details

Embeddings. Each model is loaded through the open_clip library with its own image preprocessing and tokenizer (for the DataComp checkpoints B, L and C, all come with the released configuration). Images and captions are encoded once with the model in evaluation mode, the outputs are normalized to unit length in float32, and the resulting image and text embeddings of every pool are cached once per model and reused by every experiment; no experiment re-encodes data. All simi-

Table 9: **All pairs against unique pairs on checkpoints L and C.** AUCs at $T = 1024$; the top decile is the share of the full-batch effect size of the loss retained by its top decile, and the spread is the per-pair standard deviation of the loss, both at $B = 2048$; amplification is the in-batch minus fixed-bank slope of the crossing indicator. Unique pairs are the main-text analysis.

	L		C	
Quantity	all pairs	unique	all pairs	unique
Pairs removed (%)	—	5.1	—	3.4
Per-pair loss spread	2.76	1.00	2.29	1.08
Top decile of the loss (%)	68	91	69	90
AUC, InfoNCE loss, in-batch	0.666	0.951	0.583	0.747
AUC, lower-tail mean, $q = 0.2$	0.805	0.993	0.654	0.830
AUC, mean alignment	0.880	0.952	0.690	0.749
AUC, threshold count, $q = 0.2$	0.985	0.989	0.794	0.806
AUC, expected crossing ψ_N, $N = 4096$	0.989	0.997	0.803	0.854
AUC, bank crossing score	0.973	0.989	0.775	0.811
Amplification, crossing indicator	0.39	0.23	0.35	0.17
Shift contrast $R_{\text{low}} - R_{\text{high}}$	2.28	2.32	2.26	2.34

Table 10: **Checkpoints L and C on unique pairs**, with checkpoint A for reference. Bank optimum $N_0^\star$ and its mean crossing rate from the neutral-bank sweep, with 95% intervals; crossing-to-alignment ratio of per-pair effect sizes, measured at $T = 2048$ (A: median over batch sizes) and from the near-tie model driven by the measured shift function; prediction test of Section 3.3 over ten local slopes up to $B = 65{,}536$: mean absolute deviation, coverage and nominal χ^2 against the prediction and against the flat null $\alpha(B) = \frac{1}{2}$.

Quantity	A	L	C
$N_0^\star$	4,096 [1,879, 7,510]	4,096 [2,124, 8,661]	4,096 [1,555, 23,166]
Crossing rate at $N_0^\star$	0.10	0.11	0.13
Crossing / alignment, measured	1.52	1.37	1.28
Crossing / alignment, model	1.53	1.38	1.33
Collision rate: MAD / coverage (%)	0.01 / 100	0.01 / 90	0.02 / 100
Collision rate: χ^2 prediction / flat	3.7 / 238.1	7.5 / 576.6	2.1 / 45.7
Loss: MAD / coverage (%)	0.01 / 100	0.02 / 100	0.02 / 100
Loss: χ^2 prediction / flat	1.5 / 154.2	4.3 / 879.3	4.8 / 101.8

larities are inner products of these cached vectors, computed in float32 on the GPU in chunks, and the $B \times B$ in-batch matrices are never stored.

Pools. Checkpoint A's corpus is the CLIPScore top 40% of the downloaded CommonPool-L metadata, of whose 1.28×10^9 entries about 80% were recovered, packed into $43{,}026$ shards of $10{,}000$ pairs (Gomez, 2026); 400M pairs over 32 passes is the published recipe of OpenAI CLIP (Radford et al., 2021). The checkpoint records $1{,}561{,}927$ optimizer steps, which at batch size $8{,}192$ is 1.28×10^{10} samples seen, or 32 passes over 400M pairs. The pools follow a held-out design rather than a post-hoc split: 26 shards removed before training form the non-member pool and 26 shards drawn after training from the $40{,}000$ training shards form the member pool, with a globally shuffled metadata order. Both pools pass through one preprocessing pipeline: resize to at most 512 pixels, face blurring, alt-text captions only. Checkpoint A uses the member and non-member pools of its benchmark: 20 evaluation shards per class ($185{,}550$ member and $178{,}040$ non-member pairs) and the remaining 6 shards per class as held-out pools. Checkpoint B uses the split of Appendix B, and checkpoints L and C the filter-matched splits of Appendix C. Held-out pools supply the bank draws, the class-matched control banks, and nothing else; the calibration of the detector test equation 5 splits the non-member evaluation pool into two halves. The blind model is evaluated on both pair of pools with its own preprocessing.

Batches, partitions, and intervals. At each batch size B, each class pool is partitioned into disjoint batches by a uniform random permutation, the remainder dropped; the point estimate pools

the batches of 50 independent partitions, generated from a fixed key so that every statistic sees the same batches. Confidence intervals come from 500 half-sample replicates: each draws $\lfloor n/2 \rfloor$ pairs without replacement within each class pool, re-partitions the subsample into five partitions, and recomputes every quantity, so no pair is ever duplicated inside a batch. Two families of quantities use other schemes, stated where they appear: the shift function and the per-pair estimator of Appendix J use a bootstrap over pairs, and the rival-free detectors and duplicate-sensitivity runs of checkpoints A and B use 200 set replicates over five partitions, whose point estimates differ from the 50-partition ones by up to 0.005 in AUC (Appendix N). The interval is recentred on the point estimate, $[\hat{\theta} - (q_{97.5} - m),\ \hat{\theta} + (m - q_{2.5})]$ with m the replicate mean, and paired quantities are formed inside each replicate before the percentiles are taken. Table 11 prints the replicate mean beside the point estimate for every quantity of Table 4; the two agree to the printed precision in all but one cell. The earlier protocol, five partitions and a with-replacement bootstrap, put the point estimate at the edge of its own interval: one partition realisation at $B = 2048$ inflated the pooled spread of every in-batch statistic, which the resampling could not average away. For $B \geq 16{,}384$, where a pair can appear in more batches than exist, the per-pair estimator is bootstrapped at fixed batch assignment and the interval is conditional on the partition.

Table 11: **Point estimate and half-sample replicate mean**, for the quantities of Table 4. The point estimate pools the batches of 50 partitions; the replicate mean averages 500 half-sample replicates. Under the earlier protocol these two differed by more than an interval half-width at $B = 2048$.

	Checkpoint A		Checkpoint B	
Quantity	point	replicate mean	point	replicate mean
$\hat{\alpha}$, InfoNCE loss, in-batch	0.64	0.64	0.57	0.57
$\hat{\alpha}$, crossing indicator, in-batch	0.62	0.62	0.54	0.54
$\hat{\alpha}$, crossing indicator, fixed bank	0.50	0.50	0.50	0.50
Difference, in-batch − fixed bank (InfoNCE loss)	0.14	0.14	0.07	0.06
Difference, in-batch − fixed bank (crossing indicator)	0.12	0.12	0.04	0.04
Detector AUC at $T = 2048$	0.986	0.987	0.796	0.798
Symmetrization gain in d'	1.13	1.13	1.08	1.08
Negative swap, d'/d'_{real}	1.04	1.04	0.95	0.95

Shard structure. Every interval in the paper resamples pairs, which is valid if pairs within a shard are no more alike than pairs across shards. Table 12 tests that on each pool of checkpoints A and B: the variance ratio of shard means to the value independent sampling implies, the intraclass correlation, and a permutation p-value from relabelling shards. Positive similarity shows no shard structure on any pool (variance ratios 1.03–1.22). The two bank scores and the lower-tail indicator do on checkpoint A's member pool, the same low-alignment tail that carries the signal; resampling shards instead of pairs widens their intervals by at most $1.34\times$ without changing a conclusion. Checkpoint B's member pool shows a marginal effect for the lower-tail indicator alone. A blind model's c_{ii} shows no structure on any pool, so what the bank scores see is not a property of how the shards were written.

Logit scales. Checkpoint A is evaluated at $t \in \{20, 33, 50, 75, 100, 150\}$ with $t = 100$ its trained value; checkpoints B, L and C at their trained values $t = 85.4$, 96.6 and 92.5 (B also at $t \in \{20, 50, 150\}$ where a temperature sweep is shown); the blind model at $t = 100$. The trained value is read from each checkpoint's logit-scale parameter.

Compute. Every run uses one NVIDIA A40 GPU with the pools resident in device memory; wall time per checkpoint ranges from a few minutes (detector, bank sweep) through about half an hour (in-batch statistics, large-batch estimation, fixed-bank control) to about an hour and a half (permutation nulls with 500 replicates). The analysis scripts write one CSV per quantity; the figure scripts and the table generator read only those CSVs; and every number in the text is a macro generated from a single results file that records, for each macro, the CSV row and the command that produced it.

Table 12: **Shard structure of the per-pair scores.** For each pool: VR, the variance of shard means over the value independent sampling implies; ICC, the intraclass correlation; and p_{perm}, a permutation test that relabels shards. VR above one and a small p mean pairs within a shard are more alike than pairs across shards.

Pool		c_{ii}	bank crossing	fixed-bank loss	$\mathbf{1}[c_{ii} < Q_{0.10}]$	blind c_{ii}
A, member	VR	1.03	2.22	2.65	3.46	0.91
	ICC $\times 10^4$	0.0	1.3	1.8	2.7	-0.1
	p_{perm}	0.439	≤ 0.001	≤ 0.001	≤ 0.001	0.573
A, non-member	VR	0.78	0.68	0.79	0.70	0.92
	ICC $\times 10^4$	-0.2	-0.4	-0.2	-0.3	-0.1
	p_{perm}	0.749	0.832	0.719	0.833	0.584
B, member	VR	1.15	0.89	1.11	1.67	0.97
	ICC $\times 10^4$	0.1	-0.1	0.1	0.7	-0.0
	p_{perm}	0.315	0.599	0.356	0.042	0.500
B, non-member	VR	1.22	1.16	1.11	0.82	0.73
	ICC $\times 10^4$	0.2	0.2	0.1	-0.2	-0.3
	p_{perm}	0.260	0.286	0.321	0.677	0.807

E EFFECT-SIZE ESTIMATORS AND THEIR VALIDITY

From AUC to effect size. Suppose the set statistic is Gaussian in both classes with a common standard deviation, and the class means lie d' standard deviations apart. The difference between the statistic of a random non-member set and that of a random member set is then Gaussian with mean d' and variance 2 in those units, so $\mathrm{AUC} = \Pr(\text{difference} > 0) = \Phi(d'/\sqrt{2})$ and $d'_{\mathrm{AUC}} = \sqrt{2}\,\Phi^{-1}(\mathrm{AUC})$ (Figure 6). Unlike AUC, d' does not saturate, and under averaging of independent scores it grows as $\sqrt{m}$.

Canonical estimator. Near $\mathrm{AUC} = 1$ the probit map is steep and a finite number of sets makes d'_{AUC} unreliable. All slopes in the main text therefore use the moment estimator $d' = (\hat{\mu}_{\mathrm{non}} - \hat{\mu}_{\mathrm{mem}})/\hat{s}_{\mathrm{pooled}}$, which does not saturate. The choice does not change any slope in the main text: for every statistic of Tables 4 and 2, the slopes fitted from the two estimators on the common window differ by at most 0.02, and the moment estimate lies inside the half-sample interval of the AUC-based one (Table 13). The moment estimate is slightly higher for the loss because the AUC map compresses the largest effect sizes at $B = 2048$. Restricting the fit window to $B = 64$–1024 moves each slope by less than its interval width.

Table 13: **Fitted slope $\hat{\alpha}$ by estimator**, checkpoint A, window $B = 64$–2048, with 95% half-sample intervals. $\hat{\alpha}_{\mathrm{mom}}$ uses the moment estimator across disjoint sets; $\hat{\alpha}_{\mathrm{AUC}}$ fits $\sqrt{2}\,\Phi^{-1}(\mathrm{AUC})$. The last column is the moment estimate on the shorter window $B = 64$–1024.

Statistic	$\hat{\alpha}_{\mathrm{mom}}$	$\hat{\alpha}_{\mathrm{AUC}}$	difference	$\hat{\alpha}_{\mathrm{mom}}$, $B = 64$–1024
InfoNCE loss, in-batch, $t = 20$	0.42 [0.40, 0.44]	0.42 [0.40, 0.44]	-0.00	0.41 [0.39, 0.43]
InfoNCE loss, in-batch, $t = 33$	0.45 [0.43, 0.47]	0.45 [0.43, 0.47]	+0.00	0.46 [0.44, 0.48]
InfoNCE loss, in-batch, $t = 50$	0.55 [0.53, 0.57]	0.54 [0.52, 0.56]	+0.01	0.57 [0.54, 0.59]
InfoNCE loss, in-batch, $t = 75$	0.61 [0.59, 0.64]	0.60 [0.58, 0.63]	+0.01	0.62 [0.60, 0.65]
InfoNCE loss, in-batch, $t = 100$	0.64 [0.61, 0.66]	0.62 [0.60, 0.65]	+0.01	0.65 [0.61, 0.67]
InfoNCE loss, in-batch, $t = 150$	0.66 [0.63, 0.68]	0.64 [0.61, 0.67]	+0.01	0.67 [0.63, 0.70]
Collision rate, in-batch	0.62 [0.59, 0.65]	0.63 [0.60, 0.66]	-0.01	0.64 [0.60, 0.67]
Mean alignment $\bar{A}$	0.50 [0.48, 0.51]	0.50 [0.48, 0.51]	-0.00	0.50 [0.48, 0.51]
Crossing score, fixed bank	0.50 [0.49, 0.52]	0.50 [0.49, 0.52]	-0.00	0.50 [0.49, 0.51]
InfoNCE loss, fixed bank, $t = 100$	0.50 [0.49, 0.51]	0.50 [0.48, 0.52]	-0.00	0.50 [0.49, 0.52]

Validity checks. The binormal model behind the AUC map holds where it is used. On probit axes the ROC of the batch-mean loss at $B = 1024$ and $t = 100$ is a straight line of slope 0.96 (Figure 6), so the two classes have nearly equal spread. The skewness and excess kurtosis of the batch-mean loss are large only for small B at low temperature, that is, at large t, and decay to about 0

by $B = 2048$ (Table 14); at high temperature ($t = 20$) both are small at every B. Skewness roughly halves per octave of B and excess kurtosis falls faster, as the batch mean of a skewed per-pair loss approaches normality.

