NEAR-DUPLICATE FAMILIES BREAK EXACT-RECORD MEMBERSHIP INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

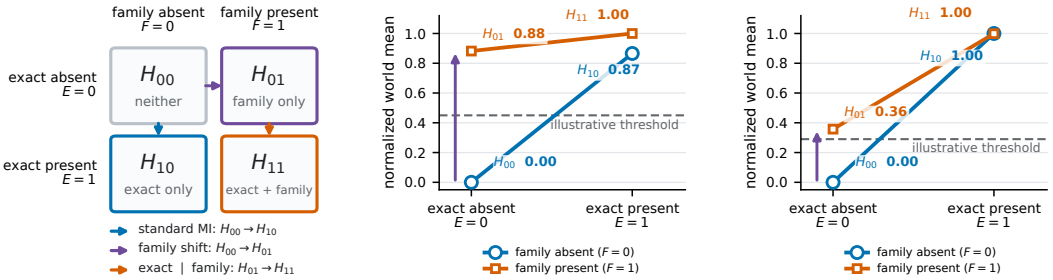
Membership inference (MI) asks whether a specific record appeared in a model's training set and is increasingly used as evidence for data provenance and copyright auditing. These applications require determining whether the exact queried record was used for training, rather than merely whether the model was exposed to similar content. Making this distinction is challenging because web-scale datasets naturally contain near-duplicates, including syndicated articles, mirrored pages, and lightly modified images. We show that this creates a fundamental confound for standard MI. A clean-reference audit typically calibrates membership against a null in which neither the queried record nor its near-duplicate family is present. In deployment, however, the queried record may be absent while a non-identical family member was used for training. We introduce a four-world audit that independently varies exact-record inclusion and family presence to separate these cases. Natural near-duplicate families cause severe false attribution. On CC-News, a clean-reference LiRA auditor labels 99.70% of family-present exact non-members as members at 1.00% false-positive rate. This failure persists across alternative scores, model architectures, and executed deduplication and retraining. Controlled interventions further reveal that the effect depends on the learning objective. In classification, faithful families largely substitute for the exact record, reducing exact-given-family inference to near chance. In autoregressive language modeling, the exact sequence retains a detectable residual, while family presence still confounds clean-reference decisions. These results show that positive model-only membership evidence may establish family-level exposure without establishing exact-record provenance.

1 INTRODUCTION

Membership inference (MI) asks whether a specific record appeared in a model's training set (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022a). While initially designed as a privacy attack, MI is increasingly repurposed as forensic evidence in copyright disputes and data provenance claims (Maini et al., 2021; Meeus et al., 2024; Ertan et al., 2026). These applications demand a rigorous standard of proof: the audit must identify the *exact* disputed record rather than merely related content. Yet, web-scale corpora inherently blur this distinction through syndicated news, mirrored pages, and minor edits (Lee et al., 2022; Kandpal et al., 2022). Standard MI ignores this reality by relying on a "clean" null hypothesis that tests exact inclusion against an idealized absence condition where no related content exists. Consequently, if a model assigns a member-like score due to the presence of non-identical near-duplicates, standard auditors risk triggering a false positive, confounding exact-record provenance with family-level exposure.

We demonstrate that this vulnerability severely compromises exact-record attribution in practice. On CC-News, natural near-duplicate families cause a clean-reference auditor to falsely label 99.70% of family-present genuine non-members as members after full fine-tuning of a Pythia-2.8B model (Biderman et al., 2023). The effect persists across alternative membership scores, including Min-K% (Shi et al., 2024), Min-K++ (Zhang et al., 2025b), and EZ-MIA (Ilic et al., 2026). It also transfers to different model and fine-tuning settings and persists after an executed MinHash deduplication-and-retraining pipeline. Similarly, natural WikiText-103 siblings independently increase the false-member rate from 1.40% to 81.00%. In both corpora, no-sibling control records remain calibrated at roughly their nominal rates (1–2%). This isolates family presence, rather than general threshold miscalibra-

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.



(a) Audit design. (b) Classification: substitution. (c) LM: residual exact signal.

Figure 1: The four-world audit isolates exact-record evidence from family presence. **(a)** Standard MI (blue, $H_{00} \to H_{10}$) tests exact inclusion against a clean absence world, ignoring the family-only shift (purple, $H_{00} \to H_{01}$); our audit instead toggles the exact record with the family already present (orange, $H_{01} \to H_{11}$). **(b)** Under classification the family substitutes, flattening the orange trajectory. **(c)** Under autoregressive modeling the exact record keeps a large additive residual, while the family shift alone (purple) already crosses the illustrative threshold (dashed). Panels (b,c): min–max-normalized pooled world means (matched AG-News/GPT-2 shadows); lines connect binary E states.

tion, as the source of the failure. A standard positive MI score can therefore provide evidence of broader content exposure without proving exact-record inclusion.

To formally isolate what membership signal, if any, survives this confounding, we introduce a four-world audit (Figure 1). This framework independently varies exact-record inclusion (E) and near-duplicate family presence (F), separating standard exact-record inference from exact-given-family inference. Applying this audit in controlled settings reveals a fundamental, objective-dependent split in the underlying mechanics. In classification tasks, faithful near-duplicate families act as substitutes for the exact example. This redundancy drives exact-record detectability down to random chance (Feldman, 2020). Conversely, autoregressive language modeling retains an additive, sequence-specific effect: a language model preserves a unique, detectable signature of the exact text even when its family is present (Carlini et al., 2021; Feldman & Zhang, 2020). This surviving signature nevertheless cannot save a standard audit: family presence alone inflates the score past any clean-calibrated threshold, triggering a false positive before the exact-sequence residual is ever weighed. Model-only audits should therefore declare their unit of inference and abstain on exact-record verdicts whenever near duplicates cannot be excluded, while still reporting family-level exposure, which remains reliably detectable (AUC 0.98 on CC-News).

This paper makes three primary contributions:

- We formulate exact-record membership as a composite hypothesis with near-duplicate family presence as a critical nuisance variable, and introduce a four-world intervention that cleanly separates exact-record evidence from family-level evidence.
- We empirically demonstrate severe natural false attribution across multiple large-scale text corpora, showing this vulnerability survives executed deduplication pipelines and transfers across different models and attack formulations.
- We identify a fundamental split in learning dynamics between substitution under classification and exposure-additive memory under language modeling, establishing that while exact provenance is empirically unidentifiable in these settings, family presence remains a robust and defensible conclusion.

2 Related Work

Membership inference and attribution. Membership inference originated as a privacy attack closely tied to model overfitting (Shokri et al., 2017; Yeom et al., 2018), and has since been strengthened by large-scale, low-FPR formulations such as LiRA (Carlini et al., 2022a) and RMIA (Zarifzadeh et al., 2024), alongside LLM-specific metrics like Min-K% (Shi et al., 2024; Zhang et al., 2025b;

Ilic et al., 2026). As these tools are increasingly repurposed for dataset inference and copyright analysis (Maini et al., 2024; 2021; Meeus et al., 2024; Ertan et al., 2026), the validity of their underlying assumptions has come under scrutiny. While recent work highlights the calibration obstacle of unsamplable production nulls (Zhang et al., 2025a), we identify a more fundamental flaw that persists even with perfect retraining access: the "absence" of a queried record is a composite state containing both family-absent and family-present conditions, which severely confounds exact attribution.

Related records and memorization. Prior work establishes that memorization is fundamentally linked to data distribution (Feldman & Zhang, 2020; Feldman, 2020; Carlini et al., 2021), with neighboring non-members often receiving member-like predictions (Rezaei & Liu, 2022). Furthermore, "privacy onion" and counterfactual memorization frameworks demonstrate that a record's leakage is deeply dependent on its surrounding training neighborhood (Carlini et al., 2022b; Zhang et al., 2023), motivating standard industry practices like corpus deduplication to mitigate extraction risks (Kandpal et al., 2022; Lee et al., 2022).

Manipulation and defenses. Optimized poisons can force arbitrary MI decisions (Mangaokar et al., 2025; Borkar et al., 2025), and active attacks like Truth Serum and Chameleon amplify leakage through mislabeled copies or adaptive poisoning (Tramèr et al., 2022; Chaudhari et al., 2024). Trainer-side defenses instead mitigate leakage through differential privacy or adversarial regularization (Abadi et al., 2016; Nasr et al., 2018; Tang et al., 2022; Chen et al., 2022; Jia et al., 2019). Unlike these active interventions, the confounding effect we evaluate arises natively without an adversary, and any defensive interpretation of our controlled experiments remains strictly secondary.

3 Near-Duplicate Families Confound Exact-Record Inference

3.1 The four-world audit framework

Standard MI treats membership as a binary question: whether the exact record was used for training. With near-duplicates, however, the observed membership signal may come from either the exact record or related family members. We therefore separate these two factors and use a four-world audit framework to distinguish three questions: standard exact-record inference, family-presence inference, and exact-record inference when the family is already present.

Let f be a model trained on dataset D, and let x be the exact record whose membership is being audited. Let $E \in \{0,1\}$ indicate whether x is present in D, and let $F \in \{0,1\}$ indicate whether a non-identical member of its near-duplicate family is present. The pair (E, F) defines four worlds, H_{00}, H_{10}, H_{01}, and H_{11}.

Standard MI compares H_{10} with H_{00}. Exact-given-family inference compares H_{11} with H_{01} and asks whether x adds detectable evidence after its family is already present. Family-presence inference compares H_{01} with H_{00} and asks whether the family itself leaves detectable evidence. When F is unknown, exact-record inference has the composite hypotheses

$$\mathcal{H}_0^E = \{H_{00}, H_{01}\}, \qquad \mathcal{H}_1^E = \{H_{10}, H_{11}\}. \tag{1}$$

A clean-reference test calibrated on H_{00} controls false positives for only one component of the null.

Proposition 1 (Clean-null calibration)**.** *Let A be any rejection region with $P_{00}(A) \le \alpha$. Its type-I error for the composite exact-absence null is*

$$\sup_{F \in \{0,1\}} P_{0F}(A) = \max\{P_{00}(A), P_{01}(A)\}, \tag{2}$$

which is not bounded by α and may equal one. If family presence among exact non-members has prevalence π, the deployment FPR is $(1-\pi)P_{00}(A) + \pi P_{01}(A)$.

Proof. Both statements follow directly by maximizing, or averaging, over the two components of $\mathcal{H}_0^E$. Calibration on H_{00} constrains only the first term, leaving the error rate on H_{01} unbounded. □

The proposition is distribution-free: it does not depend on LiRA, Gaussian scores, or a particular similarity metric. The empirical question is whether $P_{01}(A)$ differs materially from $P_{00}(A)$; our natural studies show that it can.

The four comparisons separate questions that a binary MI benchmark conflates. Standard MI measures whether x has an effect when no related content is present, but it does not identify the source of a member-like score in a corpus where F is unknown. Exact-given-family inference is the relevant controlled comparison for exact provenance: both hypotheses contain the same family and differ only in x. Conversely, family-presence inference deliberately removes x and asks whether its surrounding content is visible. This distinction also prevents a low exact-given-family AUC from being misread as content privacy; the family itself may remain highly detectable.

For a score with four-world means μ_{EF} and variances σ^2_{EF}, we define the exact record's raw conditional effect and its standardized effect as:

$$\delta_E^F = \mu_{1F} - \mu_{0F}, \qquad d_E^F = \frac{\delta_E^F}{\sqrt{(\sigma^2_{1F} + \sigma^2_{0F})/2}}. \tag{3}$$

The factorial interaction $I = \delta_E^1 - \delta_E^0$ measures how family presence shifts the exact record's absolute contribution. The ratio of the standardized gaps, $R_{\text{exact}} = d_E^1/d_E^0$, measures the retained fraction of the exact-record signal. A ratio near zero indicates near-total substitution, while a ratio near one indicates approximate additivity.

3.2 Core failure modes and baseline auditors

We distinguish two distinct failure modes under near-duplicate exposure:

- **False attribution** occurs when the presence of the family alone (H_{01}) causes model scores to cross a threshold calibrated on clean data (H_{00} vs. H_{10}), triggering a false positive for the absent exact record.
- **Erasure** occurs when the exact record's signal is completely overshadowed by the family, rendering the worlds H_{11} and H_{01} empirically indistinguishable even to an oracle auditor.

