CONTAMMORPH: DO TRANSFORMED VIEWS ADD MEMBERSHIP EVIDENCE AFTER FINE-TUNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Membership inference asks whether a passage was used to train a model. Auditors can average scores across reformatted or rewritten versions, but a detection gain may come from one transformation alone. We compare the original text, a fixed transformation, and their average to separate these effects. We randomly include or withhold short English passages during fine-tuning, using two corpora and models with 0.5 to 1.5 billion parameters. Our main case study is error-zone membership inference (EZ-MIA), which divides accumulated prediction improvements by accumulated losses. Line breaks improve this detector before averaging, largely through the released code's handling of a zero denominator. When some incorrectly predicted text pieces improve and none worsen, the code assigns zero, whereas the published rule ranks the passage first. One shared positive denominator constant restores strong detection in four model and corpus settings and in fresh training runs. At an intermediate training strength, the implementation choice moves EZ-MIA from the top two of six detectors to near the bottom. After correction, averaging rarely beats an original-text detector selected on other runs. We observe this pattern in twelve-run replications, when included passages form 5% of a larger training set, and with five separately generated paraphrases. Those paraphrases distinguish members from controls less well than the exact wording used in training, and their average performs worse. Auditors should compare every aggregate with both the original text and a prespecified transformed version.

1 INTRODUCTION

A model may perform well on a benchmark because it learned the task or because it saw the benchmark during training. Membership inference addresses part of this problem: given a passage and access to a model, an auditor asks whether the passage was used for training. The answer matters for data-use audits and for interpreting benchmark results.

One way to audit a passage is to score several versions, or *views*, and average their scores. A view might change the line breaks or rewrite the wording. Better detection then leaves two explanations open: one transformation may already help, or averaging may add evidence beyond it. We ask whether averaging improves detection beyond a single prespecified transformed view. ContamMorph is our protocol and experimental testbed for separating these contributions.[1]

Membership attacks score passages to distinguish training from withheld text. Neighbourhood comparison averages prediction losses over word-substituted versions, called neighbours (Mattern et al., 2023). Self-calibrated probabilistic variation (SPV-MIA) compares the model being audited (the target) with a separately trained reference model and averages over paraphrased neighbours (Fu et al., 2024). Semantic attacks learn from many perturbed inputs (Mozaffari & Marathe, 2024). To obtain known labels for our comparisons, we randomly include passages in fine-tuning, the additional training of a pretrained model, and withhold others. These groups are *members* and *controls*; neither label establishes pretraining history.

A line-break example shows why the comparison matters. The released error-zone membership inference (EZ-MIA) implementation (Ilić et al., 2026) scores a training passage zero as a paragraph and 2.67 with shorter lines (Figure 1), although its words and membership are fixed. Tokens are

[1] Code and data: `https://anonymous.4open.science/r/ICLR-ContamM`.

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.

Reading key. E: original text; W: fixed-width line breaks; S: sentence-end breaks; R: paraphrase; O: average score over E, S, W. U is the released rule; F is the published rule. The error zone contains incorrectly predicted tokens. Δ log-prob is the change in token log probability from the original model, measured in nats (natural-log units). AUROC (area under the receiver operating characteristic curve) measures member–control separation (.5: chance, 1: perfect). A fit is a training run; a seed fixes its random choices. A cell groups runs with the same model, corpus and training setting. Mixed-view training cycles renderings; dose matching equalizes presentations per view.

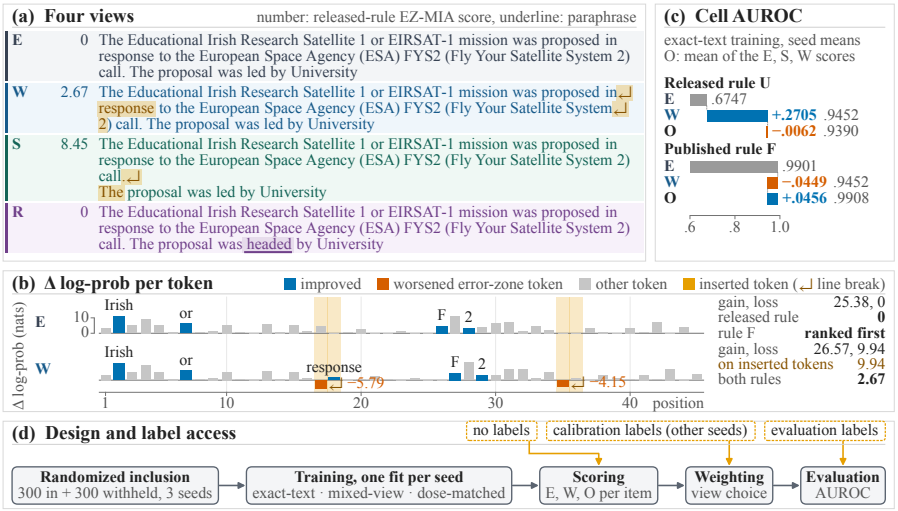


Figure 1: EZ-MIA case study and randomized-inclusion design. Line breaks change the released score through its zero-denominator rule.

the text pieces a model predicts. EZ-MIA compares log probabilities before and after fine-tuning only where the target still predicts incorrectly. It divides total improvements by total losses. If some tokens improve and none worsen, the denominator is zero. The released code then returns zero, whereas the published rule treats this case as strong membership evidence. Line breaks introduce losses that make the ratio finite.

This change affects detection across passages. After two full passes through randomly selected 32-word prefixes from WikiMIA, a collection of Wikipedia passages with date labels (Shi et al., 2024), line breaks alone raise Qwen2.5-0.5B's AUROC from .6747 to .9452, where .5 denotes chance and 1 perfect separation. We observe the same mechanism with pythia-1.4b and independently randomized arXiv abstracts. Replacing the line-break contribution to the denominator with one positive constant shared by all passages restores strong detection in four model and corpus settings; fresh training runs reproduce this recovery.

The zero-score fallback also changes detector rankings when simpler scores are imperfect. At a learning rate of 5×10^{-5} for one pass, likelihood (mean token log probability) reaches .89 to .92 AUROC. Yet the released rule drops EZ-MIA from the top two of six detectors to fourth or lower, and to a mean rank of 5.75 to 6 in a replication with twelve fresh training runs per setting.

We then test whether averaging helps after correcting the scoring rule. We compare against an original-text detector selected on other sources and training runs, using twelve fresh runs per model, corpus, and training-strength setting, average up to five separately generated paraphrases per passage, and embed members in a training mixture twenty times larger than the member set. Averaging rarely beats the selected detector. Paraphrase scores separate members and controls less well than scores on the wording used in training, and their average performs worse. The practical lesson is to report both comparisons before crediting a gain to aggregation. Our empirical conclusions concern short-passage fine-tuning of models with at most 1.5B parameters.

Related work. Transformed text can change what a benchmark measures or provide another way to score the same item. Contamination studies evaluate models on rephrased benchmarks (Yang et al., 2023), infer contamination from performance (Dekoninck et al., 2024b), or study mitigation under altered text (Sun et al., 2025). STAMP compares public watermarked rephrasings with private controls (Rastogi et al., 2025), and other methods exploit canonical order (Oren et al., 2024),

recognition across variants (Golchin & Surdeanu, 2025), or quizzes that mix verbatim text with paraphrases (Duarte et al., 2024). Aggregation across paragraphs or documents can improve membership inference (Puerto et al., 2025; Maini et al., 2024), and ReCaLL ensembles scores over several prefixes (Xie et al., 2024). These methods add content or context, whereas our views repeat one passage, so we ask what an additional rendering contributes beyond a single view.