Table 14: **Skewness / excess kurtosis of the batch-mean InfoNCE loss**, checkpoint A, by logit scale t and batch size B, pooled over both classes and 50 partitions.

t	$B = 64$	$B = 128$	$B = 256$	$B = 512$	$B = 1024$	$B = 2048$
20	0.33 / 0.18	0.16 / 0.05	0.10 / -0.00	0.05 / -0.01	0.02 / 0.04	0.02 / -0.01
33	0.82 / 1.07	0.46 / 0.34	0.28 / 0.10	0.16 / -0.00	0.08 / 0.03	0.05 / 0.01
50	1.33 / 2.63	0.77 / 0.89	0.46 / 0.29	0.27 / 0.07	0.15 / 0.07	0.07 / -0.03
75	1.69 / 4.13	0.97 / 1.35	0.56 / 0.43	0.34 / 0.16	0.20 / 0.09	0.08 / -0.05
100	1.89 / 5.21	1.07 / 1.66	0.61 / 0.51	0.37 / 0.20	0.22 / 0.09	0.09 / -0.05
150	2.12 / 6.67	1.20 / 2.09	0.67 / 0.64	0.41 / 0.26	0.24 / 0.10	0.11 / -0.03

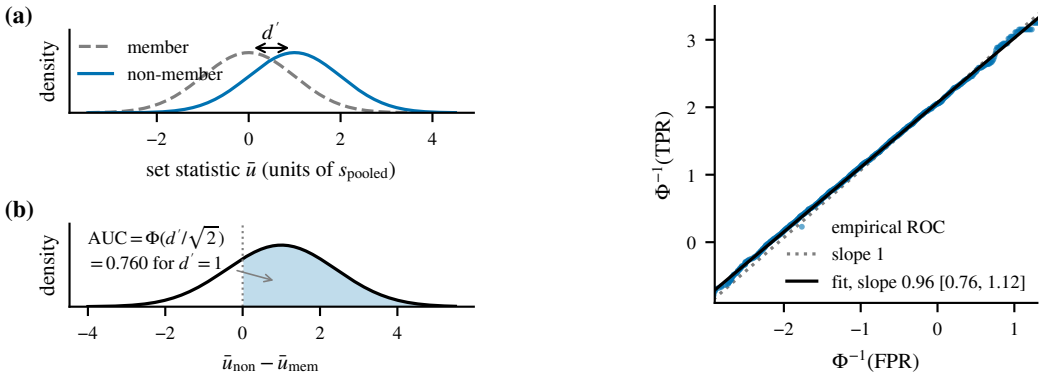


Figure 6: **Left**: **(a)** two equal-variance set-statistic densities d' apart, member (gray, dashed) and non-member (blue), in units of the pooled standard deviation; **(b)** the density of the difference $\bar{u}_{\text{non}} - \bar{u}_{\text{mem}}$, whose mass above zero (shaded) is the AUC; for $d' = 1$ it is $\Phi(1/\sqrt{2}) = 0.760$. **Right**: probit-ROC of the batch-mean loss at $B = 1024$, $t = 100$, checkpoint A; blue points: empirical ROC over the pooled batches of 50 partitions; black: least-squares fit, slope 0.96 with its 95% half-sample interval in the legend; dotted gray: slope 1 through the same intercept.

F THE DECOMPOSITION AND THE NEAR-TIE MODEL

Proof of Proposition 1. Fix a class $k \in \{\text{mem}, \text{non}\}$ and a number of rivals N. Under (A1)–(A2), each score u_i is a function of the anchor's own pair and of N rival similarities drawn from a law that depends neither on the anchor nor on the class. Hence $\mathbb{E}[u_i \mid k] = \mu_k(N)$ and $\text{Var}[u_i \mid k] = \sigma_k^2(N)$ for all i. Writing $\bar{\rho}_k$ for the average correlation between the scores of two distinct pairs in the same set,

$$\mathbb{E}[\bar{u} \mid k] = \mu_k(N), \qquad \text{Var}[\bar{u} \mid k] = \frac{\sigma_k^2(N)}{m}\big(1 + (m-1)\bar{\rho}_k\big).$$

With $\sigma_{\text{mem}} = \sigma_{\text{non}} = \sigma$ and $\bar{\rho}_{\text{mem}} = \bar{\rho}_{\text{non}} = \bar{\rho}$, the difference of the class means of $\bar{u}$, divided by its common standard deviation, is $\sqrt{m}\,\delta(N)/\sqrt{1 + (m-1)\bar{\rho}}$. □

For in-batch statistics, differentiating $\log d'_{\text{set}}(B, B-1)$ with respect to $\log B$ gives equation 3, after setting the factor $B/(B-1)$ on the middle term to one. The equal-variance assumption is supported by the probit-ROC slope of 0.96 (Appendix E).

Near-tie model for the crossing indicator. Consider one retrieval direction. Let rival similarities be i.i.d. with distribution function F, density f, and survival function $S = 1 - F$, and let an anchor have positive similarity c. The anchor is beaten with probability

$$p_N(c) = 1 - F(c)^N \approx 1 - e^{-NS(c)}.$$

Suppose the member distribution of c is the non-member distribution shifted by $\Delta c > 0$. To first order in Δc,

$$\mu_{\text{non}}(N) - \mu_{\text{mem}}(N) \approx \Delta c\, \mathbb{E}_c\big[N f(c) F(c)^{N-1}\big], \qquad \sigma^2(N) = \bar{p}_N(1-\bar{p}_N), \quad \bar{p}_N = \mathbb{E}_c[p_N(c)],$$

where the variance is that of a Bernoulli variable with the class-average rate. This is the total variance of the indicator across randomly drawn anchors: by the law of total variance it equals $\mathbb{E}_c[p_N(c)(1 - p_N(c))] + \mathrm{Var}_c[p_N(c)]$, and it is what the pooled standard deviation across pairs estimates. The within-anchor term $\mathbb{E}_c[p_N(c)(1 - p_N(c))]$ alone omits the spread of crossing probabilities across anchors and overstates δ (Table 15).

Identical anchors. If every anchor has the same c, write $\lambda = NS(c)$ and $p = 1 - e^{-\lambda}$. Then $NfF^{N-1} \approx (f/S)\,\lambda e^{-\lambda}$ and

$$\delta(N) \approx \Delta c\, \frac{f(c)}{S(c)}\,\big(-\ln(1-p)\big)\sqrt{\frac{1-p}{p}}.$$

As $p \to 0$, $\delta \propto \sqrt{p} \propto \sqrt{N}$, so $\mathrm{d}\log\delta/\mathrm{d}\log N \to \frac{1}{2}$ and the in-batch local slope approaches 1; as $p \to 1$, $\delta \to 0$. The maximum lies at $p \approx 0.8$.

Heterogeneous anchors. When c_{ii} varies across anchors, increasing N drives poorly aligned anchors to $p_i \to 1$, where their density term vanishes, before it brings well-aligned anchors into the transition region. The class-average optimum then moves to much lower crossing rates; empirically it lies at $\bar{p}^\star \approx 0.10$ (95% CI 0.07–0.13; Section 5). Evaluated with the empirical rival distribution F of the neutral bank and the empirical c_{ii} of every evaluation pair, the model reproduces the measured per-pair effect size of the symmetric crossing score, including its rise, peak, and decline, to within a few percent for $N_0 \geq 64$, and the measured crossing rate to within about a tenth for $N_0 \geq 2048$ (Table 15). At the smallest banks the model underpredicts $\bar{p}$, because the rival distribution differs between anchors and a few hub captions dominate a small bank. With the within-anchor variance $\mathbb{E}_c[p(1-p)]$ in place of $\bar{p}(1-\bar{p})$ the model overshoots the measured δ by about a factor of two near the peak.

Summary. A class shift Δc changes the crossing rate by $\Delta c\,\mathbb{E}[Nf(c)F(c)^{N-1}]$, while the score variance is $\bar{p}(1-\bar{p})$. While crossings are rare the gap grows as N and the standard deviation as $\sqrt{N}$, so $\delta \propto \sqrt{N}$ and $\alpha(B) \to 1$; once most pairs are beaten, extra rivals add variance but no signal. For identical pairs δ peaks at a crossing rate near 0.8; heterogeneity in c_{ii} moves the peak much lower, because a large N saturates poorly aligned pairs before well-aligned pairs come into contention.

Table 15: **Near-tie model against the fixed-bank measurement**, checkpoint A, symmetric crossing score, neutral bank of N_0 pairs. The model uses the empirical rival distribution and the empirical positive similarities; δ is the per-pair effect size and $\bar{p}$ the mean crossing rate; measured values are the mean over five bank draws.

N_0	16	64	256	1,024	4,096	16,384	65,536
$\bar{p}$, measured	0.004	0.012	0.027	0.057	0.108	0.191	0.307
$\bar{p}$, model	0.002	0.007	0.019	0.048	0.100	0.186	0.308
δ, measured	0.030	0.042	0.056	0.066	0.069	0.062	0.046
δ, model with $\sigma^2 = \bar{p}(1-\bar{p})$	0.027	0.041	0.057	0.069	0.071	0.062	0.046
δ, model with $\sigma^2 = \mathbb{E}[p(1-p)]$	0.032	0.054	0.085	0.120	0.136	0.129	0.104

InfoNCE as a soft crossing count. For the per-pair loss against N rivals, $\partial\ell/\partial c = -t\,q$, where $q = \sum_j e^{tc_j}/(e^{tc} + \sum_j e^{tc_j})$ is the probability mass assigned to rivals. To first order the class gap of the loss is therefore $t\,\Delta c\,\mathbb{E}[q]$. At large t, q approaches the crossing indicator and the loss inherits its dependence on N. At small t, the log-sum-exp averages over all rivals, q depends only weakly on which rivals are hardest, and δ is nearly flat in N. This is why the loss slope rises with the logit scale while that of the crossing indicator does not depend on it.

G ESTIMATING THE DESIGN EFFECT

For a set statistic $\bar{u}$ computed on disjoint sets of size m from one class, we estimate $\mathrm{DEFF}(m) = m\,\widehat{\mathrm{Var}}_{\mathrm{sets}}(\bar{u})/\widehat{\mathrm{Var}}_{\mathrm{pairs}}(u)$, where the numerator is the variance of the statistic across sets and the denominator is the variance of per-pair scores pooled over sets; independent pairs give $\mathrm{DEFF} = 1$. We compute it per class, batch size, statistic, and logit scale, average the two classes, and obtain confidence intervals from the half-sample protocol of Section 2. The correction term in equation 3 is half the local log-slope of DEFF.

Levels and slopes. Batch-mates share rivals, so the in-batch statistics have a design effect above one while the fixed-bank scores and the mean alignment sit at one (Figure 7). For the loss the level grows with temperature: it is largest at $t = 20$, where the log-sum-exp averages over many shared rivals, and reaches 1.18 at $B = 2048$ and $t = 100$. For the collision rate the level is small and flat in B: it stays close to one across the window, with intervals that exclude 1 for $B \leq 512$, and is 1.17 at $B = 2048$, where the interval is widest. What enters equation 3 is the log-slope of DEFF in B, not its level, and for both statistics that slope is consistent with zero at every interval of the window, so the correction term is small compared with the detectability term. Beyond $B = 8192$ too few batches remain per class to estimate DEFF with any precision, which is why the large-batch slopes of Appendix J use the per-pair estimator and are compared with the prediction without the design-effect term.

Independent-rival simulation. As a second estimate we replace each real batch's similarity matrix with one whose diagonal is drawn from the class distribution of positive similarities and whose off-diagonal entries are drawn independently from the pooled negative distribution. This preserves every marginal while removing the correlation that shared rivals induce (Appendix K gives the exact construction). The ratio of the simulated to the real effect size at $B = 2048$ is 1.30 at $t = 20$, 1.23 at $t = 50$, and 1.05 at $t = 100$ for the loss, and 1.06 for the collision rate. Independent rivals separate the classes better than real ones at every temperature, most at high temperature, and the ratios at $t = 100$ and for the collision rate have intervals that include one. Shared rivals therefore cost the loss some separability at high temperature and little at the trained temperature, consistent with the design-effect levels above.