The auditor has black-box access to the target model and may train shadow models on reference datasets, but cannot inspect the deployed training set or assume knowledge of every variant in the near-duplicate family. For classification, the primary score is the true-class logit margin. For language modeling, it is negative suffix loss. Our primary calibrated auditor is LiRA (Carlini et al., 2022a), which fits parametric score densities on calibration shadows to evaluate the log-likelihood ratio:

$$\ell_x(\phi) = \log p(\phi \mid E = 1, F = f) - \log p(\phi \mid E = 0, F = f). \tag{4}$$

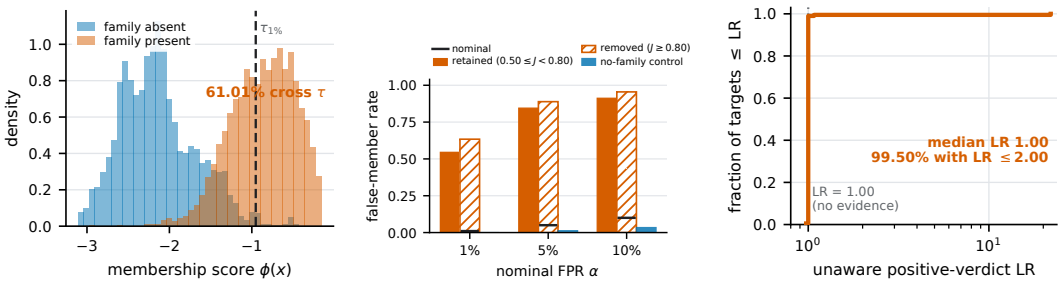
The conditioning state $F = f$ defines our two primary auditor baselines: the *clean-reference auditor* (calibrated assuming the family is absent, $F = 0$) and the *family-conditioned auditor* (calibrated knowing the family is present, $F = 1$). Standard clean-reference decision thresholds are chosen at a nominal false-positive rate α using clean OUT scores (H_{00}). False attribution then occurs precisely because the actual deployment state shifts to $F = 1$ (the family is present), blinding the clean-reference auditor to the difference.

3.3 Evaluation protocol

Our evaluation protocol follows a rigorous shadow-training pipeline. Each condition trains a fresh set of shadow models while independently randomizing each target's exact membership. Calibration shadows fit the conditional score densities and establish the decision thresholds; disjoint evaluation shadows are then used to measure false-positive and true-positive rates out of sample. We report the raw Area Under the ROC Curve (AUC), the orientation-invariant advantage $adv = 2|\mathrm{AUC} - 0.5|$, and the decision rule's positive likelihood ratio $\mathrm{LR}^+ = \mathrm{TPR}/\mathrm{FPR}$ (computed as the bootstrap-median ratio over individual resamples to handle skewness). Confidence intervals are estimated using a crossed bootstrap over both evaluation targets and shadow seeds. To prevent selection bias, vulnerable targets are identified solely on a separate set of discovery shadows.

We evaluate this protocol across both text and vision modalities:

- **Text corpora:** Near-duplicate families are mined naturally from CC-News and WikiText-103 using MinHash clustering over token 5-grams with a Jaccard similarity threshold (Lee et al., 2022; Kandpal et al., 2022).



(a) Non-member score shift. (b) Jaccard-0.8 retention bands. (c) Per-target evidence.

Figure 2: Natural CC-News/pythia-2.8B verdicts. (a) Family presence shifts pooled non-member scores: 61.01% cross a global clean 1.00%-FPR threshold (per-target calibrated FPR 99.70%). (b) False attribution stays high even in the Jaccard-0.80 dedup-retained band. (c) The unaware positive-verdict LR has median 1.00; 199/200 targets have LR $\leq$ 2.00.

- **Vision corpora:** Families are systematically constructed on CIFAR-10/100 using same-label spatial and geometric transformations of the target images.

Appendix F provides a complete map of every reported quantity to its exact construction, score, subset, and estimator.

Experimental scope. Our evaluation encompasses multiple modalities (vision, text), architectures (ResNet-9 (He et al., 2016), VGG-11 (Simonyan & Zisserman, 2015), GPT-2, Pythia-2.8B (Biderman et al., 2023), Qwen3.5-2B (Team, 2025)), and datasets (CIFAR-10/100, AG-News, WikiText-103 (Merity et al., 2017), CC-News), requiring the training of over 1,500 shadow models in total. Every intervention holds the background data and training recipe fixed, isolating the exact-record signal by independently randomizing membership. Comprehensive data construction, optimization, and compute configurations for each setting are deferred to the appendix.

4 Natural Families Cause Severe False Attribution

4.1 Widespread false attribution on CC-News

We evaluate false attribution in the wild by mining 200 natural news-syndication families and 100 no-sibling controls from CC-News. The targets span broad token-5-gram Jaccard similarities ($[0.50, 1.00)$), ensuring the audit tests non-identical, natural variation. Across 128 pythia-2.8B shadows, we establish a highly vulnerable baseline: the family-absent exact-record advantage is 0.99 (averaged per target; the pooled score AUC in Table 3 is 0.87), meaning standard MI succeeds almost perfectly in a clean world. To rigorously isolate the family effect, we score a fixed suffix of each passage, remove byte-identical copies, and evaluate against the no-sibling controls processed through the identical calibration pipeline. Because exact membership is independently randomized in every shadow, any pretraining contamination remains strictly symmetric between the IN and OUT worlds, isolating the observed failure entirely to the family-present contrast.

Severe false attribution. The clean-reference per-target LiRA auditor nevertheless labels 99.70% of family-present non-members as members at a nominal 1.00% FPR (Figure 2); the rate remains 99.70% at the 5.00% operating point used in Table 1. Its positive likelihood ratio falls from approximately 62.00 in the clean world to 1.00, because TPR and FPR become nearly equal. In the matched pooled control, no-sibling targets have only a 1.20% false-positive rate at that operating point. The effect is thus a family-specific loss of exact attribution rather than indiscriminate miscalibration.

Executed deduplication and retraining. We execute a Jaccard-0.80 MinHash deduplication pass, rebuild the corpus, and retrain rather than merely partitioning a fixed run by similarity. Sub-threshold families retained by that pipeline still produce false-member rates of approximately 92.00% and 99.00% in the $[0.50, 0.70)$ and $[0.70, 0.80)$ bands at nominal 5.00% FPR. Deduplication reduces repeated content but cannot certify that every provenance-confounding family is absent.

Robustness across scores and environments. The devastating false-attribution effect is not an artifact of a specific model, score, or training run. As detailed in Table 1, the failure persists across alternative LLM-specific auditing metrics (including Min-$K_{20\%}$ (Shi et al., 2024), Min-K++ (Zhang et al., 2025b), EZ-MIA (Ilic et al., 2026), and neighborhood comparison (Mattern et al., 2023)), where false-positive rates remain at least an order of magnitude above the nominal 5.00% operating point. It also survives transfer to a different architecture (Qwen3.5-2B LoRA) and withstands an executed Jaccard-0.80 deduplication and retraining pipeline. In stark contrast, a boundary control using meaning-preserving but lexically disjoint semantic paraphrases drops the false attribution rate to 10.00%, twice the nominal 5.00% but far below the natural lexical families' 99.70%. In this single-exposure setting the vulnerability is therefore driven primarily by surface-level lexical similarity rather than by semantic overlap (§5.3 shows a heavier fine-tuning regime where semantic families also shift the score). Complete removed-band and alternative-operating-point results are deferred to Appendix H.

Table 1: CC-News robustness at nominal 5.00% FPR. FP is the false-attribution rate on family-present exact non-members (no-sibling records for the control row). P/T denote pooled/per-target calibration.

Model/run	Score (cal.)	Targets	FP
Pythia/core	loss (P)	all	88.00%
	LiRA (T)	all	99.70%
	loss (P)	control	1.20%
Pythia/repeat	loss (P)	all	98.00%
	Min-$K_{20\%}$ (P)	all	97.00%
	Min-K++ (P)	all	82.00%
Pythia/EZ	EZ-MIA (P)	all	63.00%
Pythia/nbr.	neighbor (P)	all	58.00%
Qwen3.5 LoRA	loss (P)	$J < 0.80$	37.00%
Pythia/dedup-FT	loss (P)	$J \in [0.50, 0.70)$	92.00%
	loss (P)	$J \in [0.70, 0.80)$	99.00%
Pythia/semantic	loss (P)	all	10.00%

4.2 Replication on WikiText-103

We validate these findings on WikiText-103, fine-tuning 96 GPT-2 shadows across 56 targets with natural siblings and 32 controls. At a nominal 1.00% FPR, natural siblings inflate the clean-reference false-member rate from 1.40% to 81.00%, while no-sibling controls remain calibrated. Crucially, a family-conditioned auditor retains near-perfect exact inference (*adv* ≈ 0.99), confirming this failure is strictly false attribution rather than erasure. With the exact passage absent, the family-present and family-absent worlds separate almost perfectly (*adv* $= 1.00$), demonstrating that these scores are highly reliable indicators of broader family exposure rather than exact provenance. These studies show the composite null invalidates standard verdicts in the wild, motivating controlled interventions that isolate whether families substitute for or merely confound the exact record.

5 Four-World Interventions Explain the Failure

5.1 Causal influence of faithful content

To establish causality, we evaluate controlled same-label families on CIFAR-10 using 48 shielded targets and 256 target-disjoint bystanders, following standard shadow density estimation protocols (Carlini et al., 2022a). With exact membership randomized independently against a fixed family presence (k=8), faithful near-duplicates reduce vulnerable-target advantage from 0.55 to 0.11, a change of -0.44 (Figure 3). In contrast, lower-fidelity reinforcement variants change advantage by only -0.08, while same-class density and random controls show null effects (-0.03 and -0.05). Conflicting-label records instead increase advantage by $+0.24$ and reduce target accuracy. Global accuracy remains stable, and the target-disjoint bystanders exhibit no measured average leakage change (Appendix D.4). The intervention therefore depends strictly on reproducing target content, not merely adding records, changing class density, or increasing optimizer exposure.

5.2 Substitution versus addition regimes

The four-world audit distinguishes standard exact-record evidence, family evidence, and the exact record's residual after conditioning on the family (Table 3). Table entries use one pooled score-AUC estimator throughout and should not be interchanged with the per-target advantages reported elsewhere. In classification, family presence is detectable but exact-given-family inference falls to near

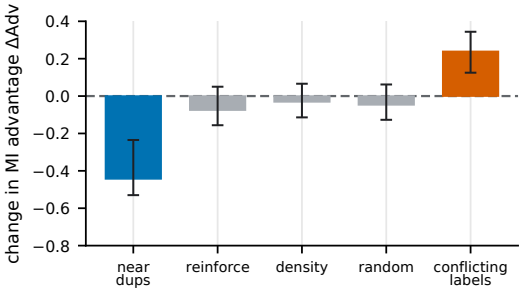


Figure 3: CIFAR-10 trichotomy. Near duplicates reduce advantage; lower-fidelity and disjoint controls remain null (95% CIs).

Calibration → evaluation	TPR	FPR	LR^+
clean → clean	0.37	0.07	5.20
copy-cal. → copies	0.11	0.11	1.00
clean → copies	0.61	0.59	1.00
conflict-cal. → conflict	0.56	0.06	9.90
clean → conflict	0.13	0.02	5.40

Table 2: CIFAR-10 verdicts at nominal 5% FPR; LR^+: bootstrap medians. Copy/conflict calibration reproduces the evaluated family.

chance. A family-conditioned auditor's TPR drops to its measured FPR and LR^+ approaches 1.00, with 88.00% of the 24 vulnerable targets falling below advantage 0.05. At a decision level (Table 2), the clean-reference auditor's FPR rises to 0.59 on faithful-family non-members. Recalibrating with the family removes this false-positive shift but cannot recover the exact bit, leaving both TPR and FPR at 0.11. Conflicting-label families provide a sign control by retaining strong separation, yielding $LR^+ = 9.90$ under conflict-conditioned calibration and $LR^+ = 5.40$ even under clean calibration. This confirms that classification exhibits substitution, where faithful coverage renders the exact example redundant (Feldman & Zhang, 2020; Feldman, 2020). Autoregressive language modeling behaves fundamentally differently. CC-News family presence is strongly detectable, yet exact-given-family AUC remains 0.75. Because every exact exposure can continue reducing sequence loss, the exact record retains a residual. The practical failure is nevertheless severe because the natural family alone crosses a boundary calibrated without families. Thus, text models demonstrate additive confounding rather than classification-style erasure.

5.3 The learning objective drives the split

Because vision and text models differ in modality and architecture, we isolate the impact of the learning objective with a matched intervention on AG-News. We hold the GPT-2 backbone, target records, family definitions (semantic paraphrase families sharing almost no 5-grams with their targets; Appendix I), exposure levels, and four-world audit design strictly constant, varying only the training task between sequence classification and autoregressive modeling. This single shift causes the exact-given-family AUC to jump from 0.53 under classification to 1.00 under autoregressive modeling, with the standardized surviving exact fraction rising from 0.27 to 0.86. Because family presence remains independently detectable under both objectives (AUC 0.62 and 0.89), and these lexically disjoint families show that even purely semantic overlap shifts scores under multi-epoch fine-tuning, the divergence cannot be explained by the family simply becoming unlearnable. Instead, this intervention definitively identifies the learning objective as the causal driver of the substitution-versus-addition split. Finally, replacing natural CC-News families with meaning-preserving paraphrases (Jaccard 0.01) reduces false attribution to 10.00% while leaving exact-record inference fully intact (AUC 1.00, Table 1), confirming that lexical near-duplication is the dominant confounding channel at single-exposure scale, though heavier fine-tuning extends it to semantic families.