For item-level inference, likelihood and zlib scores go back to extraction attacks (Carlini et al., 2021), and reference models account for item difficulty (Watson et al., 2022; Carlini et al., 2022). Our exact-text comparators include likelihood-based attacks (Shi et al., 2024; Zhang et al., 2025b) and reference-based EZ-MIA (Ilić et al., 2026). Our neighbourhood (Mattern et al., 2023) and keyword-based Tag&Tab (Antebi et al., 2025) comparators are local adaptations. SPV-MIA calibrates prediction changes against a reference model trained on target-generated text (Fu et al., 2024), and we also score Infilling Score, built from token likelihoods that also use later tokens as context (Raoof et al., 2025). We do not test the Kernel Divergence Score (Choi et al., 2025), which measures dataset-level leakage, or calibration against a reference corpus or within sliding text windows (Zhang et al., 2024; Chen et al., 2026). A false-positive rate (FPR) is the fraction of withheld passages flagged as members; the true-positive rate (TPR) is the fraction of members correctly flagged. Reporting TPR at a chosen FPR gives an operating point, complementing AUROC (Carlini et al., 2022).

Date-based proxies for members and controls differ in distribution. A blind classifier, which uses the text without querying the audited model, can therefore separate them (Duan et al., 2024; Das et al., 2025; Maini et al., 2024; Meeus et al., 2025). Randomized inclusion provides known membership labels, as with canaries or traps, sequences deliberately inserted into training (Zhang et al., 2025a; Meeus et al., 2024), and recent work defines membership effects causally (Even et al., 2026) or builds testbeds from training checkpoints (Wang et al., 2026b). Evasion and post-training suppression (Dekoninck et al., 2024a; Wang et al., 2026a) make training history essential to interpretation. We randomize inclusion within source groups, then compare views within each trained model to separate transformation from averaging.

2 Measuring what transformed views add

We compare the original input, one prespecified transformation, and the average across all views. Suppose line wrapping improves a detector before we combine any scores. Comparing the average only with the original text would credit that improvement to aggregation. Comparing all three quantities separates the change caused by wrapping from the additional change caused by averaging.

Let x be a passage and s a scoring function whose larger values indicate membership; $s(x)$ is its score on the original passage. We evaluate m views indexed by j, including the original. Each transformation T_j produces a view $T_j(x)$ intended to preserve content. We call this set the item's *orbit*; the name imposes no group structure on the transformations. Choose a transformation index w before scoring. The exact score is $s_E = s(x)$ and the fixed-view score is $s_W = s(T_w(x))$. The *orbit mean* is $s_O = m^{-1} \sum_{j=1}^{m} s(T_j(x))$. The fixed view W is the line-wrapped (width) view in the layout study and the paraphrase in the semantic study. Writing E, W, and O for the corresponding AUROCs, we decompose the total gain exactly as $O - E = (W - E) + (O - W)$: the *transformation increment* $W - E$ and the *averaging increment* $O - W$. The orbit beats both comparators exactly when $O - \max(E, W) > 0$. Paired intervals keep each passage's views together when quantifying uncertainty in these differences.

The layout study uses exact text (collapsed whitespace, no line breaks), sentence-end breaks (S), and fixed-width breaks (W, 12 words per line, 3 lines), which all preserve the words. The paired semantic study uses screened WikiMIA exact and paraphrase pairs ($m = 2$), with R denoting the paraphrase. A later study adds five paraphrases $R_1, \ldots, R_5$ generated in separate calls. Let k count the paraphrases included: O_k denotes AUROC after averaging the original score with the first k paraphrase scores, so O_5 uses six texts. Separate generation does not make the resulting scores independent.

The fine-tuned model is the *target*; its untouched pretrained checkpoint is the *base*, used as a reference for changes in prediction. We evaluate six core detectors. Likelihood (LL) is the mean token log probability, and zlib divides this value by the text's byte length after zlib compression.

In Min-$k\%$, k is a percentage (20 here), distinct from the paraphrase count above. It averages the lowest 20% of token log probabilities (Shi et al., 2024), and Min-k++ applies the same operation after normalizing scores against the model's vocabulary distribution (Zhang et al., 2025b). These four detectors query only the fine-tuned target model. Reference gain subtracts the untouched base model's mean log probability. EZ-MIA (Ilić et al., 2026) also compares target and base, but only where the target's most likely next token is wrong. With x_t the token at position t, $x_{<t}$ its prefix, v a vocabulary token, and θ the target parameters, write p_θ and p_0 for target and base next-token probabilities. Conditioning on $x_{<t}$ means using the preceding tokens. The *error zone* is $Z = \{t \geq 2 : \arg\max_v p_\theta(v \mid x_{<t}) \neq x_t\}$. Its size $|Z|$ counts the errors; the initial token is unscored. All logarithms are natural, so changes have units of nats. Let Δ_t be the change in the log probability of the observed token: $\Delta_t = \log p_\theta(x_t \mid x_{<t}) - \log p_0(x_t \mid x_{<t})$. We write the positive gain as $P = \sum_{t \in Z} \max(\Delta_t, 0)$ and the negative mass (total worsening) as $N = \sum_{t \in Z} \max(-\Delta_t, 0)$. The ratio P/N is large when gains outweigh losses, which EZ-MIA treats as evidence of membership. Membership signal means how well scores separate members from controls; it differs from average score magnitude. For a stronger comparator, our held-out selector chooses among all six original-text detectors using other training runs and source passages. It excludes the evaluated run and sources. We also report the evaluation-label best detector as a descriptive comparator.

We distinguish the released implementation, U, from the published edge-case rule, F. U returns zero when the error zone has fewer than two tokens or the denominator is zero. F instead ranks the positive-gain, zero-loss case first and specifies separate rules for short error zones (Ilić et al., 2026, App. E.5 and E.6). We use F to interpret the findings and write EZ(F) for EZ-MIA under F (and similarly for other rules). Earlier sensitivity rules M and E1 impute, or replace, selected zeros using positive scores pooled without labels within each view and fit. M uses the median for short zones and the maximum otherwise; E1 uses the maximum for both. Appendix Table 16 gives every edge case, tie rule, and averaging constant. The artifact includes the audited upstream code (commit `aa0b6b7`) and its license.

A strictly increasing score transformation preserves a single view's AUROC but can change its contribution to an arithmetic mean. We therefore also average within-view ranks, the ordered score positions within each view, giving tied scores their average rank. Arithmetic means under F replace ranked-first scores with twice the largest finite ratio in that view and fit; rank means need no such constant. Appendix A.1 derives when signal and dependence make averaging useful in a Gaussian model.

3 Experimental design

We randomize which passages enter fine-tuning, then score the same members and controls in every view. Pairing the scores of each passage across views holds its training history fixed when we compare detection accuracy.

Experimental unit and randomization. A *fit* is one training run with a specified random seed. A *cell* fixes the model, corpus, training exposure, and update strength, meaning how much fine-tuning changes the model. We vary this strength through the learning rate and number of passes. Each fit draws 150 members and 150 controls from each of two source strata, groups defined by date for WikiMIA or subject category for arXiv, for 300 members and 300 controls. Exact-text training presents only the original passage, whereas mixed-view training cycles through the three layout views.

Core natural-text studies. On a date-stratified, 764-source WikiMIA pool of 32-word prefixes, we train Qwen2.5-0.5B *base* with the AdamW optimizer, batch size 8 and three fits per cell. One pass at $5{\times}10^{-6}$ is our prespecified *weak* update, one pass at $5{\times}10^{-5}$ is *mid*, and two passes at $5{\times}10^{-5}$ are *strong*. We also score four passes and one pass at 5×10^{-7} as sensitivity settings. A pythia-1.4b replication reuses these assignments, and an arXiv replication uses three new assignments over 764 32-word prefixes of July 2026 abstracts, which postdate both models' pretraining data. We retain texts, token observations, loss curves, and realized training doses (Appendix B).

Replications frozen before training. The first frozen protocol tests replication across training runs and longer inputs. It trains 72 fits with fresh assignments: twelve per weak cell for both models and pools, three per strong cell, and runs on 96-word passages. The second protocol tests mid updates, paraphrase diversity, and training mixtures with 108 fits. It adds twelve fits per mid cell and twelve more per weak cell, and every one of these fits also scores five paraphrases per passage, generated in separate calls to llama3.1-8B, a model family distinct from both targets. It also embeds the 300 members in a 6,000-passage training mixture with Wikipedia background text, so the included passages, or members, form 5% of training (Appendix E.5).