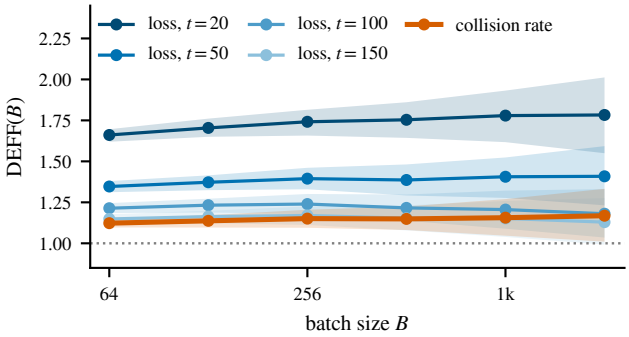


Figure 7: **Design effect of in-batch statistics**, checkpoint A. $\mathrm{DEFF}(B)$ of the batch-mean InfoNCE loss at $t \in \{20, 50, 100, 150\}$ (shades of blue, darkest for $t = 20$) and of the collision rate (vermilion) over $B = 64$–2048; bands are 95% half-sample intervals; dotted line: $\mathrm{DEFF} = 1$. The loss level rises toward high temperature (small t); all curves are flat to $B = 1024$, and the intervals widen at $B = 2048$.

H TEMPERATURE

Figure 8 reports the fitted slope of the in-batch loss at six logit scales with 95% half-sample intervals, together with the class-matched fixed-bank loss control of Section 3.2 at three bank sizes; the lower bound of the in-batch slope exceeds $\frac{1}{2}$ at the four largest values of t, while every control slope sits

at $\frac{1}{2}$ (root-mean-square deviation 0.001). Figure 9 shows the effect-size curves behind these slopes at four temperatures.

The detectability slope changes sign with t. Measured over neutral banks of $N_0 = 32$–2048, $\mathrm{d}\log\delta/\mathrm{d}\log N$ for the loss against a fixed bank is -0.091 at $t = 20$, $+0.069$ at $t = 50$, $+0.149$ at $t = 100$ and $+0.168$ at $t = 150$. Adding the design-effect term of equation 3, $-\frac{1}{2}\,\mathrm{d}\log\mathrm{DEFF}/\mathrm{d}\log B$, which is -0.010, -0.007, $+0.031$ and $+0.002$ at the same four logit scales, gives predicted in-batch slopes of 0.40, 0.56, 0.68 and 0.67. The measured in-batch slopes are 0.42 ($[0.40, 0.44]$) at $t = 20$ and 0.64 ($[0.61, 0.66]$) at $t = 100$ (Figure 8 gives all six). The prediction at $t = 20$ lies inside the measured interval; the prediction at $t = 100$ lies above the upper end of the measured interval, so the fixed-bank sweep reproduces the sign change and the size of the rise with t, which is the claim of Section 3.2, but not the slope at the trained logit scale to within its interval.

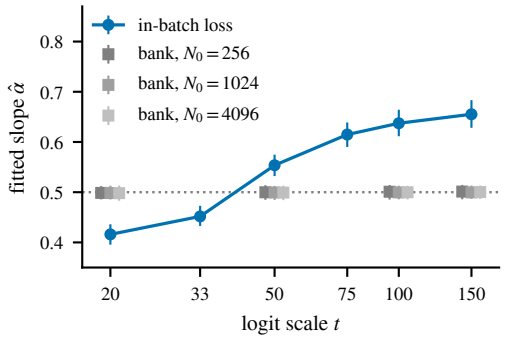


Figure 8: **Fitted slope against logit scale**, checkpoint A, window $B = 64$–2048. Blue circles: in-batch InfoNCE loss with 95% half-sample intervals. Gray squares: the same loss scored against a frozen class-matched bank of $N_0 \in \{256, 1024, 4096\}$ held-out pairs (one shade per bank size, offset slightly in t), with intervals. Dotted: slope $\frac{1}{2}$.

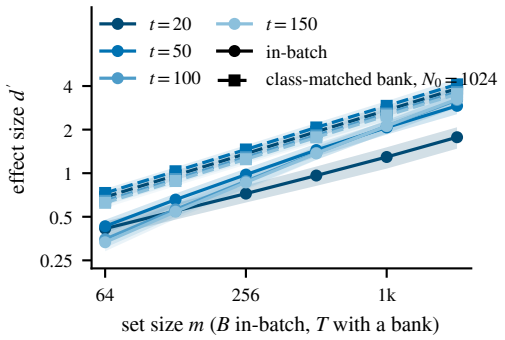


Figure 9: **Effect size against set size for the in-batch loss and its fixed-bank control**, checkpoint A. Solid lines with circles: batch-mean InfoNCE loss against batch size B at $t \in \{20, 50, 100, 150\}$ (shades of blue, darkest for $t = 20$). Dashed lines with squares: the same loss scored against a frozen class-matched bank of $N_0 = 1024$ pairs and averaged over sets of T pairs, same shades. Bands: 95% half-sample intervals. The control curves are parallel to the square-root law at every temperature; the in-batch curves fan out with t.

Temperature does more than change the number of active pairs. A simpler explanation would hold that lower temperature only reduces the fraction f_{act} of pairs with non-negligible loss, so that separability depends on B only through the number of active pairs Bf_{act}. AUC plotted against Bf_{act} would then collapse onto one curve for all temperatures. It does not (Figure 10). After rescaling, the spread across temperatures at matched Bf_{act} is 3.35 times the spread at matched B, whereas a collapse would reduce it, and at fixed Bf_{act} lower temperature still gives higher AUC at

every point of the overlapping range. Temperature changes the shape of the scaling curve, consistent with its two roles in equation 3.

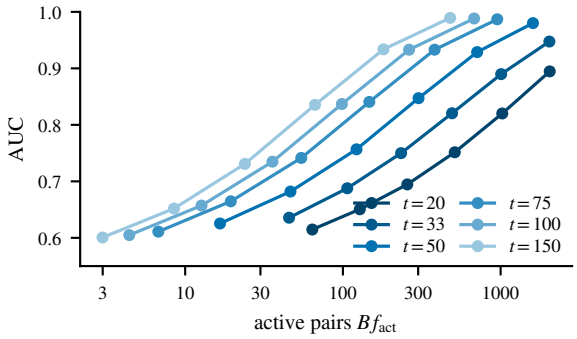


Figure 10: **Rescaling by the active fraction does not collapse the temperature curves.** AUC of the batch-mean InfoNCE loss against the number of active pairs $B \times f_{\text{act}}$ (pairs with loss above 0.01), checkpoint A, one curve per logit scale $t \in \{20, 33, 50, 75, 100, 150\}$ (darkest for $t = 20$), $B = 64$–2048. The curves separate rather than merge: at matched abscissa the spread across t is 3.35 times the spread at matched B, and larger t lies above smaller t everywhere.

I Moment decomposition of the loss slope

Write the class gap of the in-batch set statistic as $D_B = \mathbb{E}[\bar{\ell}_B \mid \text{non}] - \mathbb{E}[\bar{\ell}_B \mid \text{mem}]$ and its pooled standard deviation across batches as σ_B, so that $d'(B) = D_B/\sigma_B$. With $\sigma_B = s(B)\sqrt{\text{DEFF}(B)/B}$, where $s(B)$ is the per-pair standard deviation, taking log-derivatives gives

$$\alpha(B) - \tfrac{1}{2} = a_D - a_s - \tfrac{1}{2}\, a_{\text{DEFF}}, \qquad a_D = \frac{\mathrm{d}\log D_B}{\mathrm{d}\log B}, \quad a_s = \frac{\mathrm{d}\log s}{\mathrm{d}\log B}, \quad a_{\text{DEFF}} = \frac{\mathrm{d}\log \text{DEFF}}{\mathrm{d}\log B},$$

which is equation 3 with $a_D - a_s = \mathrm{d}\log\delta/\mathrm{d}\log B$. Because D_B, s, and DEFF are measured on the same batches as d', this identity checks the consistency of the estimates rather than providing independent evidence: at $t = 100$ the three fitted terms give $a_D - a_s - \frac{1}{2}a_{\text{DEFF}} = 0.14$, against a directly fitted excess of 0.14; the residual is the effect of the different weights of the four fits. The excess is carried by a_D: at the trained temperature the class gap grows almost as $\sqrt{B}$ while the per-pair spread grows more slowly, and a_{DEFF} contributes little.

First-order prediction of the gap. Independent content comes from predicting D_B from per-pair quantities alone. To first order in the shift of positive similarities, $D_B \approx t\,\Delta c\,\mathbb{E}[q_B]$ (Appendix F), where $\mathbb{E}[q_B]$ is the mean probability mass on rivals at batch size B. The prediction lies below the measurement at every temperature and batch size (Figure 11). It reproduces the growth of D_B with B at every t, and the ordering of the gap across temperatures at small B only. The under-prediction factor grows from $t = 20$ to $t = 50$ and then plateaus at about three for $t \geq 50$, where the loss is dominated by the hardest rival and the linearization in Δc is least accurate; at every t the factor decreases with B. On checkpoint B, whose excess slope is small, the measured and predicted gaps at $B = 2048$ are $0.078/0.070$.

Counterfactual member distributions. The near-tie model takes the non-member law of c_{ii} and a member law obtained from it, and returns everything the main text reports: the best bank size, the best per-pair effect size and its ratio to mean alignment, the factor by which the measured class gap of the loss exceeds the first-order prediction, and the in-batch slope at two temperatures. Table 16 runs it on each checkpoint's measured pair of laws, on the empirical member c_{ii}, on a uniform shift of the same mean, and on the measured quantile shift function. The uniform shift is the counterfactual that matters: on neither checkpoint does it reproduce the advantage over the mean or the gap factor, and on checkpoint A it puts the optimum beyond the largest bank evaluated. The measured shift function reproduces the ratio and the gap factor on both checkpoints and checkpoint A's optimum inside its interval; checkpoint B's optimum is flat in N_0, so no model pins it down. On checkpoint A

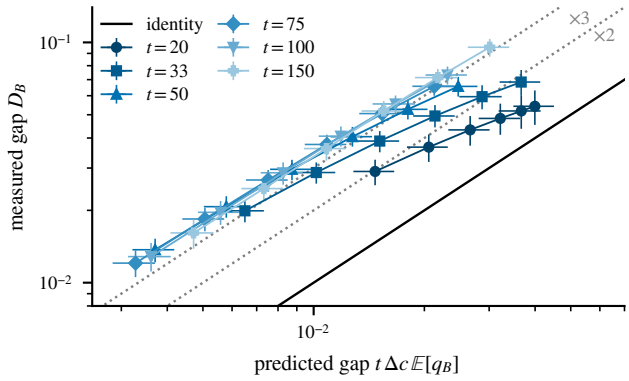


Figure 11: **Measured class gap D_B of the batch-mean loss against the first-order prediction** $t\,\Delta c\,\mathbb{E}[q_B]$, checkpoint A, log–log axes. One marker shape and shade of blue per logit scale $t \in \{20, 33, 50, 75, 100, 150\}$; the six points of a series are $B = 64$–2048, increasing along the series; bars are 95% half-sample intervals on both axes. Black: identity; dotted gray: measured $= 2\times$ and $3\times$ predicted. All points lie above the identity; the series for $t \geq 50$ cluster near the $3\times$ guide and the $t = 20$ series lies below the $2\times$ guide.

at $t = 100$ the gap factor is 3.08 under the measured shift function against 1.00 under a uniform shift, a threefold excess of the loss gap over the uniform-shift prediction. Table 17 gives the gap factor at four logit scales. The separation is a high-temperature effect: at $t = 20$ neither model reaches the measured factor on either checkpoint, while from $t = 50$ upward the measured shift function tracks it on A and a uniform shift stays at one everywhere.

Table 16: **The near-tie model under counterfactual member distributions.** For each checkpoint: the measured row, then the model driven by the empirical member c_{ii}, by a uniform shift of the same mean, and by the measured quantile shift function. $\delta^\star$, $N_0^\star$ and $\bar{p}^\star$ are the best per-pair effect size and the bank size and crossing rate that attain it, $N_0^\star$ as a continuous fit in every row (the main text rounds the measured one to the nearest grid point, 4,096); $\delta^\star/\delta_{\text{mean}}$ is a per-pair ratio to mean alignment at that optimum, not the set-level ratio of Section 4; the gap factor is the measured class gap of the in-batch loss at $B = 2048$ over its first-order prediction; the last two columns are in-batch slopes. †The optimum lies at or beyond the largest bank evaluated.

	Member distribution	$\delta^\star$	$N_0^\star$	$\bar{p}^\star$	$\delta^\star/\delta_{\text{mean}}$	gap factor, t_{train}	$\alpha(20)$	$\alpha(t_{\text{train}})$
A	measured	0.069	3,603	0.10	1.48	3.17	0.42	0.64
	empirical member c_{ii}	0.071	2,507	0.08	1.53	3.15	0.41	0.68
	uniform shift	0.038	≥65,536†	0.31†	0.81	1.00	0.53	0.76
	local shift $\Delta(F_{\text{non}}(c))$	0.071	2,634	0.08	1.53	3.08	0.42	0.72
B	measured	0.027	5,577	0.49	0.91	1.12	0.52	0.57
	empirical member c_{ii}	0.028	18,194	0.59	0.92	1.07	0.50	0.61
	uniform shift	0.024	≥65,536†	0.69†	0.80	1.00	0.51	0.56
	local shift $\Delta(F_{\text{non}}(c))$	0.028	14,361	0.57	0.93	1.05	0.50	0.61

J LARGE-BATCH ESTIMATION

At $B = 65{,}536$ the $B \times B$ similarity matrix does not fit in memory and each class pool yields only two or three disjoint batches per partition. We therefore compute row and column statistics in streamed chunks and estimate d' from per-pair moments, $d'(B) = \sqrt{B}\,\Delta\mu(B)/\sigma(B)$, using all pairs at every B and all 50 partitions. This estimator omits the design effect by construction. Where both estimators are computable, the ratio of the direct batch-level estimate to the per-pair estimate is roughly constant at 0.96 from $B = 1024$ onward (range 0.91–1.06 over the grid), against the $1/\sqrt{\text{DEFF}} = 0.92$ that the measured design effect implies, and it shifts $\log d'$ but not the slope.