Table 3: Complete four-world identification results (pooled score AUC). Columns are $E \mid F{=}0$: H_{10} vs. H_{00}; $F \mid E{=}0$: H_{01} vs. H_{00}; and $E \mid F{=}1$: H_{11} vs. H_{01}. Chance is 0.50; cls./AR denote classification/autoregressive modeling.

Setting	Standard	Family	Residual
CIFAR-10 cls.	0.56	0.69	0.52
CC-News AR	0.87	0.98	0.75
AG-News cls.	0.61	0.62	0.53
AG-News AR	1.00	0.89	1.00

5.4 Score saturation and local invariance

To formalize why the local score basin becomes invariant to exact membership, let S denote the trained family of x, and let $\phi(x; D)$ be the expected audit score of query x after training on D. The

exact record's discrete marginal contribution in the context of its family is

$$\Delta_x(S) = \phi(x; D \cup S \cup \{x\}) - \phi(x; D \cup S). \tag{5}$$

Family records raise the queried target's score more strongly when the target is OUT than when it is IN, shrinking the standardized IN/OUT gap from 1.34 to 0.10 for near-duplicates and 0.04 for exact copies. Because highly faithful records cover more of the same local signal, their benefit saturates before x is added, forcing $\Delta_x(S)$ to shrink. This saturation extends across the target's entire neighborhood: querying fixed augmentations of each target reveals their collective IN/OUT gap collapses from 2.38 without a family to 0.26 with near-duplicates (Figure 9). The effect strictly follows fidelity (Figure 7): exact copies erase the signal most strongly, while interpolating a family record toward an unrelated image smoothly removes the protection, explaining the density null (Figure 10). This erasure remains tightly localized to the target: the two nearest same-class training neighbors see their membership advantage change by only 0.01 (with a confidence interval of $[-0.02, 0.05]$). This interpretation clarifies why low exact-record leakage does not imply low content leakage: the score shift is carried by S and remains robustly observable during a family-presence test.

This mechanistic effect scales with the family budget k, appearing even with a single near-duplicate and saturating near $k = 8$ (Figure 11). The substitution dynamics consistently transfer across different datasets and architectures (driving a -0.76 advantage reduction on CIFAR-100 and a -0.30 reduction on VGG-11), while the additive residual persists across controlled language models (Figure 12). Furthermore, the classification erasure regime withstands adaptive, family-aware auditors (Figure 6). When reference shadows are explicitly trained with the family present, advanced attacks (including full-covariance LiRA, augmentation differencing, explicit family-record queries, and RMIA) all remain near chance (AUC 0.47–0.53; Appendix C). While this constitutes empirical robustness across black-box channels rather than an information-theoretic guarantee, it confirms a unified framework: fidelity dictates how strongly the family moves the target score, while the learning objective dictates whether that movement substitutes for or adds to the exact-record evidence.

6 Model Evidence Relocates to the Content Family

6.1 High detectability of family presence

Near-duplicate families redistribute rather than universally erase model evidence: what is lost in exact-record specificity is relocated to the broader content family. With the exact target absent, family presence remains highly detectable (AUC 0.69 in vision, 0.98 in text). Consequently, the very score shift that triggers a false-positive verdict for exact membership operates as a true-positive signal of family exposure. Our controlled cluster-toggle experiment (Figure 4) makes this evidence relocation explicit: as family size k sweeps from 1 to 16 in vision, exact-given-family advantage falls from 0.09 to 0.01, while family-presence advantage climbs from 0.12 to 0.48. The isolated pythia-2.8B reference point illustrates the cross-objective contrast, preserving a detectable exact-sequence residual while simultaneously exposing the content family (AUC 0.98). Near-duplicate exposure thus shifts the identifiable unit of a model-only audit from the discrete record to its surrounding content family, without any change to the underlying scores.

6.2 Limitations of composite-null calibration

A natural defense is to calibrate the auditor against both components of the composite null hypothesis in Equation (1). While a worst-case threshold can theoretically control the false-positive rate across H_{00} and H_{01}, this conservative correction cannot recreate the separation erased by the family, decimating power when the family-induced score shift is large. In our vision experiments, the resulting composite exact test achieves an AUC of only 0.53, with statistical power matching the nominal 5.00% false-positive rate (equivalent to random chance). Even an oracle four-world generalized likelihood-ratio (GLR) test reaches only 0.55 AUC. On CC-News text, the corresponding composite and oracle AUCs are 0.69 and 0.74. Here, shifting from standard clean calibration (H_{00}) toward worst-case family calibration (H_{01}) successfully reduces the worst-case false-positive rate from 0.88 to 0.05, but at an unacceptable cost: power on standard exact members (H_{10}) drops from 0.34 to 0.00, while power on family-present members (H_{11}) plummets from 0.99 to 0.17. Crucially, this limitation is not an artifact of a simplistic monotone threshold: the oracle GLR test directly models all four worlds using perfect knowledge of the score distributions, yet remains bounded by these

same low AUCs (Appendix E). In practice, incomplete family knowledge presents a further, serious deployment obstacle. If a text auditor identifies and accounts for only 1 or 2 members of an 8-sibling family, the resulting false-positive rate remains highly elevated at 0.70 or 0.19, respectively. The auditor recovers nominal calibration ($\alpha = 0.05$) only once at least half of the family (4 of 8 siblings, FPR 0.06) is reproduced and included in the reference shadows (Appendix E). Because exhaustively enumerating every syndicated variant, mirrored page, or minor edit across web-scale data is vastly more difficult than in this controlled best-case evaluation, robust composite-null calibration remains practically unfeasible.

6.3 Scoped reporting and recommended abstention

When near-duplicate variants of a queried record cannot be definitively excluded from the training corpus, model-only evidence should never be treated as dispositive for proving exact-record data provenance. Instead, we argue that a rigorous model-only audit must explicitly declare its unit of inference, distinguishing a discrete record from a broader lexical family. To evaluate an exact-record claim, the auditor should construct both clean (H_{00}) and family-present (H_{01}) reference null distributions, disclosing separate false-positive and power rates across all four worlds ($H_{00}, H_{01}, H_{10}, H_{11}$). If plausible near-duplicates cannot be exhaustively enumerated, or if the family-present null crosses the clean threshold, the only defensible outcome is a proactive abstention on exact membership, optionally accompanied by a positive family-presence finding. This reporting rule preserves what the model actually reveals without silently converting general content exposure into provenance evidence. When model-only audits must abstain, definitive exact provenance must rely on external records, such as training-set inspection, signed manifests, or system access logs.

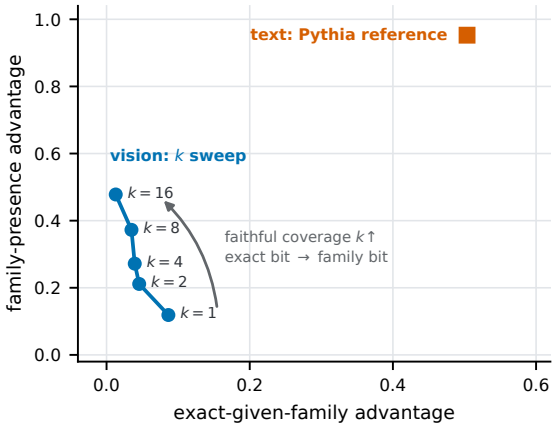


Figure 4: Information relocation. The blue path connects the five evaluated vision family sizes $k \in \{1, 2, 4, 8, 16\}$; the Pythia point is a single cross-objective reference, not a text sweep. Crossed-bootstrap intervals are reported in the appendix.

This paradigm shift reflects the empirical boundaries of black-box audits. Across nine tested scoring channels, including loss, logit, covariance, and neighborhood statistics, faithful classification families completely erase exact-record separation, whereas autoregressive text models preserve a detectable exact-sequence residual. Crucially, this residual signal does not repair standard audits because family presence still falsely crosses the clean threshold with high probability. The saturation mechanism is moreover dual use: the same data-side additions that reduce a record's exact-membership leakage can equally be used to weaken or fabricate model-only evidence, further underscoring why such verdicts are forensically invalid on web-scale corpora where natural, un-enumerated near-duplicates abound.

7 Conclusion

This paper exposes a fundamental vulnerability in using membership inference as a forensic tool for data provenance. By showing that standard audits conflate exact-record inclusion with broader content-family exposure, we demonstrate that a positive model-only score is an invalid proof of individual copyright or licensing claims. The four-world audit framework provides a rigorous methodology to separate these overlapping signals, mapping the precise boundaries of exact-record survival and erasure. Ultimately, these findings shift the conceptual baseline of machine learning audits from treating neural networks as discrete databases toward understanding them as continuous statistical representations of broad data neighborhoods. In doing so, this work establishes a new, defensible standard for data-provenance auditing, ensuring that future privacy and legal evaluations remain grounded in what models actually reveal.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318. ACM, 2016.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *International Conference on Machine Learning (ICML)*, pp. 2397–2430. PMLR, 2023.

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A. Smith, and Christopher A. Choquette-Choo. Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pp. 18703–18726. ACL, 2025.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *USENIX Security Symposium (USENIX Security)*, pp. 2633–2650. USENIX, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 1897–1914. IEEE, 2022a.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramèr. The Privacy Onion Effect: Memorization is Relative. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022b.

Harsh Chaudhari, Giorgio Severi, Alina Oprea, and Jonathan R. Ullman. Chameleon: Increasing Label-Only Membership Leakage with Adaptive Poisoning. In *International Conference on Learning Representations (ICLR)*, 2024.

Dingfan Chen, Ning Yu, and Mario Fritz. RelaxLoss: Defending Membership Inference Attacks without Losing Utility. In *International Conference on Learning Representations (ICLR)*, 2022.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking High-Accuracy Differentially Private Image Classification through Scale. *CoRR abs/2204.13650*, 2022.

Murat Bilgehan Ertan, Emirhan Böge, Min Chen, Kaleel Mahmood, and Marten van Dijk. On the Evidentiary Limits of Membership Inference for Copyright Auditing. *CoRR abs/2601.12937*, 2026.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Annual ACM Symposium on Theory of Computing (STOC)*, pp. 954–959. ACM, 2020.

Vitaly Feldman and Chiyuan Zhang. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016.

David Ilic, David Stanojevic, and Kostadin Cvejoski. Powerful Training-Free Membership Inference Against Fine-Tuned Autoregressive Language Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14077–14093. ACL, 2026.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 259–274. ACM, 2019.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In *International Conference on Machine Learning (ICML)*, pp. 10697–10707. PMLR, 2022.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8424–8445. ACL, 2022.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset Inference: Ownership Resolution in Machine Learning. In *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 *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2024.

Neal Mangaokar, Ashish Hooda, Zhuohang Li, Bradley A. Malin, Kassem Fawaz, Somesh Jha, Atul Prakash, and Amrita Roy Chowdhury. What Really is a Member? Discrediting Membership Inference via Poisoning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2025.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pp. 11330–11343. ACL, 2023.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. Copyright Traps for Large Language Models. In *International Conference on Machine Learning (ICML)*, pp. 35296–35309. PMLR, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations (ICLR)*, 2017.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine Learning with Membership Privacy using Adversarial Regularization. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 634–646. ACM, 2018.

Shahbaz Rezaei and Xin Liu. On the Discredibility of Membership Inference Attacks. *CoRR abs/2212.02701*, 2022.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18. IEEE, 2017.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating Membership Inference Attacks by Self-Distillation Through a Novel Ensemble Architecture. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2022.

Qwen Team. Qwen3 Technical Report. *CoRR abs/2505.09388*, 2025.

Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2779–2792. ACM, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-Friendly Differential Privacy Library in PyTorch. *CoRR abs/2109.12298*, 2021.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-Cost High-Power Membership Inference Attacks. In *International Conference on Machine Learning (ICML)*, pp. 58244–58282. PMLR, 2024.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual Memorization in Neural Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2023.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Position: Membership Inference Attacks Cannot Prove That a Model was Trained on Your Data. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 333–345. IEEE, 2025a.

Jingyang Zhang, Jingwei Sun, Eric C. Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Helen Li. Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025b.

ETHICS STATEMENT

The mechanism studied in this paper is dual use. A party that trained on a record without permission could misuse it to weaken later model-based evidence, or to fabricate a member-like score for a record that was never trained on. It equally lets a contributor reduce the exact-record leakage of their own data. We disclose these risks explicitly, scope every claim to the exact-record bit (content presence remains detectable), and recommend that audits treat model-only exact-record verdicts as unresolved whenever near duplicates may be present. The study uses only public datasets and public model weights and involves no human subjects, personal data, or live production systems.

REPRODUCIBILITY STATEMENT

All experimental details needed to reproduce our results, including shadow-model counts, calibration procedures, scoring, and per-experiment configurations, are specified in §3 and Appendices A–J. Upon acceptance, we will publicly release the full codebase together with per-shadow score files and a driver that reproduces every reported table and figure.

AI USE STATEMENT

We used generative AI tools to polish the writing for readability and consistency, and to generate the plotting code for the figures. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the final content of this work.