What we estimate. Randomized inclusion gives known labels for comparing detection accuracy under a stated training recipe. Table 3 identifies the primary analysis for each claim. With twelve fits per cell, training runs are the unit, and we report one-sided Student t bounds for mean differences across runs. A pointwise bound addresses one comparison; Bonferroni bounds adjust for the number of comparisons in each protocol. These estimates address variation across training runs under the tested recipe and source pool.

The core studies have three fits per cell. Their 95% confidence intervals (CIs) use a bootstrap: resampling passages and fits while preserving paired scores. They condition on observed runs and fixed preprocessing; three fits do not establish coverage for new runs. We call a difference *resolved* when its interval excludes zero, and an unresolved difference is not evidence of equality.

Analysis timing. The original eight comparisons of EZ-MIA with likelihood used Holm's stepwise correction for testing multiple hypotheses (Holm, 1979). Six of them were fixed before their fits, whereas the grouping, the .01 reference margin, and rule F came after the first two results. The margin describes effect size, not a deployment threshold, so we also report detection rates at low false-positive rates. Appendix E.1 records when each analysis was fixed.

Held-out comparator. The exact-text selector operates within five source folds, disjoint groups of passage identities, using AUROC on the other fits of the cell. For each test fold, selection excludes its sources and the evaluated fit (Section 4.3).

Appendix A.5 reports exploratory synthetic code and multiple-choice studies and a dependence model. They motivate the paraphrase study but carry no main claim.

4 Results

4.1 Why line breaks improve released EZ-MIA

Line breaks improve released EZ-MIA mainly by changing which passages receive a zero score, before any scores are averaged. In the motivating exact-training study, 257 of 288 member zeros arise when no error-zone token worsened and positive gain remains. Across all eight strong cells (four model and corpus settings under both exposures), this cause accounts for 3,262 of 3,683 member zeros, with 409 arising from short zones ($|Z| < 2$). To quantify how zeros affect detection, let q and r be the member and control zero fractions and A_{++} the AUROC among positive-score pairs. For a nonnegative score, let A denote the overall AUROC. It satisfies $A = (1-q)r + qr/2 + (1-q)(1-r)A_{++} \leq 1 - q + qr/2$, with the last contribution zero if no positive pairs exist. After exact-text training, averaging removes all 288 member zeros but lowers A_{++} from .9919 to .9390, and on items positive under both scores, exact text and orbit remain close (.9919 and .9897).

To check that the zeros come from the released rule, we use a scalar reimplementation. It matches the upstream score on all 129,600 item-view observations, including every zero, confirming the zero pattern in the released implementation (Table 1).

Line breaks remove zeros by adding losses to the denominator. The affected *boundary tokens* contain a line break or follow one, never occur in exact-text training, and carry 85.9% of member negative error-zone mass. Removing all boundary positions returns the member zeros from 1 to 277 and lowers AUROC from .9452 to .6842, but it also changes the numerator and error count. A second retrospective intervention therefore fixes both at their width values and removes only the boundary negative mass. Across four model and corpus cells, AUROC falls into the range .336 to

Here q, r are member/control zero fractions, z_1, z_0 their counts, A overall AUROC, and A_{++} AUROC among positive-score pairs. Shared-positive restricts both scores to positive items. N_b is loss on boundary tokens, c its replacement constant; memb./ctrl. mean members/controls. Zero causes list short zones/no worsening/zero gain. Ceiling gap is the support bound minus AUROC; EZ tokens gives the mean error count; CI denotes confidence interval.

Table 1: Support decomposition and denominator-only intervention for the released EZ-MIA rule.

		Zeros of 900			Positive scores		Shared-positive AUROC		Boundary	EZ tokens	
View	AUROC	z_1	Zero causes	z_0	A_{++}	Ceiling gap	exact	view	mass (%)	memb.	ctrl.
A. Exact-text training											
exact	**.6747**	288	28/257/3	1 (0/0/1)	**.9919**	.0055	–	–	0.0	7.7	23.0
sentence	.8497	105	11/92/2	2 (0/0/2)	.9617	.0337	.9918	.9687	76.0	8.8	23.6
width	.9452	1	0/0/1	0	.9462	.0537	.9919	.9563	85.9	10.7	25.2
orbit	**.9390**	0	–	0	**.9390**	.0610	.9919	.9897	–	–	–
paraphrase	.8793	88	3/83/2	1 (0/0/1)	.9744	.0230	.9918	.9752	0.0	10.0	23.6
B. Mixed-view training											
exact	.7387	225	17/208/0	0	.9848	.0113	–	–	0.0	8.3	23.1
sentence	.6843	276	9/267/0	1 (0/1/0)	.9866	.0092	.9853	.9864	3.3	8.7	23.4
width	.7581	213	0/213/0	1 (0/1/0)	.9930	.0054	.9847	.9931	15.9	9.7	24.7
orbit	.8304	145	–	0	.9899	.0084	.9848	.9923	–	–	–
paraphrase	.9002	69	3/66/0	0	.9750	.0231	.9847	.9737	0.0	10.6	23.7

Qwen2.5-0.5B on WikiMIA after two strong passes, released rule, seed means; bold: cited in the text. z_1, z_0: member and control zeros (causes: short zone ($|Z| < 2$) / no worsening / zero positive mass); A_{++}: AUROC among positive scores; ceiling gap $= 1 - q + qr/2 - A$; shared-positive AUROC: exact text and view scored on the items positive under both; boundary mass: share of member negative error-zone mass on tokens that contain or follow a view-introduced line break; EZ tokens: mean error-zone length; –: not applicable.

	Observed	Erased	1-nat replacement	Fixed-c	Exact	Exact	Exact	Exact	Exact
Model / corpus	W	N_b	[95% CI]	range	U	F	$P/(N+1)$	$-N$	$-\lvert Z\rvert$
Qwen/WikiMIA	.945	.687	.984 [.975, .991]	.936–.991	.675	.990	.973	.994	.983
pythia/WikiMIA	.836	.375	.961 [.943, .976]	.886–.969	.367	.980	.945	.995	.994
Qwen/arXiv	.953	.337	.981 [.964, .993]	.954–.985	.324	.993	.967	1.000	1.000
pythia/arXiv	.889	.336	.974 [.959, .986]	.915–.979	.309	.992	.947	.999	.999
Qwen/WikiMIA new	.955	.671	.981 [.970, .991]	.947–.985	.654	.989	.978	.993	.986
pythia/WikiMIA new	.845	.409	.963 [.948, .977]	.894–.970	.375	.986	.954	.994	.993
Qwen/arXiv new	.960	.334	.989 [.978, .998]	.959–.990	.295	.996	.974	1.000	1.000
pythia/arXiv new	.897	.323	.971 [.955, .985]	.920–.977	.291	.992	.958	.998	1.000

C. Denominator intervention. Fixed $c \in \{.1, 1, 10\}$. Erasure and replacement hold width-view positive mass and error count fixed; N_b: boundary negative mass. Replacement uses the same constant for both classes. Exact-view columns score the original text with the released ratio U, convention F, the stabilized ratio $P/(N+1)$, negative mass alone and error-zone size alone. Rows marked new are four independent strong cohorts with fresh assignments, trained and scored after the first replication protocol was frozen. Exact-text training, two strong passes, three fits per cell; paired source/fit conditional intervals.

.687. Replacing the removed mass with the *same* constant for every member and control, one nat (a natural-log unit), restores .961 to .984. The paired improvement over erasure is .297 to .644, with all four conditional intervals above zero (Table 1 and Figure 2b). We fixed one nat before the calculation, without tuning on labels. Constants of .1 and 10 nats give similar recovery (Appendix B).

The same stabilization works on the original text, which has no boundary tokens. Exact-text AUROC ranges from .31 to .68 under U and from .945 to .973 with $P/(N+1)$, while negative mass or error-zone size alone reaches .98 to 1.00. Four fresh strong cohorts trained under a frozen protocol reproduce the pattern: erasure lowers width AUROC to .32 to .67, the one-nat replacement restores .963 to .989, and $P/(N+1)$ gives .954 to .978 (Table 1C). The intervention locates the collapse in the singular ratio. Strongly trained members have fewer mispredicted tokens and those tokens worsen less, leaving many denominators at zero. A shared constant suffices without passage-specific boundary losses, although other transformations could still carry information.