Table 17: **Gap factor by logit scale.** Measured class gap of the in-batch loss at $B = 2048$ over its first-order prediction, against the same quantity under a uniform shift and under the measured shift function. A uniform shift predicts a factor of one at every temperature.

	t	measured	uniform shift	local shift
A	20	1.36	1.00	1.05
A	50	2.65	1.00	2.18
A	100	3.17	1.00	3.08
A	150	3.17	1.00	3.24
B	20	1.82	1.00	1.00
B	50	1.35	1.00	1.06
B	85.4	1.12	1.00	1.05
B	150	1.03	1.00	1.02

With few batches per class at the largest sizes the direct estimator's intervals are wide (Table 18), so apparent drifts in the ratio are noise; the per-pair estimator, whose interval comes from a bootstrap over pairs with the batch assignments held fixed, is the one compared with the fixed-bank prediction in Figure 3b. Over $B = 512$–$65{,}536$ the mean symmetric crossing rate rises from 0.037 to 0.307 and the mean margin $c_{ii} - \max_{j \neq i} c_{ij}$ falls from 0.097 to 0.019 (Figure 12). Neither has saturated within the measured range: the margin is still positive at $B = 65{,}536$. On the per-pair estimator the collision-rate slope crosses $\frac{1}{2}$ at $B \approx 2{,}784$ and its last three local slopes average 0.32; the loss slope at the trained logit scale declines from 0.59 to 0.48 on the same measure and crosses $\frac{1}{2}$ at $B \approx 13{,}127$ (Table 18).

Checkpoints B, L and C. The per-pair estimator gives the same kind of curve on the other checkpoints (Table 18); the fixed-bank prediction of their local slopes is tested in Section 3.3 and, for L and C, in Table 10. On L and C the collision-rate slope crosses $\frac{1}{2}$ at $B \approx 4{,}699$ and $4{,}931$, inside the intervals of their bank optima $N_0^\star$ as on A, and the loss slope at $B \approx 31{,}864$ and $33{,}398$; their last three collision-rate slopes average 0.34 and 0.38. On B the loss slope stays within 0.05 of $\frac{1}{2}$ at every size, and the collision-rate slope starts at $\frac{1}{2}$ and falls below it (last three: 0.40), so its crossing point is not resolved ($B \approx 1{,}045$, 95% interval 91–13,374), as its nearly flat $\delta(N)$ implies.

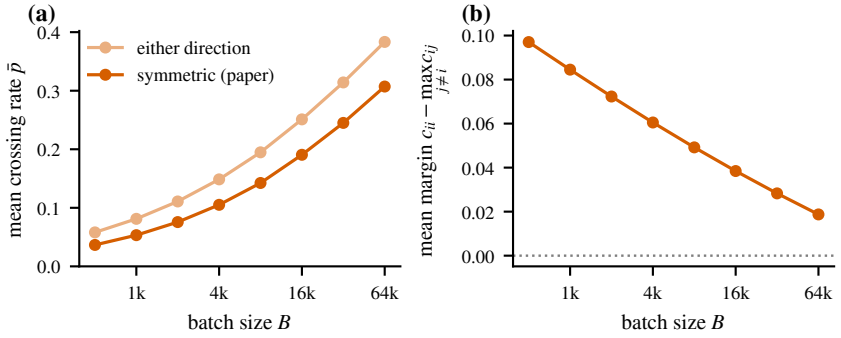


Figure 12: **Geometry of saturation**, checkpoint A, $B = 512$–$65{,}536$. **(a)** Mean crossing rate $\bar{p}$ against B: the symmetric rate used throughout the paper (dark vermilion) and the rate of being beaten in either direction (light). **(b)** Mean margin $c_{ii} - \max_{j \neq i} c_{ij}$ against B; dotted: zero. Bands are 95% half-sample intervals and are narrower than the lines. The crossing rate keeps rising and the margin keeps falling but stays positive at the largest B.

K Permutation nulls and the negative tails

Definitions. For one batch of B pairs of one class let $S^0_{ij} = c_{ij}$ be the real $B \times B$ similarity matrix (image i, caption j), with positives $c_{ii} = S^0_{ii}$. Each condition defines a row matrix (image anchor i against the B captions) and a column matrix (caption anchor i against the B images); the

Table 18: **Local slopes of d' at large batch sizes**, InfoNCE loss at each checkpoint's trained logit scale and collision rate. Checkpoint A: per-pair and direct estimators; checkpoints B, L and C (L and C on unique pairs): per-pair estimator. Intervals are 95% half-sample intervals for the direct estimator and 95% pair-level bootstrap intervals for the per-pair estimator; ‡marks direct-estimator intervals from a bootstrap over the pooled batches, a coarser interval. The direct estimator is not computed at $B = 65{,}536$ (fewer than four batches per class and partition).

B range	512–1k	1k–2k	2k–4k	4k–8k	8k–16k	16k–32k	32k–64k
Checkpoint A, per-pair and direct estimators, $t = 100$							
Loss, per-pair estimator	0.61	0.59	0.56	0.53	0.51	0.47	0.46
95% CI	[0.59, 0.63]	[0.57, 0.61]	[0.55, 0.58]	[0.52, 0.54]	[0.49, 0.52]	[0.46, 0.49]	[0.44, 0.47]
Loss, direct estimator	0.60	0.60	0.59	0.56	0.56‡	0.57‡	—
95% CI	[0.50, 0.70]	[0.47, 0.73]	[0.41, 0.76]	[0.27, 0.81]	[0.26, 0.78]	[0.46, 0.68]	—
Collision rate, per-pair estimator	0.60	0.56	0.50	0.46	0.39	0.30	0.26
95% CI	[0.57, 0.62]	[0.53, 0.58]	[0.47, 0.52]	[0.44, 0.49]	[0.35, 0.42]	[0.25, 0.34]	[0.20, 0.31]
Collision rate, direct estimator	0.57	0.57	0.51	0.45	0.51‡	0.33‡	—
95% CI	[0.46, 0.70]	[0.44, 0.72]	[0.34, 0.68]	[0.20, 0.68]	[0.22, 0.68]	[0.22, 0.44]	—
Checkpoint B, per-pair estimator, $t = 85.4$							
Loss	0.55	0.53	0.54	0.51	0.52	0.50	0.50
95% CI	[0.52, 0.57]	[0.50, 0.55]	[0.52, 0.56]	[0.49, 0.53]	[0.49, 0.54]	[0.49, 0.52]	[0.49, 0.52]
Collision rate	0.52	0.49	0.54	0.49	0.43	0.40	0.36
95% CI	[0.46, 0.58]	[0.43, 0.55]	[0.48, 0.60]	[0.42, 0.55]	[0.35, 0.50]	[0.31, 0.47]	[0.25, 0.44]
L, unique pairs, per-pair estimator, $t = 96.6$							
Loss	0.80	0.73	0.67	0.61	0.55	0.52	0.48
95% CI	[0.77, 0.84]	[0.71, 0.76]	[0.66, 0.69]	[0.60, 0.63]	[0.54, 0.56]	[0.51, 0.53]	[0.46, 0.49]
Collision rate	0.68	0.61	0.56	0.47	0.41	0.34	0.28
95% CI	[0.65, 0.70]	[0.59, 0.63]	[0.54, 0.58]	[0.46, 0.49]	[0.39, 0.43]	[0.32, 0.37]	[0.24, 0.31]
C, unique pairs, per-pair estimator, $t = 92.5$							
Loss	0.70	0.69	0.63	0.58	0.57	0.52	0.48
95% CI	[0.64, 0.78]	[0.64, 0.75]	[0.59, 0.68]	[0.55, 0.62]	[0.54, 0.61]	[0.49, 0.56]	[0.45, 0.51]
Collision rate	0.62	0.61	0.55	0.49	0.46	0.38	0.31
95% CI	[0.56, 0.69]	[0.55, 0.67]	[0.49, 0.61]	[0.43, 0.55]	[0.40, 0.52]	[0.31, 0.45]	[0.21, 0.39]

InfoNCE loss and the crossing indicator use the row matrix for image→text and the column matrix for text→image and average the two directions.

- *Real.* Row and column matrix $= S^0$.
- *Diagonal shuffle.* One uniform permutation π per batch; $S_{ii} = c_{\pi(i)\pi(i)}$, off-diagonal entries untouched, the same matrix for both directions. It destroys the link between an anchor's positive and its own rivals while keeping the exact multiset of positives of the batch.
- *Diagonal resample.* Every diagonal entry replaced by an independent draw, with replacement, from the class pool of positive similarities $\{c_{kk}\}$ of that class; off-diagonal entries untouched. It destroys the same link and the batch-level composition of positives.
- *Negative swap* (the negative swap of Table 2). Batch k of one class is paired with batch $k \bmod n_{\text{other}}$ of the other class of the same partition. Row matrix $U_{\text{own}}V_{\text{other}}^{\top}$ with the diagonal reset to the own c_{ii} (own image i against the foreign captions); column matrix $U_{\text{other}}V_{\text{own}}^{\top}$ with the same diagonal (own caption i against the foreign images), where U and V stack the image and caption embeddings of a batch. Every anchor keeps its own image, caption, and positive; only the class of its rivals changes, in both directions.
- *Caption-only variant.* The single matrix $U_{\text{own}}V_{\text{other}}^{\top}$ with own diagonal serves both directions, so the text→image direction compares c_{ii} with own images against the foreign captions. The class of the rivals therefore changes in one direction only, so the variant is not a well-posed null: on checkpoint A it agrees with the symmetric swap within the intervals, on L it keeps less of the signal (Table 19). The main text uses the symmetric swap.

- *Independent rivals* (Appendix G). Every off-diagonal entry is an independent draw from a pooled sample of 2×10^7 negative similarities (10^7 per class), and the diagonal is drawn as in diagonal resample. It destroys every correlation among rivals while keeping all marginals.

All conditions are evaluated on exactly the batches of the real matrix inside each replicate, so every ratio is a within-replicate paired ratio.

Table 19: **All null conditions**, checkpoints A, L and C (L and C on unique pairs), at each checkpoint's trained logit scale: fitted slope $\hat{\alpha}$ on $B = 64$–2048 and effect size at $B = 2048$ relative to the real matrix, d'/d'_{real}, with 95% half-sample intervals. The symmetric negative swap is the negative swap of Table 2; the collision rate does not depend on t.

	InfoNCE loss, own t		Collision rate	
Condition	$\hat{\alpha}$	d'/d'_{real}	$\hat{\alpha}$	d'/d'_{real}
A (400M pairs), $t = 100$				
Real	0.64 [0.61, 0.66]	1	0.62 [0.59, 0.65]	1
Diagonal shuffle	0.67 [0.64, 0.70]	0.93 [0.87, 0.99]	0.65 [0.62, 0.69]	0.93 [0.85, 1.01]
Diagonal resample	0.68 [0.64, 0.72]	0.96 [0.84, 1.07]	0.66 [0.62, 0.69]	0.95 [0.83, 1.06]
Negative swap, symmetric	0.65 [0.62, 0.67]	1.04 [0.97, 1.10]	0.63 [0.60, 0.65]	1.03 [0.93, 1.12]
Negative swap, caption-only	0.64 [0.61, 0.67]	1.00 [0.92, 1.07]	0.63 [0.60, 0.66]	1.00 [0.89, 1.09]
Independent rivals	0.67 [0.63, 0.71]	1.05 [0.92, 1.17]	0.65 [0.61, 0.69]	1.06 [0.92, 1.20]
L (LAION filter), unique pairs, $t = 96.6$				
Real	0.83 [0.78, 0.88]	1	0.73 [0.70, 0.77]	1
Diagonal shuffle	0.87 [0.80, 0.93]	0.84 [0.79, 0.89]	0.77 [0.73, 0.81]	0.88 [0.80, 0.95]
Diagonal resample	0.87 [0.79, 0.93]	0.84 [0.75, 0.93]	0.77 [0.73, 0.81]	0.88 [0.79, 0.97]
Negative swap, symmetric	0.84 [0.78, 0.89]	1.01 [0.95, 1.06]	0.74 [0.71, 0.77]	1.03 [0.94, 1.12]
Negative swap, caption-only	0.86 [0.78, 0.91]	0.86 [0.80, 0.93]	0.75 [0.72, 0.79]	0.90 [0.82, 0.99]
Independent rivals	0.87 [0.79, 0.94]	0.89 [0.80, 0.98]	0.77 [0.73, 0.82]	0.95 [0.85, 1.05]
C (CLIP-score filter), unique pairs, $t = 92.5$				
Real	0.73 [0.61, 0.81]	1	0.67 [0.59, 0.74]	1
Diagonal shuffle	0.75 [0.61, 0.84]	0.91 [0.82, 1.00]	0.69 [0.60, 0.77]	0.94 [0.80, 1.08]
Diagonal resample	0.75 [0.58, 0.87]	0.92 [0.75, 1.08]	0.70 [0.60, 0.80]	0.97 [0.77, 1.14]
Negative swap, symmetric	0.74 [0.63, 0.82]	1.00 [0.91, 1.10]	0.68 [0.60, 0.75]	1.02 [0.87, 1.16]
Negative swap, caption-only	0.74 [0.62, 0.82]	0.93 [0.82, 1.03]	0.68 [0.59, 0.75]	0.95 [0.78, 1.09]
Independent rivals	0.76 [0.59, 0.88]	0.95 [0.75, 1.12]	0.70 [0.60, 0.80]	1.01 [0.80, 1.19]

Diagonal resample. Figure 13 gives slopes, large-B effect-size ratios, and the decomposition of d' into class gap and pooled spread for every condition. Diagonal resample, which redraws each positive similarity from the class pool instead of permuting it within the batch, overstates the real effect size at $B = 2048$ at high temperature, by $1.14\times$ at $t = 20$, and matches it at the trained scale ($0.96\times$ at $t = 100$), while its slopes sit above the real ones (Table 19). The gain at high temperature is a spread deficit rather than a larger gap: at every temperature its class gap is at or below the real one, for $t \leq 50$ its pooled spread is reduced even more, so the quotient exceeds one, at $t = 100$ the two reductions nearly cancel, and at $t = 150$ the loss of gap dominates (Table 20). Redrawing positives from the pool destroys the batch-level correlation between a batch's positive composition and its rival structure, which inflates the between-batch variance of real batches. Diagonal shuffle preserves that correlation and is the better-posed null; the main text reports it.