A THE EVIDENCE MODEL

Terminology. Throughout the appendix, a *shield* is a record of the controlled family added around a target (§5), and a *shielded* target is one whose family is present in training.

The trichotomy, its fidelity ordering, and the saturation are not independent findings; one scalar model *organizes* them (the protective and null arms follow from a single concavity assumption; the backfire arm needs one added assumption, below). Fix target x of class y, and assign each training example z an *evidence weight* toward (x, y), $s(x)=1$; a same-label example carries $s(z)=\rho(z) \in [0, 1]$,

where ρ is its content fidelity to x (copies $\rho{=}1$, near-duplicates slightly less, content-disjoint $\rho{\approx}0$); a conflicting-label example near x carries $s(z){=}-\beta\rho(z)$ with $\beta{>}0$. In text there is no label channel, so a shared-prefix/divergent-suffix shield has small $s \geq 0$ and $\beta{=}0$. With total evidence $t(D){=}\sum_{z\in D} s(z)$, model the score $\phi(x){=}g(t(D)) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ over training randomness, g increasing.

Assumption 1 (saturating fit)**.** *g is strictly concave and bounded on $[0, \infty)$ with $g' \to 0$.*

This is cheap to defend, cross-entropy gradients on an example scale with $1{-}p_y$ and vanish as it is fit, so each added faithful copy moves the score less; it holds exactly for faithful (positive-evidence) shields in the kernel instantiation (Appendix B). The backfire arm alone needs one more ingredient.

Assumption 2 (boundary steepness)**.** *Extended to conflicting evidence $t < 0$, g is S-shaped; g' is unimodal with mode $t^\star \leq 0$, so g is convex below $t^\star$ and concave above (A1 is its restriction to $t \geq 0$).*

The protective and null arms use A1 alone; A2 is invoked only for conflict.

Proposition 2 (gap identity, mean level)**.** *With shield evidence t_0 present in both worlds, $\mu_{\text{in}} - \mu_{\text{out}} = g(t_0{+}1) - g(t_0)$. Every shield effect on the exact-record mean gap is the discrete derivative of g at t_0.*

Corollary 1 (advantage readout)**.** *For Gaussian scores the per-record advantage is a function of $\Delta\mu/\sigma_p$, where $\sigma_p{=}\sqrt{\sigma_{\text{in}}^2 + \sigma_{\text{out}}^2}$ is the pooled dispersion (AUC$=\ \Phi(\Delta\mu/\sigma_p)$); so* at pooled dispersion held fixed across the budget*, it is strictly increasing in the mean gap $\Delta\mu$.*

Concavity forces μ_{out} to rise more than μ_{in} ($g(t_0){-}g(0) > g(t_0{+}1){-}g(1)$), the asymmetry of Figure 9a ($+5.45$ vs $+3.10$).

Proposition 3 (the four arms)**.** *Let $h(t){=}g(t{+}1) - g(t)$; under A1 it is positive, strictly decreasing, and $h(t) \to 0$. Then (i) in the* protect *arm, faithful shields give $t_0{=}k\rho$, so the gap $h(k\rho)$ strictly decreases in k and ρ (Fig. 10); (ii) in the* null *arm, content-disjoint additions have $\rho{\approx}0$, so $t_0{=}0$ and the gap equals baseline regardless of count (density, random); (iii) in the* backfire *arm, under A2, a conflict count m giving $t_0{=}{-}\beta m$ with $t^\star{\leq}t_0{<}0$ yields $h(t_0){>}h(0)$, widening the gap (non-monotone in m); (iv) in* text*, $\beta{=}0$ and first-token-only conflict give $t_0{\approx}0^+$, hence null-to-slightly-protective.*

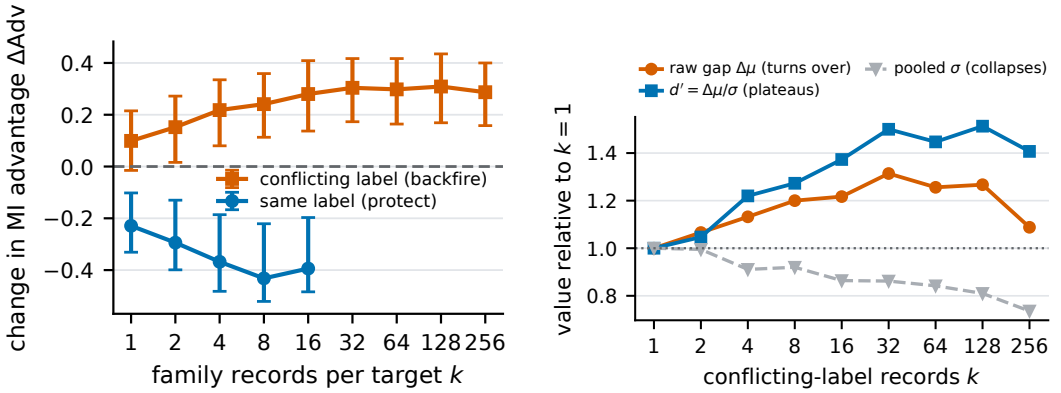
The proofs are one line ($h'{=}g'(t{+}1) - g'(t){<}0$ by strict concavity; (iii) uses A2, i.e. g' peaking at $t^\star{\leq}0$; Appendix B). Thus A1 *alone* forces the protect and null arms and the fidelity ordering, the backfire arm additionally requires A2, and the text null ($-0.09, -0.00$) follows from $\beta{=}0$; the model *organizes* the trichotomy rather than deriving all four arms from one assumption.

Tested prediction. A2 predicts the backfire is non-monotone in the conflict budget, the *advantage* should eventually turn over. It does not; sweeping k_{conf} to 256 (Figure 5), Δ*adv* plateaus at $\sim$0.30 even at $8\times$ the budget. This is a failure of *Corollary 1's budget-invariant-dispersion assumption*, not of the mean-gap model; the moment decomposition and its caveats are given once, in Appendix B (Limits). We report this plainly rather than rescue it, and the backfire *sign* needs no A2 in any case; the μ-asymmetry (μ_{out} crashes more than μ_{in}, §5.4) already delivers it. The model's *positive* record stands. It unifies the four arms under one decreasing h, quantitatively fits the budget curve ($\lambda'{\approx}3.5$, out-of-sample to $k{=}16$; §5.4), reproduces the Truth-Serum amplification at $k{=}8$ ($+0.24^*$, §D.1), and its basin analysis (Prop. 4, §5.4) located the residual score-channel surface that the full-covariance and differencing attacks then probed and found empty.

B KERNEL INSTANTIATION AND PROOFS

We instantiate the evidence model of §A with kernel ridge regression (equivalently a Gaussian process posterior mean), where Assumption 1 and the *protective* statements (Prop. 2, Prop. 3(i,ii)) hold *exactly*; the backfire arm (A2) and the basin bound (Prop. 4) need the added assumptions noted below and are *not* instantiated here.

Setup. Predict the score for x by KRR against a background value b with ridge λ, and suppose x's evidence is carried by a cluster of n points each at kernel similarity κ to x, against an uninformative



(a) Advantage change vs. budget

(b) Moment decomposition

Figure 5: Pushing the conflicting budget to k=256. **(a)** Two arms of one decreasing-h mechanism. Same-label evidence saturates *protection*, conflicting-label evidence saturates *backfire* (Prop. 3); the backfire Δadv plateaus at $\sim$0.30 with no turnover, and its k=8 point matches the Truth-Serum amplification (§D.1). **(b)** The A2-predicted turnover does not appear in the advantage; the moment decomposition and its caveats are in Appendix B. Panel (b) series are normalized to k=1. Note k=256 is a *mechanism probe* (a $\sim$25% dataset increase), not a deployment budget; the protective sweep stays at $k\leq$16 (sub-percent). Vulnerable targets, crossed-bootstrap 95% CIs.

background. The posterior mean at x is

$$g(n) = b + \Big(1 - \tfrac{\lambda}{n\kappa+\lambda}\Big)(y-b), \qquad g(0) = b, \tag{6}$$

which is strictly increasing in n toward the label y. This is exact for n *identical* copies of x (rank-one Gram $K{=}\kappa\mathbf{1}\mathbf{1}^\top$, each at cross-similarity κ to x); for distinct near-duplicates the Gram is only approximately rank-one and (6) is a scalar-evidence heuristic that treats the cluster's evidence as additive ($n\kappa$). We use it only for the protective (positive-evidence) arms.

Concavity (A1). With n continuous, $g'(n) = (y{-}b)\lambda\kappa/(n\kappa+\lambda)^2 > 0$ and $g''(n) = -2(y{-}b)\lambda\kappa^2/(n\kappa+\lambda)^3 < 0$ for $y > b$, and $g'(n) \to 0$. Hence g is strictly concave and bounded, so A1 holds exactly. Note g' is *monotone* on the valid domain $n \geq 0$. This g realizes the protective and null arms exactly but does *not* instantiate the backfire arm, which requires the S-shaped g of A2 (an interior g' peak at $t^\star \leq 0$), a separate modeling assumption the rational KRR g does not satisfy.

Gap and budget curve (Prop. 2, 3). The IN world has the original plus k copies ($n{=}k{+}1$), the OUT world only the k copies ($n{=}k$), so

$$\Delta(k) = g(k{+}1) - g(k) = \frac{(y{-}b)\,\kappa\lambda}{(k\kappa+\lambda)\big((k{+}1)\kappa+\lambda\big)},$$
$$\frac{\Delta(k)}{\Delta(0)} = \frac{\lambda'(1{+}\lambda')}{(k{+}\lambda')(k{+}1{+}\lambda')}, \qquad \lambda'{=}\lambda/\kappa, \tag{7}$$

a one-parameter curve, strictly decreasing in k, $\to 0$ (Prop. 3(i); the fit in Fig. 11). Fidelity enters as κ. A shield of similarity $\kappa_s < \kappa$ contributes $k\kappa_s/\kappa$ effective copies, so protection is monotone in fidelity; content-disjoint shields have $\kappa_s \approx 0$ and leave Δ at baseline (the null). With $\phi(x) = g(\cdot) + \varepsilon$, for Gaussian IN/OUT scores per-record AUC $= \Phi(\Delta\mu/\sigma_p)$ *exactly* (pooled $\sigma_p{=}\sqrt{\sigma_{\text{in}}^2 + \sigma_{\text{out}}^2}$; no equal-IN/OUT-variance assumption is needed). This map is only *qualitative* in practice. Our measured pairs depart from a single-σ reading by up to $\sim 2\times$ (mild non-Gaussianity and budget-varying σ_p; unequal $\sigma_{\text{in}}, \sigma_{\text{out}}$ does *not* affect it), so we treat Cor. 1's advantage claim as monotone in $\Delta\mu/\sigma_p$, not as a specific functional form.

Basin invariance (Prop. 4). For a query q at similarity κ_q to x, the IN$-$OUT shift is $\delta(q) \propto (y{-}b)\kappa_q\lambda/(n_q\kappa+\lambda)^2$, where n_q counts nearby evidence. Where κ_q is large (near x), faithful shields

make n_q large, so $\delta(q) \to 0$ by the same saturation; where $\kappa_q \approx 0$ (far), $\delta(q) \approx 0$ directly. The Neyman–Pearson bound for a Gaussian location family with *isotropic*, independent-per-query noise ($\Sigma{=}\sigma^2 I$) and mean gap $\|\boldsymbol{\delta}\|_2$ gives Prop. 4; correlated queries instead require the Mahalanobis $\sqrt{\boldsymbol{\delta}^\top \Sigma^{-1} \boldsymbol{\delta}}$ and fall outside it (handled empirically, §5.4). For exact copies ($\kappa_s{=}\kappa$), $\delta(q) \to 0$ for all q as $k \to \infty$.

Remark 1 (mean-shift decomposition)**.**

$$\underbrace{g(t_0{+}1) - g(t_0)}_{\textit{membership | cluster}} + \underbrace{g(t_0) - g(0)}_{\textit{cluster presence}}$$
$$= \; g(t_0{+}1) - g(0) \;\geq\; g(1) - g(0) \;\; (\textit{undefended}).$$

Mean-shift decomposition (Rem. 1). Equation (6) telescopes directly; the total is a lower bound, *not* a conserved quantity (it grows with k), and as k grows the point-membership term $\to 0$ while the cluster-presence term grows.