Where the fallback matters. Strong updates push likelihood and reference gain to .995 to 1.000, so there the main consequence is a misleading comparison between methods. Mid updates are different. After one pass at 5×10^{-5}, likelihood reaches only .89 to .92, yet U lowers EZ-MIA by .082 (Qwen) and .157 (pythia) relative to F, with 74 and 142 member zeros. Under F, EZ-MIA leads likelihood by .084 and .041 and ranks second of six detectors behind reference gain. Under U, it ranks fourth for Qwen and last for pythia, where it trails likelihood by .116. Twelve fresh fits per mid cell, in both corpora and models, confirm the pattern: U costs EZ-MIA .091 to .227 AUROC, while likelihood reaches only .904 to .955 (Figure 2a and Appendix Table 23).

Reading key. LL: mean token log probability; Gain: change from base to target; U/F: released/published EZ-MIA rules. Weak, mid and strong name the training recipes in Section 3. $O_k - E$ compares averaging original text with k paraphrases against original text alone; k counts paraphrases, not total views. F averages within-view ranks; LL averages raw scores. Open markers identify twelve-fit arXiv mid cells; panel (c) shows fit-level means and 95% intervals. Its "independent paraphrases" means separate generation calls, not independent scores.

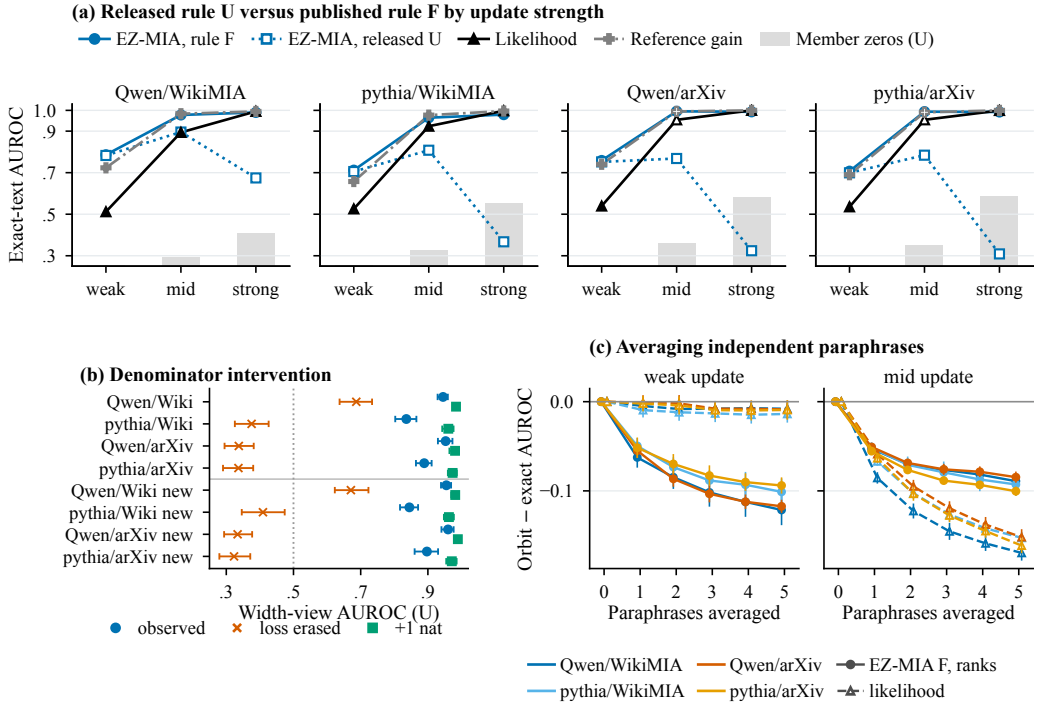


Figure 2: (a) Exact-text detection by update strength; (b) denominator intervention in original and fresh strong cohorts; (c) paraphrase averaging relative to exact text.

4.2 Reference-based scores lead after weak updates

EZ-MIA leads likelihood after weak updates and trails it after strong updates. The weak-update lead is much larger than the strong-update deficit. After one pass at 5×10^{-6}, where M, E1, and F coincide, EZ-MIA reaches .71 to .79 AUROC and likelihood .51 to .54, a lead of .17 to .27 across eight cells (Table 2). Reference gain, which also uses the base checkpoint, reaches .66 to .74, in line with difficulty calibration (Watson et al., 2022). Across twelve fresh training runs per weak cell, EZ(F) minus likelihood is .162 to .267, with one-sided lower bounds of .151 to .245. After two strong passes, likelihood and reference gain reach .995 to 1.000, and EZ-MIA under F trails likelihood by .005 to .018. Over six fits per setting, every 95% interval of this deficit lies below zero, but the deficit is small because both scores approach 1. Holm adjustment of the original three-fit family retains 5 of eight contrasts under F (four weak and one strong) and all eight under M (Appendix B). At 1% FPR, EZ(F) exceeds likelihood by .05 to .13 true-positive rate in the twelve-fit weak cells.

4.3 After correction, aggregation gains are small

Under U, width increments are .27 to .63 after exact-text training, and averaging increments are .07 to .16 after mixed-view training. After the scoring rule is corrected, the gains shrink. The following width and orbit ranges use the historical imputation rule M; the held-out comparison below uses the published rule F. Width lowers strong-update EZ-MIA by .02 to .12 after exact-text training, and the orbit returns to about the exact-text score. After mixed-view training, width and the orbit add at most .015, and neither reaches the best exact-text likelihood or reference-gain score (Appendix Table 19).

LL: likelihood; Gain: reference gain; EZ: EZ-MIA. U/M/F name scoring rules. Update magnitude uses parameter change $\Delta\theta$ relative to base parameters θ; $\|\cdot\|$ denotes the Euclidean norm. CI: confidence interval; n/a: unavailable. $*$: interval excludes zero; •: prespecified weak rate; †: paired two-pass trajectories.

Table 2: Exact-text AUROC and detector differences by update strength and scoring convention.

			Exact-text AUROC					AUROC differences			
Update	Train	Rel. update 10^{-3}	LL	Gain	EZ (F)	EZ (U)	Zeros /900	EZ (F)−LL mean [95% CI]	EZ (M) −LL	EZ (M) −Gain	$O-E$ (M)
Qwen2.5-0.5B, WikiMIA											
weak•	exact	0.30	.5119	.7233	.7861	.7828	3	**.274** [.215, .330]	.274*	.063*	−.002
weak•	mixed	0.31	.5110	.7173	.7846	.7813	3	.274 [.213, .326]*	.274*	.067*	.004
2×5e-6	exact	0.49	.5295	.7933	.8584	.8484	9	.329 [.270, .384]*	.329*	.065*	−.006
2×5e-6	mixed	0.50	.5274	.7858	.8541	.8507	3	.327 [.269, .381]*	.327*	.068*	.005
mid	exact	8.19	.8940	.9838	.9779	.8957	74	.084 [.057, .113]*	.084*	−.006	−.002
mid	mixed	8.26	.8879	.9828	.9769	.9269	45	.089 [.063, .116]*	.089*	−.006	.005*
strong†	exact	n/a	.9951	.9947	.9901	.6747	288	−.005 [−.013, .001]	−.009*	−.009*	−.001
strong†	mixed	n/a	.9949	.9950	.9875	.7387	225	−.007 [−.015, −.001]*	−.009*	−.009*	.008*
pythia-1.4b, WikiMIA											
weak•	exact	0.30	.5256	.6574	.7120	.7053	8	**.186** [.142, .229]	.186*	.055*	−.003
weak•	mixed	0.31	.5246	.6610	.7218	.7151	9	.197 [.153, .240]*	.197*	.061*	−.003
2×5e-6	exact	0.45	.5459	.7115	.7963	.7863	9	.250 [.203, .297]*	.250*	.085*	−.005
2×5e-6	mixed	0.47	.5432	.7093	.7959	.7881	8	.253 [.207, .298]*	.253*	.087*	−.003
mid	exact	6.64	.9237	.9770	.9647	.8081	142	.041 [.018, .066]*	.041*	−.012*	−.001
mid	mixed	6.69	.9212	.9792	.9739	.8750	89	.053 [.033, .072]*	.053*	−.005	.005*
strong	exact	8.56	.9979	.9968	.9795	.3672	562	**−.018** [−.034, −.007]	−.039*	−.038*	−.012*
strong	mixed	8.63	.9975	.9967	.9870	.5028	438	−.011 [−.021, −.003]*	−.020*	−.020*	.015*
Qwen2.5-0.5B, arXiv											
weak•	exact	0.32	.5394	.7404	.7586	.7519	6	**.219** [.151, .286]	.219*	.018	.000
weak•	mixed	0.32	.5380	.7362	.7497	.7419	7	.212 [.143, .278]*	.212*	.013	.001
strong	exact	12.99	1.0000	1.0000	.9931	.3242	607	−.007 [−.016, −.000]*	−.027*	−.027*	.005
strong	mixed	13.12	1.0000	1.0000	.9817	.4972	449	−.018 [−.038, −.004]*	−.018*	−.018*	.014*
pythia-1.4b, arXiv											
weak•	exact	0.31	.5355	.6902	.7071	.6993	7	**.172** [.117, .226]	.172*	.017	−.006
weak•	mixed	0.32	.5338	.6909	.7090	.7090	0	.175 [.115, .234]*	.175*	.018	.003
strong	exact	9.89	.9994	.9999	.9922	.3089	620	−.007 [−.019, .000]	−.041*	−.042*	−.007
strong	mixed	9.82	.9998	1.0000	.9946	.4490	494	−.005 [−.012, −.001]*	−.014*	−.015*	.012*