Table 20: **Diagonal resample against the real matrix at** $B = 2048$, checkpoint A: ratios of the class gap, of the pooled spread across batches, and of their quotient d', with 95% half-sample intervals.

Ratio to real at $B = 2048$	loss, $t = 20$	loss, $t = 50$	loss, $t = 100$	loss, $t = 150$	collision rate
Class gap	0.99 [0.92, 1.05]	0.85 [0.81, 0.90]	0.79 [0.74, 0.84]	0.79 [0.74, 0.85]	0.81 [0.75, 0.87]
Pooled spread	0.87 [0.80, 0.93]	0.77 [0.69, 0.83]	0.83 [0.74, 0.90]	0.89 [0.80, 0.98]	0.86 [0.77, 0.94]
d'	1.14 [1.02, 1.25]	1.11 [1.00, 1.23]	0.96 [0.84, 1.07]	0.89 [0.77, 0.99]	0.95 [0.83, 1.06]

Gap and spread under the diagonal shuffle. An effect-size ratio can hide opposite changes in its numerator and denominator, so we split it. Under the diagonal shuffle the class gap of the loss falls to 0.80 ([0.77, 0.83]) of the real one on A, 0.81 ([0.78, 0.83]) on L and 0.87 ([0.80, 0.94]) on C, and that of the collision rate alike (0.82, 0.81, 0.87). On A the pooled spread falls almost as much (0.86), which is why the effect-size ratio stays at 0.93; on L and C the spread barely moves (0.96 and 0.95), so the loss of gap shows in d'. Between an eighth and a fifth of the class gap therefore depends on the link between an anchor's positive and its own rivals, on all three checkpoints; on L the other two nulls that break that link, diagonal resample and independent rivals, lose a similar share of the effect size (Table 19). The symmetric negative swap, which keeps the link and changes only the class of the rivals, leaves the gap (0.97, 0.97 and 0.97) and the spread within 7% of the real ones. This is the sense in which (A2) holds only approximately (Section 4): the rivals carry no class information, but how a pair's positive sits among its own rivals carries part of the signal.

Negative tails. Exchanging the class of every rival changes the large-batch effect size by 4% (Table 2); the negative-similarity tails of the two classes therefore carry no detectable membership signal.

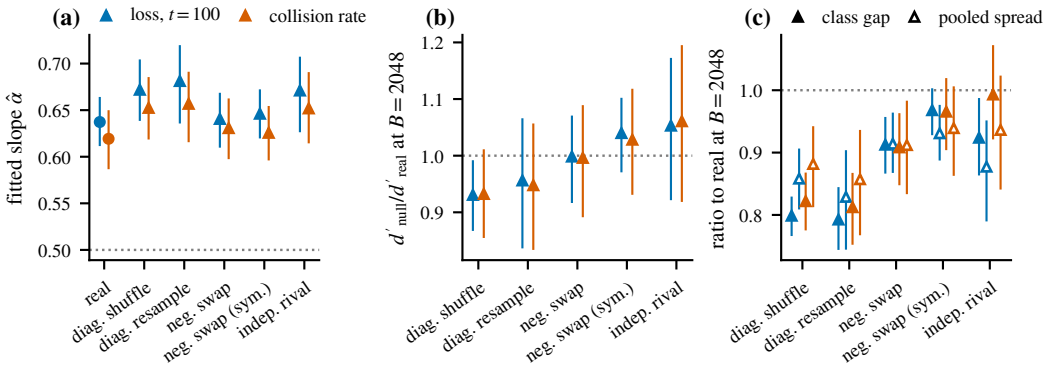


Figure 13: **Permutation nulls**, checkpoint A, InfoNCE loss at $t = 100$ (blue) and collision rate (vermilion); bars are 95% half-sample intervals. **(a)** Fitted slope on $B = 64$–2048 for the real matrix (circles) and each null (triangles); dotted: $\frac{1}{2}$. **(b)** Effect size at $B = 2048$ relative to the real matrix; line at 1. **(c)** Class gap (filled) and pooled spread (open) at $B = 2048$ relative to the real matrix; line at 1.

L Relation to U-statistics

The batch-mean loss is a symmetric function of the B pairs of a batch, so its variance admits a Hoeffding expansion $c_1/B + c_2/B^2 + \dots$ (Hoeffding, 1948). An exponent between $\frac{1}{2}$ and 1 could in principle indicate a degenerate first-order term, that is, a signal carried by pairwise rather than marginal structure. The permutation nulls of Section 4 rule this out here: the signal is marginal, and the exponent reflects the near-tie mechanism of Section 3.1.

M Tail analyses

Robust aggregation removes or exaggerates the departure from $\frac{1}{2}$. Replacing the batch mean by the batch median removes the temperature-dependent departure of the slope from $\frac{1}{2}$: the median's slopes are near $\frac{1}{2}$ at every $t \le 100$, with wide intervals that all contain $\frac{1}{2}$ and no trend in t, while the mean's slope rises from 0.42 at $t = 20$ to 0.64 at $t = 100$ (Table 21, Figure 15a). At $t = 150$ the batch median is zero in both classes at small B, because most pairs have a loss near zero at low temperature, and its fit is undefined. The median pays for this with most of the effect size (Figure 14). The 10% trimmed mean does not remove the departure but exaggerates it: its slope exceeds the batch mean's at $t \ge 100$, with disjoint intervals at $t = 150$. Trimming discards the near-tie tail, which at small B holds the whole signal, but at large B the near-tie population outgrows

the trimmed fraction and the trimmed mean recovers most of the effect; a statistic that starts lower and ends near the mean has a steeper log–log slope. Both behaviors say the same thing: the signal lies in the tail that robust statistics discard, and how much of that tail survives depends on B. The stronger evidence for localization is the top-decile sufficiency of Figure 14 and the margin analysis of Figure 15b.

Table 21: **Fitted slope of robust batch statistics of the per-pair InfoNCE loss**, checkpoint A, window $B = 64$–2048, 95% half-sample intervals. †At $t = 150$ the batch median is zero in both classes at small B, so its log–log fit is undefined.

Statistic	$t = 20$	$t = 33$	$t = 50$	$t = 75$	$t = 100$	$t = 150$
Batch mean	0.42	0.45	0.55	0.61	0.64	0.66
95% CI	[0.40, 0.44]	[0.43, 0.47]	[0.53, 0.57]	[0.59, 0.64]	[0.61, 0.66]	[0.63, 0.68]
Trimmed mean (10% each end)	0.43	0.44	0.51	0.61	0.69	0.85
95% CI	[0.41, 0.45]	[0.41, 0.46]	[0.49, 0.54]	[0.58, 0.64]	[0.65, 0.73]	[0.79, 0.90]
Batch median	0.49	0.47	0.47	0.47	0.49	undefined†
95% CI	[0.46, 0.52]	[0.43, 0.51]	[0.42, 0.51]	[0.41, 0.53]	[0.42, 0.56]	—

Top-decile sufficiency. The top decile of per-pair losses in each batch retains 97–99% of the full-batch effect size at every B, while the batch median retains a fraction of it (Figure 14).

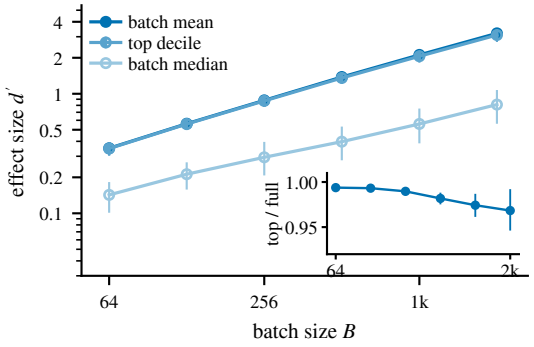


Figure 14: **The top decile of per-pair losses carries the effect size**, checkpoint A, $t = 100$. Effect size against batch size of the batch-mean loss, of its top-decile mean, and of the batch median; bars: 95% half-sample CIs; inset: ratio of the top-decile to the full-batch effect size.

Margins. Widening the crossing criterion to near-misses, $\max_{j \neq i} c_{ij} > c_{ii} - \epsilon$, dilutes the signal monotonically: the fitted slope falls from 0.62 at $\epsilon = 0$ to 0.58 at $\epsilon = 0.02$ and 0.43 at $\epsilon = 0.05$, and so does the effect size at every B (Figure 15b). The strict crossing count tracks the batch-mean loss at the trained temperature.

N Detector details

Directions. Scoring only one retrieval direction lowers the effect size by a factor of 1.13 relative to the symmetric score, and the two single directions are equally informative (Figure 16). For two directions with equal gaps and variances and correlation ρ, the symmetric score improves d' by $\sqrt{2/(1+\rho)}$, so the observed gain implies $\rho \approx 0.56$, as expected when both directions read the alignment of the same pair; the Pearson correlation of the two per-pair indicators is 0.56.

What the crossing indicator reads. The indicator $\mathbf{1}[s_{\max} > c_{ii}]$ can register a class difference through c_{ii}, the intended alignment signal, or through $s_{\max}$, the similarity to the nearest bank vector. Scoring the two terms separately with the neutral bank of the main text shows that $s_{\max}$ alone carries a small positive contribution whose interval includes or touches zero at every set size (Table 23). The provenance of the bank matters little: a member-only or a non-member-only bank moves the per-pair effect size of the crossing score by a few percent in opposite directions, within the half-sample

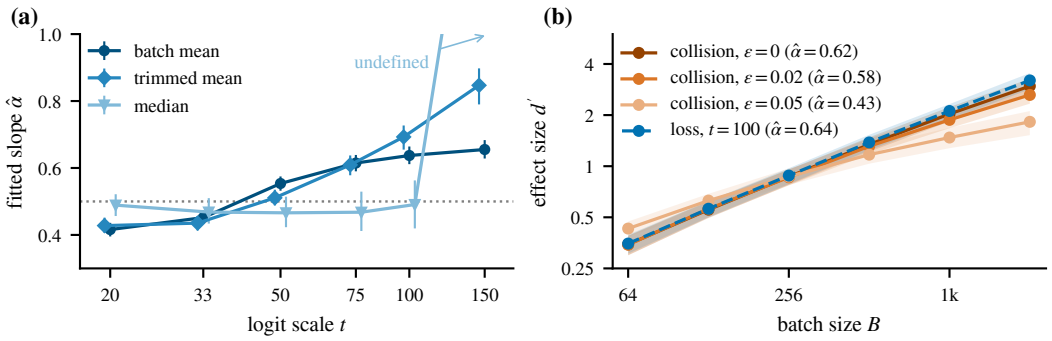


Figure 15: **Robust statistics and margins**, checkpoint A. **(a)** Fitted slope against logit scale t for the batch mean (dark blue circles), the 10% trimmed mean (diamonds), and the batch median (light blue triangles) of the per-pair loss, with 95% half-sample intervals; dotted: $\frac{1}{2}$. The median's point at $t = 150$ lies outside the axis range and is annotated with its unstable fit and interval. **(b)** Effect size against batch size for the collision rate with margin $\epsilon \in \{0, 0.02, 0.05\}$ (vermilion, darker for smaller ϵ, solid with circles) and for the batch-mean loss at $t = 100$ (blue, dashed); bands: 95% half-sample intervals; the legend gives each statistic's fitted slope.

Table 22: **Detector comparison at two set sizes** (checkpoint A), with the true-positive rate on L and C at $T = 1024$ (unique pairs); the AUC columns at $T = 1024$ are Table 3 of the main text. AUC and true-positive rate (TPR) at 1% false-positive rate for member versus non-member sets of T pairs; all scores use the same sets, and the fixed-bank scores use the same neutral bank of $N_0 = 4096$ pairs. Reference quantiles come from held-out non-member shards. For A, TPR is read from $10{,}000$ sets per class, and the rank, lower-tail, ψ_N and collision rows come from a separate 5-partition run (Appendix D).