Limits. The model assumes (a) a scalar sufficient statistic; (b) a pooled dispersion σ_p *constant across the budget*, violated in practice (σ_p shrinks as k grows, i.e. heteroscedasticity across budgets, which is what erases the predicted advantage turnover in the confusion sweep); (c) an identical-copy rank-one Gram matrix for the exact derivation, whose extension to distinct near-duplicates is a scalar-evidence heuristic, not an established bound; (d) A2's S-shape for the backfire arm, which the rational KRR g does not satisfy; and (e) isotropic query noise for Prop. 4, so correlated-query attacks are covered only empirically (§5.4). A1 is proved here for the kernel case and assumed for deep networks (cross-entropy gradient decay). It is a *mechanism* model consistent with the phenomenology, not a security proof, and speaks only to the score channel; non-score attacks (training-set access, near-duplicate counting) fall outside the query-only access considered here. Its one prediction beyond the fitted phenomenology, a conflicting-budget turnover in the *advantage* (Prop. 3(iii)), is *not* borne out; the advantage plateaus out to $k{=}256$ (Fig. 5). Only a post-hoc decomposition suggests a turnover in the unobservable mean gap ($\Delta\mu$ peaks near $k{\approx}32$ at 3.64 and declines to 3.01 by $k{=}256$; paired $+0.63$, but a single post-selected contrast on a target-only bootstrap, so suggestive not confirmatory), while the pooled σ falls $\approx 15\%$ over $k{=}32{\to}256$ (-26% over $k{=}1{\to}256$), leaving $d'{=}\Delta\mu/\sigma$, and the advantage, flat. Separately, per-instance DP already shows that records with similar neighbors enjoy tighter per-instance bounds than isolated outliers, which is precisely what a shield manufactures, so the connection to a formal per-record guarantee has analytic support, not merely conjecture; making it quantitative for shielded records is the closest route from this heuristic toward a certificate. The σ collapse itself has a mechanistic reading. Heavy conflicting injection makes the local fit near-deterministic (the model confidently predicts the wrong label regardless of x's membership), so $\phi(x)$ becomes less sensitive to the membership mask.

Reproducibility details. We fit λ' in d'-space over $k \in \{1, 2, 4, 8\}$ through the Gaussian AUC relation ($k{=}16$ treated as saturation noise), yielding $\lambda'{\approx}3.5$, $R^2{=}0.74$; the curve's apparent floor at large k is the AUC map saturating, so the fit does not pass through $k{=}1$ exactly (-0.18 predicted vs the measured -0.23). Shadow-model counts per experiment follow.

C ADAPTIVE-ATTACK DETAILS

This section gives the full adaptive-attack analysis summarized in §5.4. We build a shield-aware LiRA, a per-target diagonal-Gaussian likelihood ratio over the joint vector $z = [\phi(x), \phi(16 \text{ augmentations of } x)]$, calibrated on *discovery* shadows that themselves contain the shields, frozen, and evaluated on disjoint shadows. This attacker is stronger than naive LiRA on undefended records. On the vulnerable targets its TPR@1%FPR is 0.24 versus naive 0.17 (Figure 6), so the comparison is not against a weakened adversary. On shielded records its TPR@1%FPR falls to 0.01 (neardup) and 0.00 (copies), with AUC ≈ 0.50–0.53, at the 1%-FPR floor and a $>20\times$ reduction from the adaptive baseline. (All adaptive-attack numbers are on the same vulnerable-24 targets as RMIA below, for a like-for-like comparison.) We do not claim immunity to all future attacks; we report that this stronger, shield-aware attacker gains essentially no excess true-positive rate over the 1%-FPR floor that matters for privacy.

Table 4: Number of shadow models used in each experiment. Counts separated by a slash report the baseline and intervention arms, respectively.

experiment	shadows
baseline	96
main trichotomy arms	48
generalization (CIFAR-100, VGG; baseline/arms)	64/32
text (scratch, finetune)	32
adaptive LiRA, RMIA, mechanism	96
DP-SGD ϵ-sweep, GN control	24 each
De-style DP	16
k-sweep, interp, confusion budget	48
why-keep-x (utility; trade-off probe)	64 (32)
CC-News scale-up (*none*/*natdup*)	64/64
WikiText natural siblings	96
WikiText erasure frontier (per arm)	96
AG-News objective control (cls./AR)	64/64
CC-News semantic *natdup*	28
dedup-and-retrain (*natdup*)	64
Qwen3.5 LoRA ablation	48

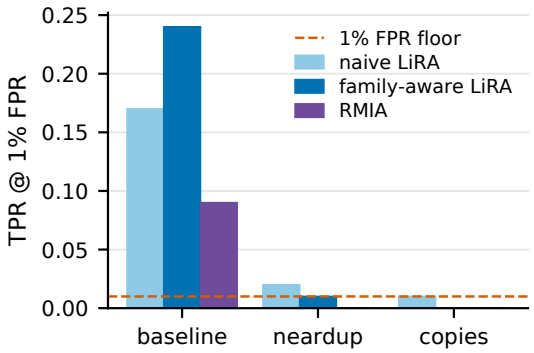


Figure 6: A shield-aware adaptive attacker is stronger than naive LiRA on undefended records (baseline) yet is driven to the 1%-FPR floor on shielded records; RMIA, a structurally different attack family, behaves likewise.

Attackers the diagonal LR cannot represent. A diagonal likelihood ratio cannot capture cross-augmentation correlation or the natural shield-aware statistic $\phi(x)$ minus the local basin level. We therefore add two stronger attacks on the same features, namely a *full-covariance* joint LiRA over z (isotropic shrinkage, intensity 0.3, for the 24-sample/17-dim estimator), and the explicit difference $\phi(x) - \overline{\phi(\text{aug})}$ (the text-domain analogue is the neighbourhood-comparison attack of Mattern et al. (2023)). On undefended records the full-covariance attack matches the diagonal one (AUC 0.84 vs 0.86), confirming it is a real attacker; the differencing statistic is weak *even there* (AUC 0.53), because subtracting the shared basin removes the membership signal along with the basin. On shielded records both collapse to chance (full-covariance AUC 0.49–0.51, difference AUC 0.47–0.50). To rule out that the 17-dimensional covariance is merely under-determined ($\sim$24 samples), we also run the optimal contrast as a *well-conditioned* 2-*dimensional* full-covariance LR on $[\phi(x), \overline{\phi(\text{aug})}]$. It is strong undefended (AUC 0.84, TPR@1% 0.22) yet also floors on shielded records (0.53/0.01 neardup, 0.51/0.00 copies), and sweeping the 17-d shrinkage over $\{0.1, 0.3, 0.5\}$ leaves the shielded AUC in $[0.49, 0.52]$ throughout. Even at x itself, where residual curvature might let signal survive, none does. The exact-record signal remains at chance against every LiRA variant we construct, diagonal or full-covariance, aggregating or differencing.

An attacker that reads the shields. The shields are treated as *non-secret*, so the strongest adaptive attacker should query the *actual* k=8 shield images, not generic augmentations of x. We record ϕ on the real shields per shadow and fit a per-target full-covariance LR over

$[\phi(x), \phi(\text{shield}_1), \ldots, \phi(\text{shield}_8)]$. It gains *nothing*. On the shielded world its AUC is 0.52 (TPR@1% 0.01), indistinguishable from naive LiRA's 0.52, and combining the shields with the 16 augmentations (25-dim) also floors (0.51). Reading the known shields does not help because they are present in *both* the member and non-member worlds and near-saturated, which makes them membership-invariant, exactly the basin of Prop. 4. This closes the most natural adaptive gap. Even an attacker who knows and queries every shield floors.

Why score-channel attacks floor. The model of §A explains this under one stated assumption. Model any black-box attacker as choosing queries $q_1, \ldots, q_m$ and observing $\phi(q_j)=g(t_j + \rho_j \mathbf{1}[x \in D]) + \varepsilon_j$, where t_j is shield evidence at q_j and ρ_j is x's own influence there. IN and OUT induce Gaussians differing only in mean, with per-query shift $\delta_j=g(t_j+\rho_j) - g(t_j)$.

Proposition 4 (basin invariance, independent-query case)**.** *If the per-query noise is independent (ε_j i.i.d., $\Sigma=\sigma^2 I$), then for any test* $\mathrm{TPR} \le \Phi\big(\Phi^{-1}(\mathrm{FPR}) + \|\boldsymbol{\delta}\|_2/\sigma\big)$. *Near x, ρ_j is large but so is t_j (faithful shields shadow x's influence profile), so δ_j is killed by saturation (A1); far from x, δ_j is killed because $\rho_j \approx 0$. For exact copies the shadowing is perfect and* $\sup_j \delta_j \to 0$ *as* $k \to \infty$.

So for (near-)independent queries, small $\boldsymbol{\delta}$ bounds *every* such attack. This is why the shield-aware diagonal LiRA and RMIA floor (the theorem is *consistent with*, not proof of, those empirical results). The bound does *not* cover correlated queries. For those, the Neyman–Pearson-optimal statistic uses the Mahalanobis distance $\sqrt{\boldsymbol{\delta}^\top \Sigma^{-1} \boldsymbol{\delta}}$, which can exceed $\|\boldsymbol{\delta}\|_2/\sigma$ along a low-variance contrast direction such as $\phi(x) - \overline{\phi(\text{aug})}$. That is precisely the residual surface, and the full-covariance and differencing attacks above are built to exploit it; that they nonetheless collapse to chance on shielded records is therefore *empirical* evidence that $\boldsymbol{\delta}$ itself vanishes there, not a consequence of the bound. For near-duplicates ($\kappa_s < \kappa$) this empirical robustness, not the theorem, is the operative statement.

A different attack family, RMIA. A skeptic might worry the defense merely evades LiRA's specific statistic. We therefore also evaluate RMIA (Zarifzadeh et al., 2024), which uses a fundamentally different test. Instead of a per-target Gaussian on the target's own loss, RMIA computes the online pairwise likelihood ratio of the target against a *population* of reference points z, scoring $\Pr_z[\,(p_\theta(x)/\bar{p}(x))/(p_\theta(z)/\bar{p}(z)) > \gamma\,]$ (our port of the official implementation, $\gamma=2$, bystanders as the population). On the same vulnerable-24 records RMIA is a working but *weaker* attack than LiRA on this subset (AUC 0.68, TPR@1%FPR 0.09, versus naive LiRA's $0.80/0.17$; Figure 6). On shielded records it too is driven to chance, with AUC 0.48 (neardup) and 0.50 (copies), with TPR@1%FPR $= 0.00$. That a second, structurally different attack family, even if weaker here, fails on the same shields indicates the effect is not specific to LiRA's statistic.

D Secondary reading: the family intervention as exact-record camouflage

This appendix collects experiments that read the controlled family intervention of §5 from the contributor's side. These results are *not* part of our main claims. As stated in §2, any defensive interpretation is strictly secondary, and the mechanism is dual use (Ethics statement). The same data-side addition that suppresses exact-record evidence for a contributor is available to any party that can place records in a training corpus, which is precisely why model-only verdicts should not be treated as proof of exact inclusion. We report the protective effect, its cost, and its limits so that readers can judge how cheaply exact-record evidence can be moved, not to propose a privacy defense. Content presence remains detectable throughout (§6).

D.1 Truth Serum poisoning

Truth Serum (Tramèr et al., 2022) is the offensive mirror of shield data. An adversary with the ability to inject a few points inserts *mislabeled replicas* of the target to *amplify* its leakage (Chameleon (Chaudhari et al., 2024) does the same for label-only attacks via adaptive poisoning). We reproduce it ($k=8$ clean copies of each target with a wrong label) and confirm amplification. Vulnerable advantage rises by $+0.24^*$; indeed our confusion arm (§5.1) is the same mechanism, which is why it backfires. We then ask whether an added faithful family counteracts such an attacker. Adding neardup shields on top of the Truth-Serum poison reduces the advantage by -0.34^* versus the no-poison baseline,

turning the attacker's $+0.24$ amplification into net protection, close to the unpoisoned shield's -0.44. Faithful, correct-label coverage counteracts the mislabeled poison, so even under active poisoning the exact-record signal falls back close to the unpoisoned shielded level. Exact-record evidence can thus be pushed in either direction by cheap data-side additions.

D.2 Why keep the record? A negative result on local influence

If faithful shields make x's content persist, the natural challenge is why the real x should be kept at all. In a separate 64-shadow run of the cluster assay (Figure 8; hence exact-record advantages here read 0.15 vs. that figure's 0.14), we measure the marginal value of keeping x as margin(x+shields) $-$ margin(shields-only) on x and on 16 held-out neighbors of x (fresh augmentations, never used as shields), paired over targets. At high fidelity (k=8 neardup) x is *task-redundant*; keeping it changes accuracy and margin by $\Delta\approx 0$ (n.s.), because the shields already saturate its neighborhood, which is the substitution mechanism of §5.2 seen from the utility side. On these local metrics, withholding x and keeping its shields matches x+shields while making exact membership definitively false (0 vs 0.15). The margin gain from keeping x reappears only as fidelity falls ($+0.86^*$ at half-fidelity, $+1.23^*$ at quarter-fidelity, target-only bootstrap), and it is essentially x's counterfactual self-influence, the very quantity MI exploits, so the two axes are two views of one number. We do not measure any downstream-task advantage from keeping x; where the literal record must be trained (e.g. under a contractual provenance obligation verified out of band), that requirement is external to these metrics.