Reading the table. Rel. update: $\|\Delta\theta\|/\|\theta\|$; Train: exact-text or mixed-view exposure. LL: likelihood; Gain: reference gain; EZ (F/U/M): published/released/imputed rules. F is primary, and M columns give the originally prespecified contrasts. Zeros: released-rule member zeros; n/a: weights not retained. **Marks.** • prespecified weak update; † paired two-pass trajectories; $*$ unadjusted 95% interval excludes zero. Bold: prespecified primary contrasts surviving Holm adjustment at .05 under F (five of eight, all eight under M).

Held-out exact-text comparison. We use the source- and run-held-out exact-text selector from Section 3. Strong cells leave at most .005 of headroom below perfect AUROC, so the informative comparisons are the weak and mid ones. In the original three-fit cells, 16 of these 18 width and orbit comparisons have pointwise one-sided 95% upper bounds below .01, and 9 have Bonferroni bounds below .01 (Table 12). The twelve-fit replications test variation across new training runs. In the twelve-fit weak cells, all 12 pointwise and 9 of 12 Bonferroni bounds fall below .01, and the remaining cases are arXiv orbits, where only Qwen has a positive mean (.005, Bonferroni bound .013). In the new twelve-fit mid cells, 12 of 12 pointwise and 12 Bonferroni bounds fall below .01. At 1% FPR, the orbit gains over exact EZ-MIA are unresolved in three weak cells and reach .017 and .021 true-positive rate for the two Qwen arXiv orbits, the only weak cell with a resolved gain.

The averaging rule and other detectors. When mixed-view training matches the number of presentations of each view in exact-text training (dose matching), arithmetic averaging under F retains a resolved $O-W$ of .017 [.006,.032] and .028 [.015,.044], but within-view ranks lower it to −.017 [−.038,.003] and −.012 [−.031,.008], so this gain depends on the averaging rule (Table 17). Likelihood, zlib, Min-k%, and Min-k++ show no resolved view-over-exact gain in the 16 weak and strong cells, and averaging often only recovers a transformation's loss: width lowers Min-k++ by up to .27, and the orbit beats width by .07 to .19. Our Tag&Tab adaptation trails the selected exact detector in all 20 prespecified rows, and Infilling Score stays near chance after weak updates (Appendix B).

Adapted SPV-MIA gives the single resolved orbit gain over its own exact score (.023 in the Qwen arXiv weak cell), but its gain over the width view is unresolved, and prospective own-exact gains remain unresolved (.003 and .008, Table 20).

Table 3: Claims, primary analyses and outcomes.

Claim	Primary analysis	Outcome
Width helps released EZ-MIA by removing zeros	One shared constant in the denominator, in 8 cohorts	AUROC returns to .961 to .989 in all eight cohorts
At mid updates, the fallback reorders the detectors	U against F in four mid cells, twelve fits each	U loses .091 to .227 AUROC and ranks EZ-MIA near last
After weak updates, EZ-MIA leads likelihood	EZ(F) minus likelihood over twelve fits in each cell	Leads by .16 to .27 with every lower bound above zero
Layout orbits add little	Orbit minus the held-out exact detector, twelve fits	Bounds below .01 in 9 of 12 weak and 12 of 12 mid tests
Paraphrases dilute the evidence	Five-paraphrase average, 96 fits	Below exact text by .085 to .121
Weak-update lead persists in a mixture	5% member share, 12 fits	Weak lead holds, and orbit bounds reach up to .021

4.4 Diverse paraphrases and training mixtures

Separately generated paraphrases. Layout views repeat almost the same evidence, so we also averaged up to five separately generated paraphrases per passage, in twelve fresh fits per weak and mid cell. We summarize dependence by the *effective number of views*, the number of equally variable independent scores with the same variance reduction under averaging. This diagnostic measures redundancy, not detection accuracy. The paraphrase scores are less redundant than the layout scores (an effective 1.41 to 1.86 independent views out of six for likelihood), but each paraphrase separates members from controls less well than the exact wording used in training. The average also performs worse. Averaging the original with one paraphrase lowers EZ-MIA by .050 to .063 AUROC, averaging with five lowers it by .085 to .121 (Figure 2c), and the likelihood average changes by $-.170$ to $-.008$. The five-paraphrase average stays below the held-out exact detector (AUROC differences $-.120$ to $-.084$), and averaging all eight views does not beat the best single view chosen on other fits (AUROC differences $-.071$ to $-.041$). This agrees with the dependence model: less redundant views can still weaken detection if they separate the classes less well (Appendix A.5).

A training mixture. When the 300 members form 5% of a one-pass fine-tuning set, EZ-MIA under F still leads likelihood, by .183 to .188 after weak and .073 to .128 after mid updates, and U still costs .037 to .121 AUROC at the mid rate. No averaging gain over the held-out exact detector is resolved; with three fits per cell the weak-update bounds reach .021, so our frozen prediction that all bounds fall below .01 fails (Appendix Table 25).

4.5 What an auditor should report

To separate transformation from averaging, report the scoring rule, zero rates, per-view scores, query cost (Table 6), and comparisons with both the exact text and a fixed transformed view. Date labels do not verify membership: frozen embeddings predict WikiMIA's date labels at .973 to .990 AUROC, as blind baselines anticipate (Das et al., 2025).

5 Conclusion

Implementation and deployment. Our neighbourhood, Tag&Tab, and SPV-MIA comparators are local adaptations, Infilling Score lacks released code, and the Kernel Divergence Score is untested. Some historical weights are unavailable (Table 5). We do not infer misconduct from statistical anomalies, and all data are public.

Conclusion. A useful transformation need not imply useful aggregation. Released EZ-MIA's largest gains arise from a singular ratio. Fresh cohorts reproduce this mechanism, which also reorders detectors at mid updates where simpler scores are not saturated. Corrected averages rarely improve on the selected exact-text detector across replications, paraphrases, and the mixture, whose bounds remain wider. Auditors should disclose scoring edge cases and compare aggregates with both original text and a prespecified transformed view.

REPRODUCIBILITY STATEMENT

We keep the texts, model outputs, and code behind every result. Scripts regenerate each table, figure, and reported number from these saved outputs, and a short test runs on an ordinary CPU. Repeating the training itself requires the model weights and the recorded settings. We keep the weights of 18 fine-tuned models and the saved outputs, without weights, of the other runs, and Table 5 marks which experiments can be rerun from scratch. We will share an anonymized copy of these materials with reviewers through OpenReview. All models are open-weight, and we used no paid service.