	A, $T = 256$		A, $T = 1024$		L	C
Score	AUC	TPR	AUC	TPR	TPR	TPR
No rivals						
Mean alignment $\bar{A}$ (Ko et al., 2023)	0.705	5%	0.858	18%	52%	7%
Rank score against the reference	0.666	3%	0.808	12%	42%	7%
Lower-tail mean, $q = 0.2$	0.837	15.7%	0.977	67.3%	88%	16%
Expected crossing ψ_N, $N = 4096$	0.826	14.1%	0.971	60.4%	95%	19%
Fixed rivals						
Crossing score $\bar{S}_T$	0.784	9%	0.942	43%	82%	13%
InfoNCE loss (trained t)	0.816	13%	0.964	57%	82%	12%
In-batch rivals						
InfoNCE loss (trained t)	0.735	7%	0.933	40%	57%	8%
Collision rate $\bar{C}_B$	0.734	4%	0.927	34%	69%	10%

intervals (Table 24), which is why the main text quotes the neutral bank's 0.069. Relative to mean alignment, the crossing score's set-level effect size is $1.52\times$ with the neutral bank, $1.41\times$ with the member-only bank and $1.50\times$ with the non-member-only bank, at every batch size. None of this affects the exponents, which follow from each pair's indicator being independent of T. The blind model gives $+0.05$ for the $s_{\max}$ term on checkpoint A's pools and -0.61 on checkpoint B's, so the rival term reads pool provenance where the pools differ, in the direction that lowers checkpoint B's membership estimates (Section 6); on L's and C's splits it gives -0.29 and -0.38, smaller and in the same direction.

Bank curation and operating point. Figure 18 compares four bank-selection rules at $T = 1024$. Stacking the bank with high-crossing "hub" vectors (Radovanović et al., 2010) is worse than random selection at small banks and converges to it from $N_0 = 4096$ on, and ranking vectors by near-tie density, the Fisher-information criterion for a small shift in c, is never better than random selection and is worse at small banks. Diversity-aware near-tie selection, by greedy submodular coverage, lies

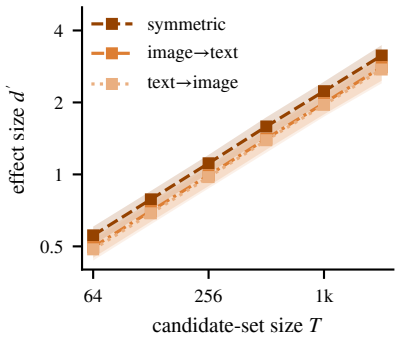


Figure 16: **Direction ablation**, checkpoint A, neutral bank of $N_0 = 4096$ pairs. Effect size of the bank crossing score against candidate-set size T for the symmetric score (dark vermilion), image→text only, and text→image only (lighter), all dashed with squares; bands: 95% half-sample intervals. The two single directions coincide and the symmetric score lies above both by a constant factor.

Table 23: **The rival term alone**, checkpoint A, neutral bank of $N_0 = 4096$ pairs: signed effect size of the set mean of $s_{\max}$ (positive means non-member sets are closer to the bank) against candidate-set size T, 95% half-sample intervals.

T	64	128	256	512	1,024	2,048
Signed d' of $s_{\max}$ alone	+0.05	+0.07	+0.10	+0.14	+0.20	+0.29
95% CI	[-0.00, 0.10]	[-0.01, 0.14]	[-0.01, 0.20]	[-0.01, 0.29]	[-0.02, 0.41]	[-0.03, 0.57]

Table 24: **Bank provenance**, checkpoint A: per-pair effect size δ of the symmetric crossing score and mean crossing rate $\bar{p}$ against banks of $N_0 = 4096$ held-out pairs of three provenances, mean over five bank draws, 95% half-sample intervals.

Bank provenance ($N_0 = 4096$)	per-pair δ	$\bar{p}$
Neutral (half held-out member, half held-out non-member)	0.069 [0.062, 0.075]	0.108
Member-only	0.068 [0.062, 0.075]	0.108
Non-member-only	0.071 [0.064, 0.077]	0.108

within the intervals of random selection at every bank size and never exceeds the random ceiling; it reaches the confidence band of that ceiling with a $4\times$ smaller bank on the coarse grid of the figure and a $2\times$ smaller bank against the ceiling of the fine grid of Figure 17b. The operative failure is redundancy: any per-vector ranking concentrates the bank on one subpopulation of anchors. All four rules peak between $N_0 = 2048$ and 8192 and decline by 16,384 as the crossing rate grows. Our bank pool, the union of the two held-out pools, holds 102,961 vectors with a maximum per-vector crossing rate of 1.3% of the evaluation pairs, so larger pools with stronger hubs may behave differently.

Calibration of the detector test equation 5. Table 25 reports the rejection rate of the one-sided test over 10,000 draws per cell, each cell reading nominal 1% / nominal 0.1%. In the *random-assignment* design, each draw samples $T+R$ pairs without replacement from the whole non-member evaluation pool and assigns T to candidates and R to references at random; the bank is fixed, so the test is valid conditionally on it. Every score is at nominal at every set size, and the power against all-member candidates matches the sample-size formula of Section 5 to inside its Wilson interval: the crossing score rejects 1.0%, 1.0% and 0.9% of the time at nominal 1% for $T = R = 1024$, 4096 and 16,384, and its power at $T = 4096$ is 79.6% against the formula's 79.3%. The *fixed-split* design is the one an auditor is tempted to use: candidates always from one half of the pool, references from the other. A single random halving leaves a per-pair offset of 0.007 standard deviations for the crossing score and 0.011 for mean alignment, a 1.5 to 2.4 standard-error draw, and that offset alone multiplies the type-I error with T, exactly as $1 - \Phi(z_{1-\alpha} - \epsilon\sqrt{T/(1+T/R)})$ predicts. It is

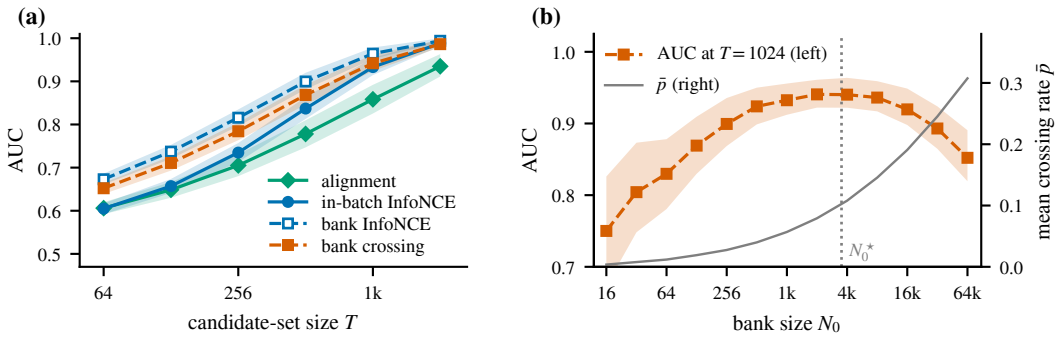


Figure 17: **The fixed-bank crossing detector.** Checkpoint A. **(a)** AUC of member versus non-member sets against candidate-set size T for the four scores of Table 3; bands are 95% half-sample CIs. **(b)** AUC at $T = 1024$ of the crossing score and mean crossing rate $\bar{p}$ (right axis) against random-bank size N_0; dotted: $N_0^\star$.

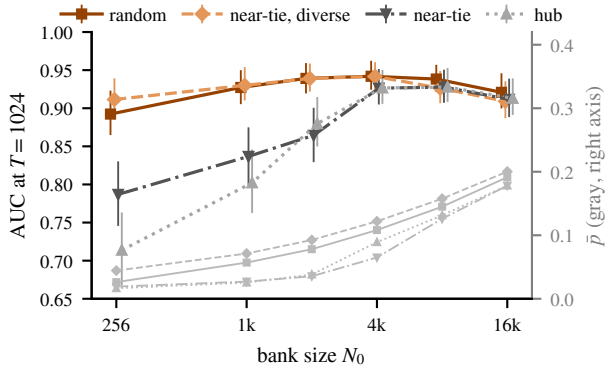


Figure 18: **Bank-selection rules**, checkpoint A. AUC of member versus non-member sets of $T = 1024$ pairs (left axis, with 95% half-sample intervals) and mean crossing rate $\bar{p}$ (gray, right axis) against bank size N_0 for random selection (vermilion squares, solid), diversity-aware near-tie selection (light vermilion diamonds, dashed), near-tie density ranking (dark gray triangles, dash-dotted), and hub selection (light gray triangles, dotted). Random and diverse near-tie selection coincide within their intervals at every N_0; the two per-vector rankings fall below them at small banks and converge at $N_0 \geq 4096$.

the paper's own warning about reference sets, measured on its own data: the reference set, not the statistic, is what breaks an audit.

The rival-free family. Tables 26 and 27 score every detector of Section 5 on all four checkpoints: the per-pair effect size in units of mean alignment, the set-level AUC, and the true-positive rate at a 1% false-positive rate. Reference quantiles Q_{ref} and F_{ref} come from held-out non-member shards. Checkpoints A, L and C rank the family in one order and B in the opposite one. On A, L and C, whose shifts are concentrated in the low-alignment tail, every statistic that reads only the lowest fifth has a larger per-pair effect than the mean, significantly so except the sparsest threshold ($q = 0.05$) on C, whose interval reaches 0.97, and the best score is the lower-tail mean on A and the expected crossing score on L and C; on checkpoint B, whose shift is uniform, the same statistics fall below the mean and no score separates significantly better than it. In units of mean alignment, every checkpoint-A score weighting the lowest fifth lands between 1.39 ($[1.26, 1.55]$) and $1.81\times$ ($[1.64, 2.01]$), every score reading the whole distribution at or below it (rank score $0.80\times$, $[0.75, 0.83]$), and the best detector needs no rivals at all: which quantiles of c_{ii} a score reads matters more than its form.

Table 25: **Calibration and power of the one-sided test**, checkpoint A only, 10,000 draws per cell, neutral bank of $N_0 = 4096$ pairs for the fixed-bank scores. Upper block: candidates and references both drawn from the non-member pool; a calibrated test rejects at the nominal rate. Lower block: member candidates against non-member references; the number in parentheses is the power predicted from the per-pair effect size of each score.

Score	$T = R = 1{,}024$	$T = R = 4{,}096$	$T = R = 16{,}384$
Random assignment from one pool: non-member candidates			
Crossing score	1.0 / 0.11	1.0 / 0.12	0.9 / 0.10
InfoNCE, fixed bank	0.8 / 0.06	1.0 / 0.10	1.1 / 0.10
Mean alignment	1.1 / 0.07	1.2 / 0.17	1.0 / 0.06
InfoNCE, in-batch	0.9 / 0.06	0.9 / 0.09	1.0 / 0.16
Member candidates (formula in parentheses)			
Crossing score	22.7 (22.5) / 6.51 (6.43)	79.6 (79.3) / 52.02 (52.06)	100.0 (100.0) / 99.97 (99.93)
InfoNCE, fixed bank	30.2 (29.8) / 9.78 (9.76)	89.8 (89.7) / 69.42 (69.14)	100.0 (100.0) / 100.00 (100.00)
Mean alignment	10.8 (10.7) / 2.25 (2.23)	44.1 (43.5) / 17.66 (17.68)	98.1 (97.7) / 89.82 (89.15)
InfoNCE, in-batch	20.1 (20.3) / 4.95 (5.53)	84.6 (84.3) / 59.92 (59.56)	100.0 (100.0) / 100.00 (100.00)
Fixed split of the pool: non-member candidates			
Crossing score	1.5 / 0.07	2.0 / 0.26	3.3 / 0.32
InfoNCE, fixed bank	1.4 / 0.14	2.1 / 0.28	4.0 / 0.31
Mean alignment	1.7 / 0.24	3.4 / 0.43	8.0 / 1.13
InfoNCE, in-batch	1.5 / 0.17	3.0 / 0.35	9.1 / 1.30

Table 26: **Rival-free and fixed-rival detectors on checkpoints A and B.** Per-pair effect size relative to mean alignment with 95% half-sample intervals, set-level AUC, and true-positive rate at a 1% false-positive rate; checkpoint A at $T = 1024$ and checkpoint B at $T = 2048$. ‡For in-batch scores the pairs of a set are not independent, so the ratio is formed from the set-level moment estimator instead of the pooled per-pair one.