D.3 DP-SGD comparison details

We compare the shield's exact-record effect with DP-SGD on the targeted records. This is a comparison of mechanisms, not a proposal to replace DP. Against a *tuned* DP-SGD (Abadi et al., 2016) baseline (Opacus (Yousefpour et al., 2021), GroupNorm, augmentation, tuned at ϵ=8, swept $\epsilon \in \{1, \ldots, 64\}$, strong evaluation per; Table 5), *no* operating point reaches the shield's corner (advantage 0.11 at accuracy 0.90), because a targeted data intervention escapes the global accuracy-vs-noise tradeoff; a non-private GroupNorm control reaches advantage 0.37, so part of DP's effect is architectural. We do *not* claim Pareto-dominance. DP gives an all-record worst-case guarantee whereas the shield helps only its 48 targets, and the full De et al. (2022) recipe at scale is untested. We claim only that on the targeted records the shield beats every DP-SGD point *we test*.

Table 5: Tuned DP-SGD vs shield on the protected (vulnerable) records, CIFAR-10/ResNet-9. Baseline vulnerable advantage 0.55; accuracy on the clean holdout. The non-private control (GroupNorm pipeline, ϵ=∞) indicates how much of the advantage reduction is architectural versus from privacy.

defense	test acc.	remaining adv. (Δ*adv*)
no defense	0.90	0.55
shield (neardup)	**0.90**	**0.11** (-0.44^*)
non-private (GN, ϵ=∞)	0.80	0.37 (-0.18)
DP-SGD ϵ=64	0.68	0.31 (-0.24^*)
DP-SGD ϵ=16	0.62	0.23 (-0.32^*)
DP-SGD ϵ=8	0.57	0.21 (-0.34^*)
DP-SGD ϵ=4	0.51	0.24 (-0.31^*)
DP-SGD ϵ=1	0.30	0.27 (-0.27^*)

Along the ϵ sweep, accuracy and protection trade off as they must. Larger ϵ buys accuracy ($0.57\rightarrow 0.62\rightarrow 0.68$ at ϵ=$8, 16, 64$) but advantage rises back toward the non-private 0.37 ($0.21\rightarrow 0.23\rightarrow 0.31$; Table 5). A truncated run of the high-accuracy De et al. (2022) recipe (wide WRN-16-4, augmentation multiplicity, 120 epochs, a budget tractable for a shadow-model study) reaches the same 0.57 accuracy at ϵ=8 with a vulnerable-target advantage no lower than the simpler DP baseline, so it does not reach the shield's corner either; its published headline (~ 0.81) needs a hundreds-of-epochs regime that a per-record LiRA study of dozens of models cannot afford. We therefore claim only that every DP configuration we *successfully evaluated* stays off the $0.90/0.11$ corner, not that a better architecture could not improve accuracy at fixed ϵ.

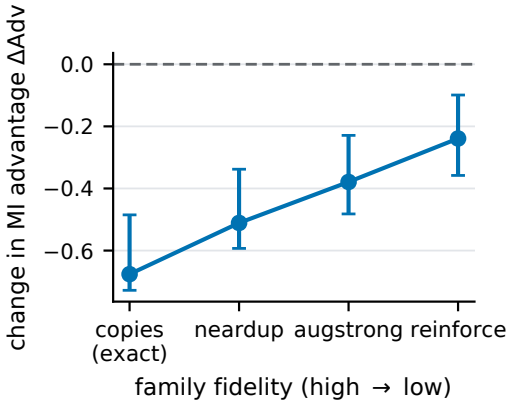


Figure 7: Controlled fidelity ordering, online augmentation disabled. Exact copies reduce exact-record advantage most, then near duplicates, strong augmentations, and reinforcement variants; the ordering isolates stored family fidelity from online augmentation.

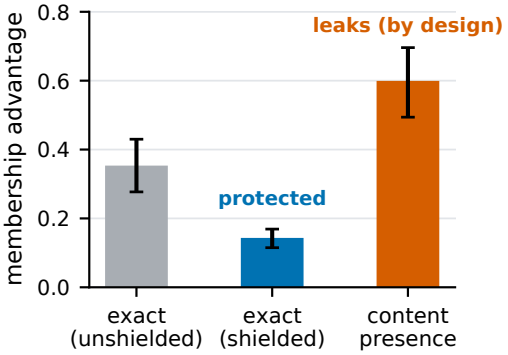
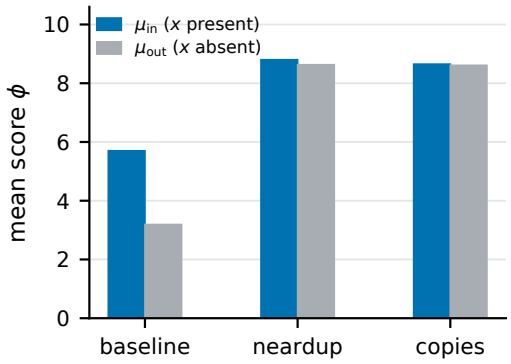
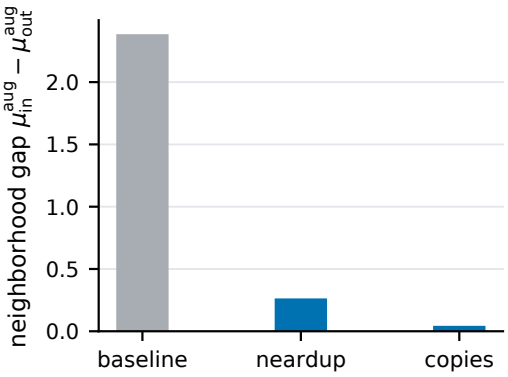


Figure 8: What the family intervention changes. Faithful families weaken *exact-record* inference (advantage $0.35 \to 0.14$) while *content presence* stays easily detectable (0.60, scoring $\phi(x)$ with x absent), because the family makes x's content persist (48 shielded targets, CIFAR-10).



(a) Score collapse



(b) Neighborhood gap

Figure 9: Mechanism. (a) Shields raise the OUT-world score μ_{out} to meet μ_{in}, collapsing the IN/OUT gap (d', $1.34 \to 0.10 \to 0.04$). (b) The target's neighborhood (its augmentations) becomes confident and membership-invariant under shielding.

D.4 Utility and bystander effects

Across all arms, global test accuracy is statistically unchanged from baseline, and the target's own accuracy when OUT is preserved for every *same-label* arm (e.g., OUT-accuracy ≥ 0.97 in the fidelity sweep); only confusion harms it. Critically, the *bystander* (non-shielded) targets show $\Delta adv \approx 0$ in every arm and dataset; shielding one record does not expose others. This is a powered null, not absence of data. Over the 256 bystander targets the neardup effect is $\Delta adv = -0.00$ with a 95% CI half-width of ± 0.02, so any externality larger than ~ 0.02 would have shown up on average. Unlike the onion effect's removal, additive shielding has no measured *average* externality on random bystanders; we do not test neighborhood-matched, rare, or worst-case-tail records, where a localized defense would most plausibly act, so this bounds the average, not the worst case.

E Supplementary figures

Figures 9 and 10 give the mechanism detail for §5.4; Figures 11 and 6 support the budget sweep (§5.4) and the adaptive-attack analysis (§5.4).

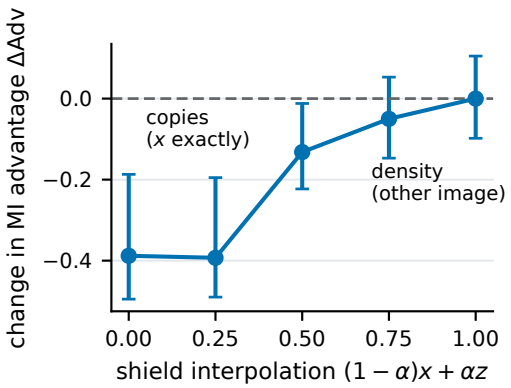


Figure 10: Causal test. As shields interpolate from the exact target (α=0) toward a random same-class image (α=1), protection decays monotonically to the density null, so proximity to x is the lever.

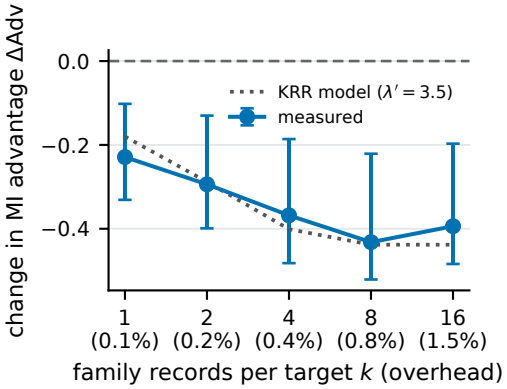


Figure 11: The exact-record advantage reduction grows with the shield count per target and saturates near k=8; even one near-duplicate (0.1% overhead) yields a significant reduction. Vulnerable targets, CIFAR-10/ResNet-9, 95% CI.

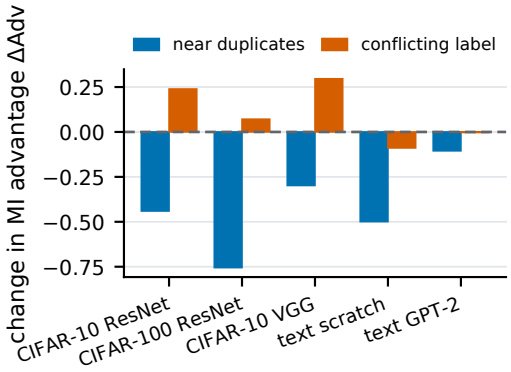


Figure 12: Faithful-coverage protection holds across datasets, architectures, and modalities (including a finetuned 124M GPT-2). The confusion backfire is vision-specific; in text it is null.

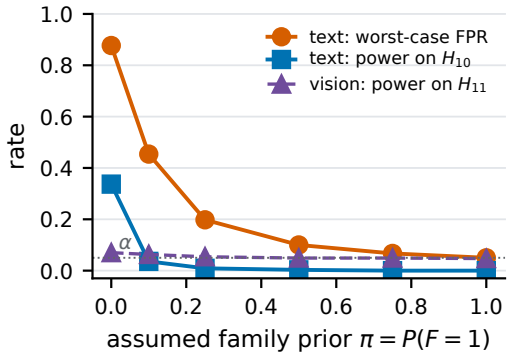


Figure 13: Sweeping the assumed family prior π (§6.2). In text, worst-case FPR control (orange) drives H_{10} power (blue) to 0 as $\pi \to 1$; in vision, H_{11} power (purple) stays $\approx\alpha$.

Calibration–power tradeoff vs. the assumed family prior (Fig. 13). The nuisance-aware auditor of §6.2 controls FPR on the composite exact-record null by assuming a family prior π=$P(F$=1) and calibrating one threshold τ on the mixture $(1-\pi)H_{00} + \pi H_{01}$; π=0 is the clean-reference auditor and π=1 the minimax (worst-case-H_{01}) auditor of §6.2. Sweeping π over $[0, 1]$ shows the section's two failure modes are endpoints of a strict tradeoff. In *text* (pythia-2.8b), the worst-case FPR over $\{H_{00}, H_{01}\}$ falls from 0.88 (π=0, false attribution) to 0.05 (π=1, controlled) exactly as power on the *family-absent* exact alternative H_{10} collapses from 0.34 to 0 (and on H_{11} from 0.99 to 0.17). No prior both controls the family-present null and keeps exact-record power. In *vision* power stays $\approx\alpha$ (0.05–0.07) for *every* π, because faithful coverage has removed the exact bit, so no calibration can recover it.

Oracle four-world GLRT (not a monotone-threshold artifact). The auditors of §6.2 threshold the scalar ϕ. To rule out that a *monotone* threshold is simply the wrong test once there are four score distributions, we give the auditor a generalized likelihood-ratio test with every advantage. We fit per-world multivariate Gaussians $p(z \mid H_{00}), \ldots, p(z \mid H_{11})$ on a feature vector z (loss ϕ and, for text, Min-K%), and score the composite exact-record ratio $\mathrm{LR}(z)$=$[p(z \mid H_{10})$+$p(z \mid H_{11})]/[p(z \mid H_{00})$+$p(z \mid H_{01})]$ (equal F prior), fitting the densities on half the shadow seeds and evaluating the ratio's AUC on the *held-out* half (so the number is honest, not fit=eval). The GLRT does *not* change the picture. In *vision* it separates E at AUC 0.55 (vs. 0.53 for the monotone threshold), still chance,

so the exact bit is genuinely *absent* rather than hidden by a poor threshold; in *text* it reaches 0.74 (vs. 0.69; the Min-K% feature adds the 0.02), confirming that text *retains* exact information but that recovering it requires resolving the latent family state. The substitutive/additive split is thus robust to the form of the test.