STATEMENT ON THE USE OF AI ASSISTANTS

We used AI assistants to polish the writing and as coding assistants. They did not produce any detector score or train the evaluated models. An open-weight model (llama3.1-8B) generated the paraphrases used as study inputs. We reviewed every AI-assisted edit, checked assisted code against saved outputs and tests, verified citations and numbers against their sources, and take responsibility for the final text, claims, code, and data.

REFERENCES

Sagiv Antebi, Edan Habler, Asaf Shabtai, and Yuval Elovici. Tag&Tab: Pretraining data detection in large language models using keyword-based membership inference attack. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5273–5286, 2025. doi: 10.18653/v1/2025.findings-emnlp.283. URL https://aclanthology.org/2025.findings-emnlp.283/.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649. URL https://doi.org/10.1109/SP46214.2022.9833649.

Yuetian Chen, Yuntao Du, Kaiyuan Zhang, Ashish Kundu, Charles Fleming, Bruno Ribeiro, and Ninghui Li. Window-based membership inference attacks against fine-tuned large language models. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 1187–1206, 2026. URL https://www.usenix.org/conference/usenixsecurity26/presentation/chen-yuetian.

Hyeong Kyu Choi, Maxim Khanov, Hongxin Wei, and Yixuan Li. How contaminated is your benchmark? measuring dataset leakage in large language models with kernel divergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 10666–10682, 2025. URL https://proceedings.mlr.press/v267/choi25b.html.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 118–125, 2025. doi: 10.1109/SPW67851.2025.00016. URL https://doi.org/10.1109/SPW67851.2025.00016.

Jasper Dekoninck, Mark Niklas Müller, Maximilian Baader, Marc Fischer, and Martin Vechev. Evading data contamination detection for language models is (too) easy. *arXiv preprint arXiv:2402.02823*, 2024a. URL https://arxiv.org/abs/2402.02823.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. ConStat: Performance-based contamination detection in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92420–92464, 2024b. doi: 10.52202/079017-2935. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a7f89793b9e6f8c6568dbbb6ff727b9b-Abstract-Conference.html.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=av0D19pSkU.

André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-COP: Detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 11940–11956, 2024. URL https://proceedings.mlr.press/v235/duarte24a.html.

Mathieu Even, Clément Berenfeld, Linus Bleistein, Tudor Cebere, Julie Josse, and Aurélien Bellet. Causal evaluation of membership inference attacks. *arXiv preprint arXiv:2602.02819*, 2026. URL https://arxiv.org/abs/2602.02819.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *Advances in Neural Information Processing Systems*, volume 37, pp. 134981–135010, 2024. doi: 10.52202/079017-4290. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f36ad694188bb4c4bbbd61e2038e069e-Abstract-Conference.html.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830, 2025. doi: 10.1162/tacl.a.20. URL https://aclanthology.org/2025.tacl-1.37/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

David Ilić, David Stanojević, and Kostadin Cvejoski. Powerful training-free membership inference against fine-tuned autoregressive language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14077–14093, 2026. doi: 10.18653/v1/2026.acl-long.640. URL https://aclanthology.org/2026.acl-long.640/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4. URL https://doi.org/10.1016/S0047-259X(03)00096-4.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems*, volume 37, pp. 124069–124092, 2024. doi: 10.52202/079017-3941. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e01519b47118e2f51aa643151350c905-Abstract-Conference.html.

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 2023*, pp. 11330–11343, 2023. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719/.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. Copyright traps for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 35296–35309, 2024. URL https://proceedings.mlr.press/v235/meeus24a.html.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. SoK: Membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401, 2025. doi: 10.1109/SATML64287.2025.00028. URL https://doi.org/10.1109/SATML64287.2025.00028.

Hamid Mozaffari and Virendra J. Marathe. Semantic membership inference attack against large language models. In *Red Teaming GenAI Workshop at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=pHEGIrqcNx.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations*, pp. 16354–16372, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/46e624c244cff669223d488defd4e835-Abstract-Conference.html.

Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4165–4182. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.234. URL https://aclanthology.org/2025.findings-naacl.234/.

Negin Raoof, Litu Rout, Giannis Daras, Sujay Sanghavi, Constantine Caramanis, Sanjay Shakkottai, and Alexandros G. Dimakis. Infilling score: A pretraining data detection algorithm for large language models. In *International Conference on Learning Representations*, pp. 35128–35144, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5709163243753c9c9ab7f4b4e5a8766d-Abstract-Conference.html.

Saksham Rastogi, Pratyush Maini, and Danish Pruthi. STAMP your content: Proving dataset membership via watermarked rephrasings. In *International Conference on Machine Learning*, volume 267, pp. 51247–51272, 2025. URL https://proceedings.mlr.press/v267/rastogi25a.html.

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*, pp. 51826–51843, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e32ad85fa27be4a9868d55703f01323e-Abstract-Conference.html.

Yifan Sun, Han Wang, Dongbai Li, Gang Wang, and Huan Zhang. The emperor's new clothes in benchmarking? a rigorous examination of mitigation strategies for LLM benchmark data contamination. In *International Conference on Machine Learning*, volume 267, pp. 57728–57753, 2025. URL https://proceedings.mlr.press/v267/sun25t.html.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models. In *International Conference on Learning Representations*, pp. 100419–100460, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a3860475ddbfb8c644c43f0dcd266b66-Abstract-Conference.html.

Jeffrey George Wang, Jason Wang, Marvin Li, and Seth Neel. CheckMIABench: Firm foundations for membership inference attacks on language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 364–370, 2026b. doi: 10.18653/v1/2026.acl-short.30. URL https://aclanthology.org/2026.acl-short.30/.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3eIrli0TwQ.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. ReCaLL: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, 2024. doi: 10.18653/v1/2024.emnlp-main.493. URL https://aclanthology.org/2024.emnlp-main.493/.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023. URL https://arxiv.org/abs/2311.04850.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Position: Membership inference attacks cannot prove that a model was trained on your data. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 333–345, 2025a. doi: 10.1109/SATML64287.2025.00025. URL https://doi.org/10.1109/SATML64287.2025.00025.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-K%++: Improved baseline for detecting pre-training data from large language models. In *International Conference on Learning Representations*, pp. 64845–64862, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a2e3b4132ab2e0b7a21e6e75da7f91a9-Abstract-Conference.html.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, 2024. doi: 10.18653/v1/2024.emnlp-main.300. URL https://aclanthology.org/2024.emnlp-main.300/.

A Additional derivations and results

The appendix develops the explanation and documents the supporting checks. The Gaussian derivation identifies when averaging can help; the experiment inventory and calibration studies show how we test that explanation. Appendix B gives the scoring conventions, denominator interventions, and detector comparisons, and Appendix C covers changes to wording. Supplement Tables S1 to S18 provide the full grids. Study names identify the model, corpus, update strength, and task.

A.1 Gaussian aggregation derivation

Averaging can reduce noise, but it can also dilute the difference between members and controls. The following Gaussian calculation makes that tradeoff explicit. It is an explanatory model, not an assumption required by the empirical AUROC comparisons.