	Checkpoint A ($T = 1024$)			Checkpoint B ($T = 2048$)		
Score	$\delta/\delta_{\text{mean}}$	AUC	TPR@1%	$\delta/\delta_{\text{mean}}$	AUC	TPR@1%
Mean alignment $-c_{ii}$	1.00 [1.00, 1.00]	0.863	18%	1.00 [1.00, 1.00]	0.825	18%
Rank score $1 - F_{\text{ref}}(c_{ii})$	0.80 [0.75, 0.83]	0.808	12%	1.04 [1.00, 1.08]	0.834	18%
Trimmed mean (10% each end)	0.78‡ [0.73, 0.81]	0.801	10%	1.05‡ [0.98, 1.10]	0.836	19%
Threshold, $q = 0.05$	1.56 [1.40, 1.76]	0.954	46%	0.39 [0.20, 0.57]	0.642	3%
Threshold, $q = 0.10$	1.54 [1.39, 1.73]	0.954	45%	0.48 [0.31, 0.65]	0.680	5%
Threshold, $q = 0.20$	1.39 [1.26, 1.55]	0.936	39%	0.75 [0.60, 0.90]	0.762	9%
Threshold, $q = 0.50$	0.39 [0.28, 0.48]	0.661	4%	0.89 [0.77, 1.02]	0.801	13%
Lower-tail mean, $q_0 = 0.2$	1.81 [1.64, 2.01]	0.977	67%	0.57 [0.42, 0.74]	0.706	6%
Lower-tail mean, $q_0 = 0.5$	1.54 [1.43, 1.67]	0.953	48%	0.87 [0.77, 0.97]	0.791	13%
Expected crossing, pooled rivals, $N = 4096$	1.73 [1.58, 1.91]	0.971	60%	0.96 [0.87, 1.07]	0.816	16%
Expected crossing, own rivals, $N = 4096$	1.69 [1.55, 1.87]	0.966	59%	0.96 [0.86, 1.06]	0.813	16%
Bank crossing score, $N_0 = 4096$	1.45 [1.32, 1.61]	0.946	43%	0.86 [0.75, 0.98]	0.797	12%
InfoNCE loss, fixed bank, $t = t_{\text{train}}$	1.66 [1.51, 1.84]	0.965	57%	0.92 [0.83, 1.01]	0.804	15%
InfoNCE loss, in-batch, $t = t_{\text{train}}$	1.34‡ [1.22, 1.60]	0.932	40%	0.89‡ [0.79, 1.02]	0.797	14%
Collision rate, in-batch	1.30‡ [1.14, 1.54]	0.927	34%	0.83‡ [0.72, 1.01]	0.782	13%

Pooled rivals against an anchor's own. The expected crossing score $\psi_N(c_{ii}) = 1 - F(c_{ii})^N$ can be evaluated with F estimated once from the pooled held-out rivals (d_1) or, exactly, from each anchor's own bank through the rank of c_{ii} among those rivals (d_2). Under (A2) the two agree. Table 28 sweeps N on checkpoints A and B: the difference $d_2 - d_1$ never reaches 0.01 in per-pair effect size, and every interval that excludes zero does so in the direction that makes the own-rival version *worse*, except at $N = 16$ on checkpoint B, where both versions are far below any operating point. Under (A1)–(A2) the expectation of the crossing score over random banks of N held-out pairs is thus a function of c_{ii} alone, computable from each candidate's rank among the held-out rivals, and averaging out the bank draw raises the per-pair effect size by $1.16\times$ ($[1.11, 1.22]$ at $N = 4096$ on checkpoint A) over the Monte-Carlo bank crossing score. An anchor's own rivals carry no information the pooled law does not, and the main text quotes the pooled version.

Table 27: **Rival-free and fixed-rival detectors on checkpoints L and C** (unique pairs, $T = 1024$), as in Table 26.

	L, unique pairs ($T = 1024$)			C, unique pairs ($T = 1024$)		
Score	$\delta/\delta_{\text{mean}}$	AUC	TPR@1%	$\delta/\delta_{\text{mean}}$	AUC	TPR@1%
Mean alignment $-c_{ii}$	1.00 [1.00, 1.00]	0.954	52%	1.00 [1.00, 1.00]	0.751	7%
Rank score $1 - F_{\text{ref}}(c_{ii})$	0.91 [0.89, 0.93]	0.938	42%	0.94 [0.87, 1.01]	0.738	7%
Trimmed mean (10% each end)	0.86‡ [0.82, 0.89]	0.927	36%	0.91‡ [0.82, 0.99]	0.731	6%
Threshold, $q = 0.05$	1.33 [1.24, 1.45]	0.987	77%	1.20 [0.97, 1.54]	0.790	11%
Threshold, $q = 0.10$	1.48 [1.37, 1.59]	0.994	87%	1.40 [1.16, 1.75]	0.828	15%
Threshold, $q = 0.20$	1.38 [1.29, 1.47]	0.989	82%	1.29 [1.10, 1.55]	0.806	12%
Threshold, $q = 0.50$	0.56 [0.51, 0.62]	0.828	15%	0.65 [0.49, 0.79]	0.669	3%
Lower-tail mean, $q_0 = 0.2$	1.45 [1.36, 1.56]	0.993	88%	1.41 [1.19, 1.72]	0.830	16%
Lower-tail mean, $q_0 = 0.5$	1.46 [1.39, 1.54]	0.993	88%	1.44 [1.29, 1.67]	0.835	17%
Expected crossing, pooled rivals, $N = 4096$	1.65 [1.56, 1.76]	0.997	95%	1.56 [1.35, 1.89]	0.854	19%
Expected crossing, own rivals, $N = 4096$	1.70 [1.61, 1.82]	0.998	96%	1.49 [1.27, 1.81]	0.843	17%
Bank crossing score, $N_0 = 4096$	1.37 [1.27, 1.48]	0.990	82%	1.31 [1.09, 1.61]	0.809	13%
InfoNCE loss, fixed bank, $t = t_{\text{train}}$	1.32 [1.23, 1.42]	0.987	82%	1.21 [1.01, 1.50]	0.794	12%
InfoNCE loss, in-batch, $t = t_{\text{train}}$	0.98‡ [0.87, 1.13]	0.951	57%	0.99‡ [0.79, 1.25]	0.749	8%
Collision rate, in-batch	1.18‡ [1.05, 1.37]	0.977	69%	1.07‡ [0.89, 1.40]	0.766	10%

Table 28: **Expected crossing score with pooled and with own rivals.** Mean crossing rate $\bar{p}$ and per-pair effect size δ against the number of rivals N, with the difference between the own-rival and the pooled-rival version and its 95% half-sample interval.

	N	16	256	2,048	4,096	16,384	65,536
A	$\bar{p}$, own rivals	0.004	0.027	0.078	0.108	0.191	0.307
	δ, pooled rivals (d_1)	0.050	0.077	0.085	0.083	0.071	0.052
	δ, own rivals (d_2)	0.050	0.068	0.081	0.081	0.070	0.048
	$d_2 - d_1$	+0.000	-0.009	-0.004	-0.002	-0.001	-0.004
	95% CI	[-0.005, 0.005]	[-0.013, -0.005]	[-0.006, -0.001]	[-0.004, 0.000]	[-0.004, 0.001]	[-0.007, -0.001]
B	$\bar{p}$, own rivals	0.101	0.262	0.411	0.465	0.577	0.687
	δ, pooled rivals (d_1)	0.018	0.027	0.029	0.029	0.030	0.029
	δ, own rivals (d_2)	0.021	0.026	0.028	0.029	0.029	0.025
	$d_2 - d_1$	+0.003	-0.000	-0.001	-0.000	-0.001	-0.004
	95% CI	[0.001, 0.005]	[-0.002, 0.001]	[-0.002, 0.001]	[-0.002, 0.001]	[-0.003, 0.001]	[-0.007, -0.001]

Permuting the positives across candidates. If the crossing score's advantage over mean alignment came from the coupling between an anchor and the rivals it happens to face, destroying that coupling would destroy the advantage. Permuting c_{ii} across candidates while each pair keeps its own bank maximum does not: the advantage over mean alignment stays at 0.95 of its value ($[0.90, 1.00]$; Table 29); repeating the permutation with the crossing rate matched to the real one gives the same answer. The advantage is a property of where the shift sits in the distribution of c_{ii}, not of any pair's own rivals.

Table 29: **Permuting c_{ii} across candidates.** Per-pair effect size δ of mean alignment and of the bank crossing score on the real assignment and after permuting c_{ii} across candidates with each pair's bank maximum held fixed, their ratios, and the ratio of the two; 95% half-sample intervals. "Matched rate" repeats the permutation with the mean crossing rate matched to the real one. Neutral bank of $N_0 = 4096$ pairs.

		δ, alignment	δ, crossing (real)	δ, crossing (shuffled)	ratio, real	ratio, shuffled	shuffled / real
A	plain	0.048 [0.042, 0.054]	0.069 [0.063, 0.076]	0.066 [0.062, 0.069]	1.45 [1.32, 1.61]	1.38 [1.30, 1.45]	0.95 [0.90, 1.00]
	matched rate			0.065 [0.061, 0.068]		1.35 [1.28, 1.43]	0.93 [0.88, 0.98]
B	plain	0.030 [0.024, 0.036]	0.026 [0.020, 0.032]	0.027 [0.024, 0.029]	0.86 [0.75, 0.98]	0.89 [0.81, 0.97]	1.03 [0.94, 1.13]
	matched rate			0.027 [0.025, 0.029]		0.90 [0.82, 0.97]	1.04 [0.95, 1.13]

O THREAT MODEL DETAILS AND PRE-REGISTRATION

Dilution. With a fixed bank the per-pair statistic does not depend on the rest of the candidate set, so a set with member fraction π has expected score $\mu_{\text{non}} + \pi(\mu_{\text{mem}} - \mu_{\text{non}})$, linear in π by construction, and $d'_{\text{set}} = \pi\delta\sqrt{T}$. The additional variance term $\pi(1-\pi)\delta^2$, in units of the per-pair variance, is at most $\delta^2/4$, about a tenth of a percent at $\delta = 0.069$. The required T therefore scales as $1/\pi^2$, and π is not identifiable from a positive result without calibration against a known-member set. In the weak-signal regime the per-example log-likelihood ratio is approximately affine in the per-example score, so no statistic can beat $\pi\delta\sqrt{T}$ when δ is homogeneous. If memorization is concentrated on a few examples, a top-k statistic can beat the mean under dilution. The per-pair shift is not homogeneous on checkpoints A, L and C (Section 4), which is why lower-tail statistics beat the mean there; under dilution the same holds within the member subpopulation, and the optimal statistic is the likelihood ratio, which an auditor cannot estimate without known members.

Duplicate sensitivity. Duplicated images are the obvious artefact that manufactures ties, so every statement above is repeated on unique pairs only. Flagging pairs whose image hash or exact caption occurs more than once within a class, 2.8% and 3.1% of checkpoint A's pools and 9.0% and 9.1% of checkpoint B's are duplicated by image (Appendix B). On unique pairs every signature of Section 4 survives and strengthens slightly on checkpoint A: the paired slope difference is 0.13 ($[0.10, 0.16]$, against 0.12 on all pairs), the crossing-to-alignment ratio 1.54, the in-batch loss slope 0.66, and the shift-function contrast 2.65 ($[2.21, 3.25]$). The flagged pairs themselves carry no such advantage (ratio 0.98), so the near-tie signal is not a duplicate artefact. On checkpoint B the unique-pair values are unchanged (0.04, ratio 0.86, contrast 0.28). On checkpoints L and C, whose 2026 downloads contain placeholder images, the same rule with the hash taken over the downloaded bytes flags 5.1% and 3.4% of the pairs; there the flagged pairs weaken every unbounded statistic, so unique pairs are the primary analysis, with all pairs reported beside them (Appendix C).

Adversarial trimming. Removing the lowest-aligned fraction q of a Gaussian population shifts the mean of the remainder by $\phi(\Phi^{-1}(q))/(1-q)$ standard deviations: 0.033 for $q = 1.25\%$. At $T = 10^4$ a shift of $3.2/\sqrt{T} = 0.032$ standard deviations yields $d' = 3.2$, a $p < 10^{-3}$ verdict equivalent to $\pi \approx \frac{1}{2}$, although every remaining item has genuine non-member provenance.

Pre-registration protocol. An external audit on an unknown set is sound only if the auditor (1) defines the candidate set by a provenance rule that can be evaluated without the model under audit; (2) publishes a hash of the sorted content identifiers and their count, timestamped independently; (3) registers the set size, statistic, bank provenance and size, reference pool, and decision threshold together, since each of them moves the verdict; (4) applies attrition (dead links, placeholder images, decoding failures, deduplication) mechanically and reports $N_{\text{registered}} \to N_{\text{evaluated}}$; (5) reports sensitivity to composition by subsampling and by removing the top and bottom deciles of alignment; and (6) reports the same test against a blind model trained on neither pool. The checkpoints in this paper do not need this protocol, because membership is known from the training pipeline and the splits were fixed by provenance in advance; the one analysis choice made after seeing results, unique pairs for L and C, is reported beside its alternative (Appendix C).

P THE LADDER: PRE-REGISTERED PREDICTIONS AGAINST MEASUREMENTS

This appendix holds the full record of the prospective test summarized in Section 6: what was registered, the agreement criterion attached to each quantity, the outcome on the 40M rung, and the 4M rung as a boundary probe. The two rungs share checkpoint A's recipe and pools and vary only the number of unique training pairs; their class gaps follow the $O(\kappa/n)$ bound with slope one, that is, ten times fewer pairs give a ten times larger gap (Appendix A, Figure 5). The 40M rung keeps A's shift profile (contrast 2.20 [2.16, 2.25] against A's 2.49 [2.10, 3.00]; Figure 4a). Throughout, Kendall's rank correlation is written τ_K; the subscript distinguishes it from the logit scale t.

What was registered, and when. Before the first ranking quantity was measured on either rung, we committed to the supplementary repository (i) the predictions of the near-tie model for the 40M rung, (ii) one agreement criterion per predicted quantity, and (iii) for the 4M rung the expected

direction and size of the model's failure, with no criterion. The predictor takes three label-free inputs of the 40M checkpoint, its non-member c_{ii} law, its rival law and its class gap Δc (a first moment), plus one input from checkpoint A: the quantile shift shape $\Delta_A(q)/\overline{\Delta_A}$, rescaled to the 40M gap. No quantile of the 40M member distribution was computed before registration. The uniform-shift alternative replaces the shape by a constant and was registered beside the primary prediction so that a miss could be attributed to the shape rather than to the mechanism. Run with A's own inputs and A's own shape, the predictor reproduces the counterfactual model of Appendix I; this validation was done before registering.