Partial family awareness is insufficient in text. The family-conditioned auditor of §3 is an oracle that reproduces a target's *entire* near-duplicate family in its calibration shadows; a realistic auditor reproduces only part of it. We sweep the number m of siblings the auditor's shadows reproduce (pythia-2.8b/CC-News natdup at $\max$-sib$=m$; $m{=}8$ is the deployed full family), calibrate its family-present threshold there at $\alpha{=}5\%$, and measure the false-attribution rate and power on the *deployed* full-8-sibling model (exact-given-family AUC 0.90). False attribution stays far above α until the auditor reproduces at least half of the family. The FPR is $0.70/0.19/0.06/0.05$ at $m{=}1/2/4/8$ over 12 seeds per setting. Power on true members is $0.99/0.84/0.56/0.50$, and LR^+ is $1.40/4.50/8.80/10.00$. Because the exact bit is *additive* in text (§5.2), *full* family recall controls the FPR while retaining power, which is the sound family-conditioned outcome; in vision, power stays $\approx\alpha$ at every m. Nominal calibration, however, requires reproducing at least half of the family ($m{=}4$), and enumerating every natural duplicate of a target is infeasible at corpus scale. A realistic text audit therefore operates at low m and false-attributes. This is the quantitative form of the abstention argument (§6.3). Min-K++ calibration is weaker and noisier (deployed AUC 0.69), so we report the suffix-loss auditor.

F Experiment map and reconciliation of reported leakage numbers

The exact-record signal is reported under several constructions, and the numbers are *not* interchangeable. Table 6 maps every headline value to the three canonical comparisons of §3 (*standard exact*, *exact-given-family*, *family presence*), plus *content presence* $C{=}E{\vee}F$ (H_{00} vs. the other three worlds), and to its measurement coordinates, namely target subset (vulnerable-24 = top-50% by discovery advantage; all-48 = every shielded target), faithful-coverage k, the score, and the estimator (a *pooled* one-threshold LiRA log-LR audited by AUC, vs. a *per-target* advantage averaged over targets). This factorization also positions the audit relative to recent proposals that respond to near-duplicates by relaxing membership itself to a semantic neighborhood (Mangaokar et al., 2025): rather than replacing the exact-record question with a coarser one, the four-world design reports the exact bit and the family bit separately, and Mangaokar et al. (2025) show that even the relaxed definition remains unreliable under dataset manipulation. Values *within* a quantity differ by these coordinates and are not meant to coincide numerically; they should only share sign and order of magnitude. In particular the four vision *exact-given-family* numbers (0.11, 0.03, 0.14, $0.09 \to 0.01$) are one near-chance quantity seen through four lenses (subset, pooling, score, and k), not a discrepancy; the single text *exact-given-family* value (0.50) sits an order of magnitude above all of them, reflecting the substitutive/additive split (§5.2). "Advantage" is $2|\mathrm{AUC} - 0.5|$; where a construction is naturally read as a raw AUC (the four-world audits) we give AUC and its advantage in parentheses.

G Text erasure frontier across fidelity and exposure

Vision faithful copies *erase* the exact-record verdict near-totally (aware LR^+ 1.00, 88% of targets driven to advantage ≤ 0.05 at $k{=}8$; §5.2). Does the same hold in text, and is the natural passages' non-erasure (§4.2) merely their mid-fidelity? We answer both by sweeping sibling *fidelity* and *exposure* independently on the 56 WikiText targets, constructing optimization-free near-verbatim siblings (a controlled fraction of tokens resampled from the corpus unigram distribution) at token 5-gram Jaccard $\in \{0.59, 0.82, 0.96\}$ and total exposure $E = n_{\text{sib}} \cdot K_{\text{shield}} \in \{2, 8, 16\}$, always present while the exact target's membership is toggled; the auditor is the per-target aware LiRA of §3 (96 shadows/arm, clean reference LR^+ 66.00, advantage 0.90).

Table 7 shows the outcome. At matched low exposure ($E{=}2$) fidelity is nearly inert. Aware LR^+ moves only $58.00 \to 65.00$ across Jaccard $0.59 \to 0.96$, so the natural passages' survival is *not* explained by their fidelity; a near-verbatim sibling at the same exposure is no more hiding than a mid-fidelity one. Erosion is driven by *exposure*. Holding Jaccard 0.82 fixed, $E{=}2 \to 16$ drives aware

Table 6: Experiment map. Every reported exact/family leakage value, keyed to a canonical quantity (E=exact-record bit, F=family presence) and its measurement coordinates. Arrows are defense effects (unshielded→shielded, or coverage k swept). The numbers are consistent once the coordinates are read. For example, vision exact-given-family is near chance under every lens, while text retains an order-of-magnitude-larger signal. Vision uses CIFAR-10/ResNet-9, and text uses pythia-2.8b/CC-News with the suffix NLL score.

Canonical quantity	value	modality, subset (k)	score, estimator	appears
Standard exact	0.55	vision, vuln-24	LiRA log-LR, per-target adv.	§5 (baseline)
H_{10} vs. H_{00}	0.35	vision, all-48	confidence, per-target adv.	Fig. 8 (unshielded)
(x alone, family absent)	0.99	text, all shielded	suffix NLL, per-target adv.	§4.1
	0.74 (AUC 0.87)	text	suffix NLL, pooled AUC	Tab. 3
Exact-given-family	0.11	vision, vuln-24 (k=8)	LiRA log-LR, per-target adv.	§5 (neardup)
H_{11} vs. H_{01}	0.03 (AUC 0.52)	vision, pooled (k=8)	LiRA log-LR, pooled AUC	Tab. 3, §5.2
(family present,	0.14 (0.15 in App. D.2)	vision, all-48	confidence, per-target adv.	Fig. 8 (shielded)
toggle x)	0.09→0.01	vision, all-48 (k=1→16)	LiRA log-LR, pooled	Fig. 4 (frontier)
	0.50 (AUC 0.75)	text	suffix NLL, pooled AUC	Tab. 3, §5.2
Family presence	0.60	vision, all-48	confidence, per-target adv.	Fig. 8
H_{01} vs. H_{00}	0.38 (AUC 0.69)	vision, all-48	log-LR, pooled AUC	Tab. 3
(x absent, toggle family)	AUC 0.98	text	suffix NLL, pooled AUC	Tab. 3, §6
	LR^+ 3.20 / 8.10	vision / text (WikiText)	log-LR verdict, 5% FPR	App. J
Content presence	AUC 0.65	vision	log-LR, pooled AUC	this table only
C=$E\vee F$	AUC 0.95	text	suffix NLL, pooled AUC	this table only

Table 7: Aware exact-record verdict on text as sibling fidelity (Jaccard) and exposure (E) vary. WikiText/GPT-2 uses α=1%, with clean-reference LR^+ 66.00 and advantage 0.90. The CIFAR k=8 reference has LR^+ 1.00 and 88% erased. Exposure, not fidelity, is the lever. At E=2, raising Jaccard $0.59 \to 0.96$ barely moves the verdict; erosion appears only at high exposure and remains *partial* even at 16 near-verbatim copies, and never reaches vision's near-total erasure.

	E=2	E=8	E=16
Jaccard 0.59	LR^+ 58.00 (adv 0.85)	–	–
Jaccard 0.82	LR^+ 56.00 (adv 0.81)	LR^+ 23.00 (adv 0.62)	LR^+ 10.00 (adv 0.39)
Jaccard 0.96	LR^+ 65.00 (adv 0.79)	–	LR^+ 7.00 (adv 0.34)

LR^+ $56.00 \to 10.00$ (advantage $0.81 \to 0.39$). But even the extreme cell, 16 copies of a Jaccard-0.96 sibling, only *partially* erodes the verdict (LR^+ 7.00, advantage 0.16–0.34, 11% of targets erased), far from vision's near-total collapse at k=8. Across every cell the *unaware* auditor is driven to fabrication (FPR $\geq$0.85). We read this as a substitutive/additive split. Autoregressive suffix-loss memorization pins the *exact* token sequence, so a distribution-aware auditor keeps a residual grip on the true member that near-duplicates dislodge only slowly, whereas vision's boundary memorization is fully supplied by faithful copies. In text the operative corruption of the exact-record verdict is therefore *fabrication* (universal, one to a few duplicates suffice), not *erasure* (weak, exposure-gated, partial); the fidelity-ordered mechanism (§A) is shared across modalities but its reach into the exact-record verdict is not.

A matched-exposure control shows that memorization is exposure-additive. A naive fidelity sweep is confounded by exposure, because inserting the target K extra times makes it detectable by count alone. We therefore control for it, with one target insertion against *eight* sibling copies always present, varying only sibling fidelity. Even with eight *exact* copies present (the exact-copy multiplicity endpoint, 8 vs. 9), adding the exact instance once still yields exact-given-family advantage 0.74 (WikiText/GPT-2, 85 passages), barely below a different-suffix sibling (0.83) and matching the 0.75 at pythia-2.8b scale (§5.2), far above vision's 0.03 (§5.2). Eight identical copies do *not* saturate the exact instance's contribution, and the per-token residual $\Delta_t = \mathbb{E}[\ell_t \mid H_{01}] - \mathbb{E}[\ell_t \mid H_{11}]$ is spread roughly uniformly across the suffix (mean 0.08, flat; Fig. 14) rather than concentrated at distinctive tokens. Autoregressive suffix-loss memorization is thus *exposure-additive*; near-duplicate coverage cannot substitute the exact sequence, which is the mechanism behind the text-vs-vision asymmetry (vision's boundary memorization *is* saturated by faithful copies, collapsing the exact bit to chance).

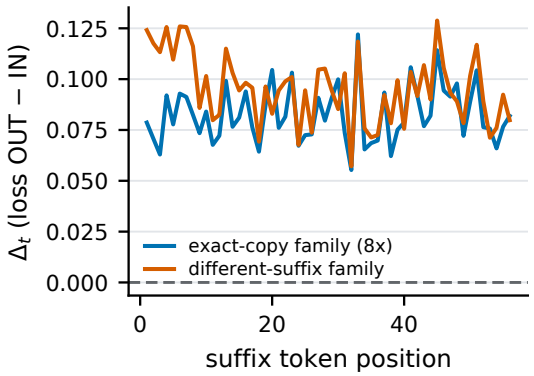


Figure 14: Per-token residual $\Delta_t = \mathbb{E}[\ell_t \mid \text{target OUT}] - \mathbb{E}[\ell_t \mid \text{target IN}]$ across the 56 suffix positions, with eight sibling copies always present (matched exposure). The exact target's presence lowers loss *across the whole suffix* rather than at localized tokens. This distributed occurrence-count signal is consistent with exposure-additive memorization.

H Details of the CC-News scale-up and deduplication

Setup. We stream CC-News (the *vblagoje/cc_news* release), tokenize with the model tokenizer to fixed 256-token passages (exact duplicates removed), and mine natural near-duplicate families by token-5-gram MinHash (P=128, 32 bands) + LSH + union-find (§4.2 pipeline). The family-size distribution over 40k passages contains 1726 families at Jaccard 0.5, 921 at 0.8, 508 at 0.9 (median shielded best-sibling Jaccard 0.89). A *standard* near-dedup (Lee et al., 2022) removes exact duplicates and near-duplicates at Jaccard $\geq$0.8; we keep the 58 *dedup-surviving* targets (best sibling in $[0.5, 0.8)$), 141 *dedup-removed* ($\geq$ 0.8), and 100 no-sibling bystanders, plus a 6k-passage deduped background; one of the 200 families has best-sibling Jaccard below 0.50 and is excluded from the banded analyses. We full-finetune pythia-2.8b (bf16 AMP, gradient checkpointing, AdamW 10^{-5}, single target exposure) over 128 shadows (64 *none* / 64 *natdup*), toggling each exact target's membership while its natural siblings are always present in *natdup*.

Calibrated verdict. Per-target LiRA log-LR, threshold calibrated to α-FPR on calibration OUT, FPR measured out of sample with a crossed bootstrap over calibration seeds, evaluation seeds, and targets.

Table 8: Calibrated CC-News exact-record verdicts. Thresholds are calibrated on family-absent non-members; the unaware and aware auditors are evaluated with natural near-duplicate families present. LR^+ is the crossed-bootstrap median of per-resample TPR/FPR and need not equal the ratio of the printed point rates.

α	auditor	TPR	FPR	LR^+
1%	(A) clean → clean	1.00	0.02	**61.67**
1%	(B) unaware → natdup	1.00	**1.00**	**1.00**
1%	(C) aware → natdup	0.64	0.02	28.36
5%	(B) unaware → natdup	1.00	1.00	1.00
5%	(C) aware → natdup	0.78	0.06	13.69

The unaware verdict is fabricated ($\text{LR}^+ \to 1$); the aware auditor retains power (LR^+ 28, paired advantage 0.90→0.75, only 2% of targets driven below advantage 0.05). Text is therefore not erased (App. G), but awareness requires infeasible enumeration. Family presence stays a usable verdict (LR^+ 9.1 at α=1%, 7.3 at 5%; bootstrap-median). A four-world audit (over {x-only, dup-only, both, neither}) recovers family presence (dup-only vs. neither, AUC 0.98) far better than exact inclusion given the family (both vs. dup-only, AUC 0.75). The exact bit is thus partially recoverable, unlike vision's near-chance 0.52 (§5.2), which is the substitutive/additive split again.

Fabrication by dedup band. A realistic unaware auditor's false-"member" rate on genuine non-members, by whether a standard MinHash dedup (Jaccard ≥ 0.80) would remove the siblings. The *surviving* band is not removed by dedup yet still fabricates; bystanders (no siblings) are the control.