Let $m \geq 1$ be the number of views and $\mathbf{S} \in \mathbb{R}^m$ their real-valued scores. The class label Y is 1 for a member and 0 for a control, and y denotes its value. Let $\boldsymbol{\mu}_y$ be the class mean vector and $\boldsymbol{\Sigma}$ the positive definite covariance shared by both classes: every nonzero weighted score has positive variance. With $\mathcal{N}$ denoting a multivariate normal distribution, suppose $\mathbf{S} \mid Y = y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$. The member-minus-control mean shift is $\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. We orient scores so that larger values indicate membership and assume independent member and control draws. To apply this model to data, our Gaussian plug-in substitutes parameter estimates from other fine-tuning seeds. Shrinkage stabilizes the empirical covariance estimate (Ledoit & Wolf, 2004). Let $\mathbf{S}_1, \mathbf{S}_0$ be independent member and control score vectors, and $\mathbf{w}$ a nonzero vector of aggregation weights. A superscript $\top$ denotes transpose. AUROC is the probability that a member score exceeds an independent control score, with half credit for ties. Gaussian scores here have no ties. The weighted score $A_{\mathbf{w}} = \mathbf{w}^\top \mathbf{S}$ satisfies $\mathbf{w}^\top(\mathbf{S}_1 - \mathbf{S}_0) \sim \mathcal{N}(\mathbf{w}^\top \boldsymbol{\Delta}, 2\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w})$, so, with Φ the standard normal cumulative distribution function,

$$\mathrm{AUROC}(A_{\mathbf{w}}) = \Phi\left(\mathbf{w}^\top \boldsymbol{\Delta} \big/ \sqrt{2\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}\right). \tag{1}$$

With $\mathbf{\Sigma}^{-1}$ denoting the inverse covariance and c a nonzero scaling constant, Cauchy–Schwarz gives $(\mathbf{w}^\top \mathbf{\Delta})^2/\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \leq \mathbf{\Delta}^\top \mathbf{\Sigma}^{-1} \mathbf{\Delta}$, with equality, for $\mathbf{\Delta} \neq 0$, exactly when $\mathbf{w} = c\mathbf{\Sigma}^{-1}\mathbf{\Delta}$, $c \neq 0$. The AUROC is maximized by $c > 0$ and minimized by $c < 0$, which reverses the ranking. This optimum allows negative weights and uses the class means and covariance; it is not a guarantee for an unweighted mean. If $\mathbf{\Delta} = 0$ every nonzero aggregate has AUROC $1/2$. Let Δ_1 be the first view's mean shift and Σ_{11} its within-class variance. Writing $\mathbf{1}$ for the all-ones vector, the orbit mean $\mathbf{w} = \mathbf{1}/m$ beats the first view exactly when

$$\mathbf{1}^\top \mathbf{\Delta} / \sqrt{\mathbf{1}^\top \mathbf{\Sigma} \mathbf{1}} > \Delta_1 / \sqrt{\Sigma_{11}}, \tag{2}$$

a signed comparison in which $1/m$ cancels. We may square it only when both sides are nonnegative. With $\mathbf{I}$ the identity matrix and covariance $\mathbf{\Sigma} = \mathbf{I}$ and $\mathbf{\Delta} = (1, -4)$, the squared form holds $(9/2 > 1)$ while the mean's AUROC is $\Phi(-3/2) = .067$ against $\Phi(1/\sqrt{2}) = .760$ for the first view. Even an uncorrelated view can lower AUROC when its class difference is weak or points in the opposite direction. We use the signed condition, record both shifts' signs, and test predictions on held-out runs. For $m > 1$, suppose every view has mean shift δ, within-class standard deviation $\sigma > 0$, and the same pairwise correlation ρ. Positive definiteness requires $-1/(m-1) < \rho < 1$. Averaging then divides the single-view variance by m_{eff}, the effective number of independent views. The orbit-mean AUROC is $\Phi\{\delta\sqrt{m_{\text{eff}}}/(\sqrt{2}\sigma)\}$ with $m_{\text{eff}} = m/[1 + (m-1)\rho]$. For $m = 1$, set $m_{\text{eff}} = 1$; no pairwise correlation is needed. The singular correlation endpoints are outside the assumptions above. With unequal shifts, effective view count alone no longer determines AUROC relative to the first view (Supplement Table S13). Let $\Pr$ denote probability and Cov covariance. With member prevalence $\pi = \Pr(Y = 1)$, the law of total covariance gives the pooled covariance: $\mathrm{Cov}(\mathbf{S}) = \mathbf{\Sigma} + \pi(1-\pi)\mathbf{\Delta}\mathbf{\Delta}^\top$. The first term describes variation within each class; the second comes from the class means. Thus pooled correlations mix redundancy with membership signal.

A.2 Experiment inventory and provenance

Table 8 names each synthetic configuration by model size, compute backend, batch size, and task. A seed fixes pseudorandom choices; the listed seed suffixes follow 202609. Exposure E/M denotes exact/mixed training with paired items and seeds and is distinct from scoring convention M. The natural-text studies cover Qwen and pythia on WikiMIA and arXiv under weak, strong, dose-matched, and paraphrase evaluations. The synthetic code/multiple-choice-question (MCQ) studies score correct answers, whereas the natural studies score passages. Released model documentation is linked for Qwen2.5, Pythia, and SmolLM2. A checkpoint is a saved parameter set; pinned identifiers fix its version and tokenizer. GPU means graphics processing unit. CUDA and Metal Performance Shaders (MPS) are the NVIDIA and Apple compute backends. SDPA means scaled dot-product attention, the implementation used for these runs. Table 4 gives the corresponding model and training abbreviations.

Q: Qwen2.5; I/B: instruction-tuned/base; MCQ: multiple-choice question; CUDA/MPS: NVIDIA/Apple backends; SDPA: scaled dot-product attention. In/out count members/controls; dose counts presentations per member; seed suffixes follow 202609. B/M in model sizes mean billion/million parameters.

Table 4: Inventory of known-inclusion experiments, models, tasks, item counts, views, seeds, doses, and compute stacks.

Study	Model / task	Members / controls	Views	Seed IDs	Dose /member	Stack; batch
Arithmetic control	Q .5/1.5B-I / arithmetic	60/60	5	05–07	24	CUDA; 16/16
0.5B CUDA	Q .5B-I / code, MCQ	45/45	6	11–13	8	CUDA 12.1; 8
1.5B CUDA	Q 1.5B-I / code, MCQ	45/45	6	21–23	4	CUDA 12.5; 16
0.5B MPS-b16	Q .5B-I / code, MCQ	45/45	6	11–20	8	MPS eager; 16
7B MPS-b16	Q 7B-I / code; MCQ	45/45	6	11–15;11–16	8	MPS eager; 16
0.5B MPS-b8	Q .5B-I / code, MCQ	45/45	6	11–13	8	MPS eager; 8
Qwen multi-pass	Q .5B-B / natural	300/300	3	27–29	2,4	MPS SDPA; 8
Qwen one-pass	Q .5B-B / natural	300/300	3	27–29	1	MPS SDPA; 8
pythia	pythia-1.4b-B / natural	300/300	3	27–29	1,2	MPS SDPA; 8
arXiv replication	Q .5B-B, pythia-1.4b-B / arXiv	300/300	3	16–18	1,2	MPS SDPA; 8
WikiMIA paraphrase	Qwen two-pass checkpoints	300/300	2	27–29	2	MPS SDPA; 8
Qwen added fits	Q .5B-B / natural	300/300	3	27–29	2; M_d 2	MPS SDPA; 8
pythia added fits	pythia-1.4b-B / natural	300/300	3	27–29	2; M_d 2	MPS SDPA; 8

We record missing historical revisions as unspecified. In Table 5, a *rerunnable* family has the pinned revision, prepared inputs, runner, and hashes required to repeat the run. A *replay-only* family retains scores for reanalysis but cannot reconstruct the exact run. Frozen protocols fix the primary contrast

Pres./member counts presentations per member; Pad is the padded token length; rate is the learning rate; M_d is dose-matched mixed training.

Table 5: For each experiment family, the table records whether runs are rerunnable or replay only and lists the pinned models, recipes and retained weights or inputs.