Registered criteria. The register fixes, for each quantity, the rule that separates "pass" from "miss"; Table 31 prints the rule beside every row, and in full it reads as follows. *Bank optimum:* pass if the measured N_0^* lies within one octave of the prediction, 3,989; strong pass if within half an octave. Near its peak $\delta(N_0)$ is flat, so the tolerance is set against the resolution of the optimum itself: checkpoint A's own 95% interval for $N_0^\star$ (1,879–7,510) spans two octaves. *Crossing rate:* pass if $\bar{p}^*$ is within ± 0.05 of 0.16. *Near-tie gain*, $\delta(N_0^*)/\delta_{\text{mean}}$ and $\delta(4096)/\delta_{\text{mean}}$: pass if the predicted value lies inside the measured 95% interval, or within 15% of the measured point. *In-batch slope:* pass if $|\hat{\alpha} - 0.70| \leq 0.05$ at $t = 100$, the trained scale; the same tolerance is applied at $t = 20$, 50 and 150, which the register predicts but for which it states no separate rule. *Ordering:* report τ_K between the predicted and measured orderings of eight detectors at each T; pass if $\tau_K \geq 0.71$ (at most 4 of 28 pairs discordant) at $T = 64$ and $T = 256$; $T = 16$ and 1024 are reported, and the same rule is applied to them for information. *Shape:* pass for the preserved-shape hypothesis if the measured 40M contrast $R_{\text{low}} - R_{\text{high}}$ has an interval overlapping A's $[2.10, 3.00]$; if it does not, the misses are diagnosed by re-running the predictor with the measured 40M shape, a consistency check reported as such and never as a prediction. Two consequences of these rules are visible in the table. The contrast passes although its measured interval $[2.16, 2.25]$ excludes the point value 2.49, because the registered rule is overlap with A's interval, not agreement with A's point. The near-tie gain misses because the predicted 1.60 lies outside the measured interval $[1.30, 1.34]$ and farther than 15% from the point 1.32; the two rules differ, so the two verdicts differ.

Biases stated before measuring. The register lists three. First, the synthetic member population carries the non-member SD, and the real 40M members are tighter, so the pooled-SD moment effect sizes of alignment-type scores can exceed the predicted per-pair δ without any change of shape; absolute δ is compared after rescaling by the measured over synthetic pooled SD, and ratios, N_0^*, $\bar{p}^*$, slopes and orderings are compared as is. Second, the in-batch scores assume DEFF $= 1$; on A the loss design effect is flat in B (Appendix G), which moves the in-batch slope by less than the tolerance, and the ordering between the two in-batch scores and their neighbors is within that margin and is not a strong prediction. Third, the model locates N_0^* by a parabolic argmax on octaves, with a resolution of about half an octave, while the measurement uses a local-quadratic fit on a finer grid. The fixed-rival scores obey the square-root law by construction, so the test bears on $\delta(N)$ and on the ranking, not on their $\hat{\alpha}$.

Core measurements on the rungs. Table 30 repeats the core measurements of Table 4 for the two rungs, beside A and B: the loss-side signatures (the in-batch minus fixed-bank slope differences for the loss and the crossing indicator, the symmetrization gain and the negative swap) appear on both rungs.

Outcome on the 40M rung. The bank optimum, its crossing rate, the in-batch loss slope at all four t and the contrast pass under the registered rules, and the eight detectors of Appendix N appear in the predicted order at $T = 16$ ($\tau_K = 1.00$), with $\tau_K = 0.93$ and 0.86 at the two registered T and 0.71 at $T = 1024$. The optimum $N_0^* = 6{,}109$ [5,411, 6,734] lies within one octave of 3,989 but not within half an octave, so it is a pass and not a strong pass. The near-tie gain misses: $\delta(N_0^*)/\delta_{\text{mean}} = 1.32$ [1.30, 1.34] against a predicted 1.60, and $\delta(4096)/\delta_{\text{mean}} = 1.31$ against 1.60 in the same direction. Driven by the rung's own measured shape, the consistency check gives 1.37 at the optimum and 1.37 at $N_0 = 4096$, both within the register's 15% tolerance of the measured points, so the miss is the residual shape difference between the rungs (lowest-quintile ratio 2.30 on 40M against 2.72 on A) and not the aggregation mechanism. The uniform-shift alternative puts the optimum at the grid boundary ($N_0^* = 65{,}536$, $\bar{p}^* = 0.44$), the gain below one (0.85) and the mean-type scores first at every T; the measurements reject all three. Its in-batch loss slope at $t = 100$, 0.69, passes the

Table 30: **Core measurements on checkpoint A, its two ladder rungs and B.** Canonical estimator and fit window (Section 2); differences within half-sample replicates, 95% CIs; detector: bank crossing score, AUC at $T = 2048$; blind rows (a model that saw neither pool, signed $d' > 0$ in the membership direction) at $T = 2048$; the rungs share A's pools, so their blind rows are A's. Control row: checkpoint B scored on A's pools.

Quantity	A	40M rung	4M rung	B
$\hat{\alpha}$, InfoNCE loss, in-batch	0.64	0.68	0.66	0.57
$\hat{\alpha}$, crossing indicator, in-batch	0.62	0.65	0.68	0.54
$\hat{\alpha}$, crossing indicator, fixed bank	0.50	0.50	0.50	0.50
Difference, in-batch − bank (loss)	0.14 [0.11, 0.16]	0.18 [0.17, 0.19]	0.16 [0.15, 0.16]	0.07 [0.03, 0.10]
Difference, in-batch − bank (crossing)	0.12 [0.08, 0.15]	0.15 [0.14, 0.17]	0.18 [0.17, 0.19]	0.04 [−0.02, 0.10]
Detector AUC	0.986	1.000	1.000	0.796
Symmetrization gain in d'	1.13	1.11	1.09	1.08
Negative swap, d'/d'_{real}	1.04	1.04	1.02	0.95
Blind, full detector, signed d'	+0.36 [+0.08, +0.66]	as A	as A	+0.06 [−0.22, +0.34]
Blind, rival term only, signed d'	+0.05 [−0.24, +0.35]	as A	as A	−0.61 [−0.89, −0.35]
Control (B on A's pools), full detector	+0.70 [+0.42, +0.99]	as A	as A	—

same tolerance, as do its slopes at $t = 50$ and 150; the slope separates the two shapes only at $t = 20$. The two discriminating questions of the register, whether a rival-based score is first and whether the lower-tail mean beats mean alignment, are both answered as the preserved-shape hypothesis predicts.

The 4M rung as a boundary probe. At $\Delta c/\sigma = 2.206$ the first-order model is out of range, and the register asked only for the expected direction and size of its failure, in four items. (1) Variance before shape: the synthetic members carry the non-member SD while the real members are tighter, so the alignment-type scores should exceed their predicted effect sizes, and the lower-tail mean, clipped at the reference quantile, should collapse to last place; both occurred. (2) N_0^* and $\bar{p}^*$ should be unresolved, with an interval spanning more than two octaves that need not contain 642; the measured optimum is instead sharply located, $9{,}564$ [$8{,}962$, $10{,}356$] at $\bar{p}^* = 0.49$ against a predicted 0.39, so the failure is in the location and not in the resolution. (3) The in-batch loss slope should miss the 0.05 tolerance because the loss distributions are far from Gaussian at this gap; it holds at every t (0.66 against 0.64 at $t = 100$). (4) The ordering at the top and between the two in-batch scores should hold; the expected crossing is first and the clipped lower-tail mean last as predicted, but the bank crossing score climbs to second, and τ_K falls from 0.71 at $T = 16$ to 0.50 at $T \geq 256$. The contrast falls to 0.90 [0.90, 0.91] although Δc is the largest of the four checkpoints, because at this gap the whole member distribution has moved past the width of the non-member one: no quantile is special relative to the mean shift, and the shift function in units of the mean shift flattens by construction as the gap grows beyond the width of the distribution. Tail concentration is a small-Δc property, and the ladder crosses out of that regime between 40M and 4M.

Table 31: **Pre-registered predictions of the near-tie model for the two ladder rungs against the measurements, with the registered criterion beside each verdict.** Inputs: each rung's own non-member c_{ii} law and rival law, its measured gap Δc (a first moment), and checkpoint A's quantile shift shape $\Delta_A(q)/\overline{\Delta_A}$ rescaled to that gap ("A shape"). Predictions and criteria were committed to the supplementary repository before any ranking quantity was measured on either rung; the criterion column quotes the register in compact form. Measured values carry 95% half-sample intervals. The contrast row passes although its interval excludes the point prediction, because the registered rule is overlap with A's interval $[2.10, 3.00]$, not agreement with the point; the near-tie gain rows miss because their rule is interval membership or 15% of the point. τ_K is Kendall's rank correlation between the predicted and measured orderings of eight detectors (mean alignment, rank score, lower-tail mean $q = 0.20$, expected crossing $N = 2048$, bank crossing $N_0 = 4096$, fixed-bank InfoNCE $t = 100$, in-batch InfoNCE $t = 100$, in-batch collision) at each T. Rows with no rule at all ($\bar{p}(4096)$ and the crossing-indicator $\hat{\alpha}$) carry no verdict. For the 4M rung the register applies no criterion and asks for the direction of failure (text above); its verdict column applies the 40M rules for information only. The uniform-shift alternative (not tabulated) predicts $N_0^* = 65{,}536$, $\bar{p}^* = 0.44$, $\delta(N_0^*)/\delta_{\text{mean}} = 0.85$ and $\hat{\alpha} = 0.69$ at $t = 100$ on the 40M rung, with the mean-type scores first at every T. Consistency-check values (the model driven by each rung's own measured shape, run after the measurement) are reported in the text, never as predictions.

Quantity	Predicted (A shape)	Measured	Registered criterion	Verdict
40M rung ($\Delta c/\sigma = 0.413$)				
N_0^*, symmetric bank crossing	3,989	6,109 [5,411, 6,734]	within one octave of the prediction; strong pass within half an octave	pass
$\bar{p}^*$	0.16	0.20 [0.19, 0.21]	within ± 0.05 of the prediction	pass
$\delta(N_0^*)/\delta_{\text{mean}}$	1.60	1.32 [1.30, 1.34]	prediction inside the measured 95% interval, or within 15% of the point	miss
$\delta(4096)/\delta_{\text{mean}}$	1.60	1.31 [1.29, 1.33]	same rule as $\delta(N_0^*)/\delta_{\text{mean}}$	miss
$\bar{p}(4096)$	0.17	0.17 [0.17, 0.17]	none registered	
$\hat{\alpha}$, in-batch loss, $t = 20$	0.45	0.46 [0.45, 0.47]	$\lvert\hat{\alpha} - \text{pred}\rvert \leq 0.05$; rule registered at $t = 100$, applied at every t	pass
$\hat{\alpha}$, in-batch loss, $t = 50$	0.67	0.64 [0.63, 0.65]	as at $t = 20$	pass
$\hat{\alpha}$, in-batch loss, $t = 100$	0.70	0.68 [0.66, 0.69]	$\lvert\hat{\alpha} - \text{pred}\rvert \leq 0.05$ (registered rule)	pass
$\hat{\alpha}$, in-batch loss, $t = 150$	0.71	0.68 [0.67, 0.70]	as at $t = 20$	pass
$\hat{\alpha}$, crossing indicator, in-batch	0.68	0.65 [0.64, 0.67]	none registered	
Contrast $R_{\text{low}} - R_{\text{high}}$	2.49	2.20 [2.16, 2.25]	measured interval overlaps A's $[2.10, 3.00]$ (shape preserved)	pass
τ_K, $T = 16$	order	1.00	reported; rule $\tau_K \geq 0.71$ registered at $T = 64$ and 256, applied here for information	pass
τ_K, $T = 64$	order	0.93	$\tau_K \geq 0.71$ (≤ 4 of 28 pairs discordant)	pass
τ_K, $T = 256$	order	0.86	$\tau_K \geq 0.71$ (≤ 4 of 28 pairs discordant)	pass
τ_K, $T = 1024$	order	0.71	as at $T = 16$	pass
4M rung, boundary probe ($\Delta c/\sigma = 2.206$)				
N_0^*, symmetric bank crossing	642	9,564 [8,962, 10,356]	boundary probe (no criterion)	miss
$\bar{p}^*$	0.39	0.49 [0.48, 0.49]	boundary probe (no criterion)	miss
$\delta(N_0^*)/\delta_{\text{mean}}$	1.05	1.44 [1.43, 1.45]	boundary probe (no criterion)	miss
$\hat{\alpha}$, in-batch loss, $t = 20$	0.55	0.60 [0.59, 0.61]	boundary probe (no criterion)	pass
$\hat{\alpha}$, in-batch loss, $t = 50$	0.63	0.66 [0.64, 0.67]	boundary probe (no criterion)	pass
$\hat{\alpha}$, in-batch loss, $t = 100$	0.64	0.66 [0.64, 0.67]	boundary probe (no criterion)	pass
$\hat{\alpha}$, in-batch loss, $t = 150$	0.64	0.66 [0.64, 0.67]	boundary probe (no criterion)	pass
$\hat{\alpha}$, crossing indicator, in-batch	0.55	0.68 [0.67, 0.69]	none registered	
Contrast $R_{\text{low}} - R_{\text{high}}$	2.49	0.90 [0.90, 0.91]	boundary probe (no criterion)	miss
τ_K, $T = 16/64/256/1024$	order	0.71 / 0.64 / 0.50 / 0.50	boundary probe (no criterion)	pass/miss/miss/miss