Table 9: False-member rates for CC-News exact non-members by natural-family Jaccard band. The threshold is calibrated on family-absent non-members at each nominal FPR α.

target band	base adv.	α=1%	α=5%	α=10%
dedup-surviving (J 0.50–0.80)	0.98	**0.55**	**0.85**	**0.91**
dedup-removed ($J \geq 0.80$)	0.99	0.63	0.89	0.95
bystander (no siblings, control)	1.00	0.00	0.01	0.03

Min-K% (Shi et al., 2024) reproduces the surviving-band fabrication (0.59/0.82 at α=1%/5%), so it is not an artifact of the loss score.

Neighbourhood-comparison attack (§4.1). The most semantically relevant baseline for text MI (Mattern et al., 2023) scores $\phi_{\mathrm{nbr}}(x) = \overline{\mathrm{NLL}(8 \text{ Qwen paraphrases of } x)} - \mathrm{NLL}(x)$. This score normalizes the suffix loss with the target's *own* paraphrase set. On matched four-world shadows with 6 *none* and 6 *natdup* seeds and 1,800 scores each, it does not escape the confound. Under a common pooled clean-reference calibration at α=5%, its false attribution on family-present non-members is 0.58 versus 0.99 for the raw loss. The result is lower but still $\gg \alpha$. The lexical family lowers $\mathrm{NLL}(x)$ but not the loss of the untrained paraphrase neighbors, so normalization removes only part of the shift. It also weakens the exact-given-family signal from AUC 0.90 to 0.58, while standard family-absent MI stays strong at 1.00 to 0.92. Normalizing by neighbors trades exact-record power for false attribution that it only partly suppresses.

Executed dedup-and-retrain. The partition above holds the training set fixed. As the *direct* test we instead *remove* every best-sibling-Jaccard-$\geq$0.8 family from the corpus and *re-finetune* pythia-2.8b on the resulting τ=0.8-deduplicated corpus (64 *natdup* shadows; the family-absent worlds are unchanged, so the *none* shadows are reused). Fabrication does not abate; it *sharpens*, because the confounding high-Jaccard families of *other* targets are now absent from training. The *surviving* families (J<0.8, which a τ=0.8 pass keeps) are false-attributed at 0.92/0.99 (5% FPR, $[0.50, 0.70)$/$[0.70, 0.80)$ bands; 0.96/1.00 at 10%), *above* the fixed-corpus partition estimate (0.85), while the exact-given-family signal is *retained and if anything sharpened* (four-world AUC 0.90 on the full corpus $\rightarrow$ 0.97 after the τ=0.8 dedup, over the 200 CC-News families of this strong-memorization finetune; naive H_{10} vs $H_{00} \approx 1.0$). The main-text 0.75 (Tab. 3) is a *separate*, weaker-memorization pythia run (naive 0.87); exact-given-family tracks how strongly the finetune memorizes, but is well above chance in every run. Even the *removed* band ($J \geq 0.8$, whose closest siblings are deleted) still fabricates at 0.66–0.71 (5% FPR) from residual sub-threshold near-duplicates. So the failure survives an *executed* deduplicate-and-retrain, not merely a measured-Jaccard partition of a fixed model.

LoRA ablation. To confirm the effect is not specific to full finetuning of pythia, we repeat the test with LoRA finetuning (r=16, 10.9M trainable parameters) of a different-architecture 2B model (Qwen3.5-2B), on the same CC-News families (48 shadows). Baseline exact-record MI is again strong (advantage 1.00), and the surviving-band fabrication reproduces. The unaware auditor's false-"member" rate on dedup-surviving non-members rises to 0.16/0.37/0.77 at α=1%/5%/10% (removed band 0.16/0.39/0.78), while bystanders stay clean (0.00). The magnitude is lower than pythia's full finetuning at the 1% operating point, but the effect, dedup-surviving-band fabrication with a clean bystander control, holds across architecture and finetuning method.

I SEMANTIC-FAMILY EXPERIMENTS, THE OBJECTIVE CONTROL, AND THE LEXICAL CONFOUND

Paraphrase families. To decouple *meaning* from *tokens* we build *semantic* families. For each shielded target we detokenize it, prompt Qwen2.5-7B (few-shot) for candidate rewrites that preserve the facts, and keep the k=8 with the *lowest* word-5-gram Jaccard to the target. Mean lexical Jaccard

is 0.00 (AG News) / 0.01 (CC-News), versus ≈ 0.9 for the near-duplicate families. They share the topic but almost no n-grams.

Objective control (§5.3). On AG News we hold the backbone (GPT-2), records, semantic family, and exposure fixed and swap *only* the objective, comparing GPT-2 with a sequence-classification head against the same GPT-2 as an autoregressive LM, each in the four-world design (64 shadows). Both have a strong family-absent exact signal (naive H_{10} vs H_{00}, classification AUC 0.61, LM 1.00). They diverge on *exact-given-family* (H_{11} vs H_{01}), with classification at 0.53 (chance; $R_{\text{exact}}{=}0.27$, 95% CI $[0.19, 0.36]$) and the LM at 1.00 ($R_{\text{exact}}{=}0.86$, $[0.82, 0.89]$). We read the split from AUC and the standardized R_{exact} rather than the raw interaction I (both ≈ -2.5), because the LM's suffix NLL saturates at a floor for memorized targets, so the raw mean interaction is a ceiling artifact while identifiability is not. With everything but the objective fixed, classification erases the exact bit and the LM retains it, so the objective is causal.

Lexical vs. semantic families (§5.3). Re-running the pythia-2.8b four-world with the CC-News *semantic* paraphrase family in place of the lexical one (natdup arm re-finetuned; family-absent worlds reuse the *none* shadows; 28 shadows) collapses the clean-reference fabrication to $\text{FPR}_{H_{01}}{=}0.10$ ($\approx\alpha$), versus 0.88 for the lexical family, while exact-given-family stays 1.00. Paraphrases that share meaning but not tokens barely move the target's suffix loss in this single-exposure regime, so they neither fabricate nor mask; the residual 0.10 is still twice the nominal α. In this regime the confound is predominantly *lexical* and the recovered quantity is near-duplicate *family* presence. The AG-News objective control (above) shows the boundary is regime-dependent, because under multi-epoch fine-tuning the same paraphrase construction is detectable at family-presence AUC 0.89 (Table 3). Concurrent work reaches the complementary conclusion from the adversarial side: Ertan et al. (2026) show that an accused developer can deliberately paraphrase training data to degrade state-of-the-art MIAs while preserving utility, with residual leakage that varies across fine-tuning regimes. Their deliberate obfuscation and our natural and controlled paraphrase families probe the same lexical dependence of the membership signal, and both observe the regime dependence documented in §5.3.

Natural family features and the confound. The controlled fidelity/k sweeps (§5.4) have a natural analogue. On the CC-News families, the OUT-score shift rises monotonically with best-sibling fidelity (mean shift $1.19/1.22/1.35/1.38$ across the $[0.50, 0.70)/[0.70, 0.80)/[0.8, 0.9)/[0.9, 1.0)$ Jaccard bands; Spearman $+0.19$ of best-sibling Jaccard with the shift), while false attribution is high in every band (false-"member" rate $0.80/0.88/0.82/0.92$ at $\alpha{=}5\%$). Family size adds to the shift once ≥ 2 siblings are present.

J Additional quantitative results from the full study

This section records the secondary operating points, confidence intervals, and construction statistics from the full result presentation. They are included for completeness and do not change the headline comparisons in the main text.

Study scale and natural-family construction. For the primary controlled CIFAR-10 experiment, the 48 targets and $k{=}8$ family records add 384 records, or approximately 0.8% of the base training set. Across all experiments, the study trains approximately $1{,}500$ shadow models; the per-experiment allocation is reported in Table 4. For WikiText-103, target-to-sibling token-5-gram Jaccard similarity has median 0.60 and range $[0.40, 0.81]$ after manual verification that the passages contain copied text.

Complete controlled-vision verdict. All five decision-level rows of the controlled CIFAR-10 verdict appear in Table 2. In particular, the clean-reference auditor evaluated on a conflicting-label family has TPR 0.13, FPR 0.02, and $\text{LR}^{+}{=}5.4$. On CIFAR-10, a fixed-threshold family verdict reaches $\text{LR}^{+}{=}3.20$ with TPR 0.57 at nominal 5% FPR.

Under the family-conditioned faithful comparison, per-target advantage declines by 0.37, with crossed-bootstrap 95% CI $[0.20, 0.52]$, and 88% of the 24 vulnerable targets fall below advantage 0.05. This family-conditioned 0.37 is not interchangeable with the clean-reference -0.44 of §5.1 or the independent budget sweep's -0.43 at $k{=}8$; the three constructions differ in comparison world and training run and are expected to agree only in sign and magnitude (App. F). In the four-world mean

decomposition, family presence changes the vision exact-record effect from δ_E^0=1.08 to δ_E^1=0.18; the interaction is I=−0.90 with 95% CI $[-1.46, -0.32]$, and R_{exact}=0.23. WikiText is closer to additive, with I=−0.02, CI $[-0.06, 0.03]$, and R_{exact}=0.89; the pythia-2.8B experiment gives I=−0.36 and R_{exact}=0.60. The conflicting-label arm increases advantage from 0.55 to 0.79, a change of +0.24.

Mechanism, budget, and generalization details. The target's two nearest same-class training neighbors receive an OUT-score shift of +0.20, with 95% CI $[0.05, 0.34]$, while their membership advantage changes by only +0.01, CI $[-0.02, 0.05]$. In the near-duplicate budget sweep, the advantage change grows from −0.23 at k=1 to −0.43 at k=8 and saturates at −0.39 for k=16; the kernel model fitted on $k \leq 8$ predicts −0.44 at k=16. With online augmentation, the reinforcement-variant effect weakens from −0.24 to −0.08.

Before vulnerable-target selection, CIFAR-100 has baseline advantage 0.73. Near duplicates change the selected targets' advantage by −0.76, the density control remains null, conflicting labels change it by +0.07, and bystanders remain unchanged. On CIFAR-10/VGG-11, near duplicates and conflicting labels change advantage by −0.30 and +0.30, respectively. In controlled language modeling, copies and near duplicates change advantage by −0.44 and −0.50 for the model trained from scratch, and by −0.32 and −0.11 for finetuned GPT-2. Conflicting evidence is null in these text settings, with changes of −0.09 and 0.00.

Complete WikiText operating points. With natural siblings present, the family-conditioned auditor retains exact-record advantage approximately 0.99 and LR^+=97.00. At nominal 1% FPR, the clean-reference false-member rate rises from 1.4% to 81%, with 95% CI [71%, 90%]; no-sibling controls have FPR 1.9%. At nominal 5% FPR, the family-present false-member rate is 85% and the positive-verdict LR^+ falls to 1.20 (from 66.00 at the clean reference's 1% operating point). Across retraining environments, the observed false-attribution magnitude ranges from approximately 28% to 85%. With the exact passage absent, toggling family presence gives advantage 1.00, versus 0.07 for no-sibling controls. Its fixed-threshold family verdict has LR^+=8.10, TPR 0.88, and realized FPR 0.11 at nominal 5% FPR.

Complete CC-News operating points. The family-absent exact-record advantage is 0.99. With natural families present, the clean-reference false-member rate is 99.7% and LR^+ falls to 1.00, compared with 19.60 in the clean world at 5% FPR and approximately 62.00 at 1% FPR. At nominal 5% FPR, the four original best-sibling Jaccard bands $[0.50, 0.70)$, $[0.70, 0.80)$, $[0.80, 0.90)$, and $[0.90, 1.00)$ have false-member rates 80%, 88%, 82%, and 92%, respectively; no-sibling controls remain at 1.2%. Across these bands, family-presence AUC lies between 0.92 and 0.97, while the clean-reference false-member rate lies between 0.59 and 0.94. A fresh pythia-2.8B run gives 98%, Min-K% gives 97%, Min-K++ gives 82%, and EZ-MIA gives 63% false attribution at the stated 5% operating point. Across the four Jaccard bands, the Min-K variants range from 86% to 100%.

Information-relocation frontier. Across k=1 to 16, the exact-given-family point estimate falls from 0.09 to 0.01 while the family-presence estimate rises from 0.12 to 0.48. The smallest observed maximum of the two advantages is 0.12 at k=1. Neighboring bootstrap intervals overlap, so this finite sweep supports a descriptive signal-redistribution pattern; it does not establish a precise monotone law or a universal lower bound. In k=1, 2, 4, 8, 16 order, crossed-bootstrap 95% CIs are respectively $[0.01, 0.21]$, $[0.00, 0.17]$, $[0.00, 0.15]$, $[0.00, 0.12]$, and $[0.00, 0.14]$ for exact-given-family advantage, and $[0.01, 0.24]$, $[0.10, 0.34]$, $[0.15, 0.41]$, $[0.24, 0.52]$, and $[0.33, 0.62]$ for family-presence advantage. The Pythia reference intervals are $[0.45, 0.56]$ and $[0.93, 0.97]$.