Study	Model / task	Model and tokenizer revision	Weights retained	Inputs	Status
Arithmetic control	Q.5/1.5B-I / arithmetic	not pinned	no	scores+manifests	replay only
0.5B CUDA	Q.5B-I / code, MCQ	not pinned	no	scores+manifests	replay only
1.5B CUDA	Q1.5B-I / code, MCQ	not pinned	no	scores+manifests	replay only
0.5B MPS-b16	Q.5B-I / code, MCQ	unspecified	no	scores+manifests	replay only
7B MPS-b16	Q7B-I / code, MCQ	unspecified	no	scores+manifests	replay only
0.5B MPS-b8	Q.5B-I / code, MCQ	`7ae5576`	no	scores+manifests+runner	rerunnable
Qwen multi-pass	Q.5B-B / natural	`060db64`	no	prepared texts+tokens	rerunnable
Qwen one-pass	Q.5B-B / natural	`060db64`	yes (18)	prepared texts+tokens	rerunnable
pythia	pythia-1.4b-B / natural	`fedc38a`	yes (18)	prepared texts+tokens	rerunnable
WikiMIA paraphrase	Qwen two-pass checkpoints	`060db64`	no (six)	paraphrase texts+tokens	rerunnable
temporal-label audit	6 untuned checkpoints	not pinned	n/a	per-token scores	replay only
Qwen added fits	Q.5B-B / natural	`060db64`	yes (9)	prepared texts+tokens	rerunnable
pythia added fits	pythia-1.4b-B / natural	`fedc38a`	yes (9)	prepared texts+tokens	rerunnable

Pinned study	Base model	Passes × rate	Steps	Pres./member	Pad	Weights / inputs kept
Qwen multi-pass	Qwen2.5-0.5B base	2 or 4 × 5e-5	76 / 152	2 or 4	128	no / yes
Qwen one-pass	Qwen2.5-0.5B base	1 × {5e-7,5e-6,5e-5}	38	1	128	yes / yes
pythia	pythia-1.4b base	{1,2} × {5e-6,5e-5}	38 / 76	1 or 2	144	yes / yes
Qwen added fits	Qwen2.5-0.5B base	2×5e-6; M_d 2×5e-5	76 / 226	2 / 6	128	yes / yes
pythia added fits	pythia-1.4b base	2×5e-6; M_d 2×5e-5	76 / 226	2 / 6	144	yes / yes

and the replication rule before the pythia and arXiv fits, and a later protocol fixes the added two-pass and dose-matched stage. We retain weights, loss curves and token observations for all 18 fits of that stage (Table 15). The 54 paraphrase scorings require no training and cover 36 earlier checkpoints and the 18 checkpoints of the added stage. Later fits are outside these tables. The 24 SPV-MIA target refits and one trained reference model per target ran on H100, with target and reference weights kept on the compute host and score rows retained, while the 18 prospective and 1.5B fits ran locally on MPS and retain rows, reference corpora and manifests but not weights. In Table 6, "labels" are known membership labels used to choose the score. Every AUROC uses labels for evaluation, and refits (new runs of an earlier training recipe) are never paired with retained checkpoints.

RC: retained-context adaptation. NB-2 and NB-3: neighbourhood scores using two or three alternative views. T5: text-infilling model. Base: the untouched pretrained model. LDA: linear discriminant analysis. An evaluation-selected envelope takes the best candidate on evaluation labels.

Table 6: For each method, the table lists the model access it needs, whether labels choose its score, its implementation and its role in comparisons.

Method	Model access	Labels	Implementation	Comparator use
LL, zlib, Min-k, Min-k++	Target	None	Standard formulas	Six-candidate selector
Reference gain	Target and base	None	Standard formula	Six-candidate selector
EZ-MIA U, M, F	Target and base	None	U released, M and F local	F in six-candidate selector
Tag&Tab RC	Target	None	Retained-context adaptation	Separate envelope
NB-2, NB-3	Target	None	Retained-view analogue	Exploratory comparison
SPV-MIA	Target, reference, T5	None	Adapted pipeline, refits	Own exact, refit envelope
Gaussian, LDA, selection	Score vectors	Calibration	Diagnostic models	Held-out calibration

A.3 Rendered-prompt validation

An independent prompt-only parser evaluates arithmetic/MCQ expressions and restricted Python abstract syntax trees (ASTs). Mutation tests change operations, operands, moduli, inputs, and choices while retaining stale answer metadata. The parser checks 17,640 reconstructed views across 32 authored runs and 6,300 item/stage specificity records, and renderer bytes match the source snapshot. Because authored rows omit submitted prompts, reconstruction cannot verify the bytes sent to the model. Natural runs do retain their texts. The stored specificity field subtracts the mean of the other $m - 1$ views, a positive rescaling of exact minus orbit that preserves AUROC for fixed m. We generate and verify answers before teacher-forced scoring, which supplies the observed preceding tokens when predicting each next token. Membership labels are used only for evaluation.

A.4 Arithmetic scoring-pipeline positive control

The arithmetic control uses strong exposure by design (each of 60 members repeated eight times for three epochs). Near-perfect separation verifies an inclusion-sensitive scoring pipeline and does not

estimate attainable audit performance. Table 7 retains the negative aggregation outcomes. It reports three fits per model size, 60 members and 60 controls per fit, five views, and 24 presentations per member. We compute seed/item-paired intervals from 2,000 draws.

A identifies the arithmetic control; AUC means AUROC, Plain means original input, and TPR5 is TPR at 5% FPR. $O - E$ compares averaged and original scores; CI denotes confidence interval.

Table 7: Plain and orbit performance in the arithmetic positive control.

A / model	Plain	Orbit	$O - E$ AUC [95% CI]	$O - E$ TPR5 [95% CI]
Qwen2.5-0.5B-Instruct	.9956	.9742	-.0214 [-.0667,.0017]	-.0722 [-.2667,.0222]
Qwen2.5-1.5B-Instruct	1.0000	.9933	-.0067 [-.0191,.0000]	-.0222 [-.1333,.0000]

A.5 Synthetic factorials and the dependence model

These exploratory studies ask whether less dependent views provide useful evidence, and they motivate the paraphrase study of Appendix E.5. For each seed, we fine-tune Qwen2.5-Instruct on 45 members, withhold 45 controls, and score six fixed code and multiple-choice (MCQ) views: five renderings and a composed view always held out from training (Appendix A.7). Crossing exact and mixed training with exact and orbit scoring gives four paired cells per seed. With Δ_E and Δ_M the $O - E$ under exact and mixed training, the interaction is $\Delta_M - \Delta_E$. Models span 0.5B, 1.5B and an exploratory 7B study (Table 4), and an arithmetic positive control uses strong exposure by design. Mean effects use 10,000-draw joint seed and item percentile intervals, which form a three-cluster bootstrap without calibrated coverage in three-seed studies.

Across ten factorials, the interaction is positive in all ten and resolved in three. It mostly reflects a lower exact-view AUROC after mixed exposure rather than a higher orbit AUROC (Tables 8 and 10). Our shared-covariance Gaussian model shows that lower dependence alone does not help when a view also loses signal. The majority baseline always predicts the more frequent gain sign in its calibration runs. Calibrated on other seeds or item halves, the Gaussian model's held-out sign accuracy (.607 and .596) trails the corresponding majority baselines (.750 and .720), and neither linear discriminant analysis (LDA), which learns a weighted score to separate the classes, nor calibrated view selection reliably outperforms the fixed mean (Tables 11 and 12). The model thus explains a possible benefit without providing a reliable view-selection rule. Table 8 reports mean factorial contrasts; the following analyses distinguish their magnitude from consistency across seeds. Sign counts es-

Δ_E and Δ_M are orbit-minus-exact AUROC after exact and mixed training. Their difference is the interaction. CI: confidence interval; $*$: excludes zero; †: sixth seed added after five; ‡: rerunnable from pinned inputs.

Table 8: Each synthetic factorial reports its orbit effects under exact and mixed training and their interaction, with joint seed/item intervals. Seed-only comparisons are in Table 10.

						AUROC effects: mean [joint 95% CI]		
						Interaction	Exact training	Mixed training
Configuration	Size	Stack	Batch	Task	Seeds	$\Delta_M - \Delta_E$	Δ_E	Δ_M
MPS-b16/code	0.5B	MPS	16	code	10	.0419 [.0072, .0780]*	−.029 [−.048, −.013]*	.013 [−.013, .037]
MPS-b16/MCQ	0.5B	MPS	16	MCQ	10	.0355 [−.0183, .0918]	−.020 [−.068, .028]	.016 [−.022, .053]
7B MPS/code	*7B*	*MPS*	*16*	*code*	*5*	.0232 [−.0188, .0717]	−.009 [−.026, .006]	.014 [−.029, .065]
7B MPS/MCQ	*7B*	*MPS*	*16*	*MCQ*†	*6*	.0319 [−.0486, .1065]	−.037 [−.081, .008]	−.005 [−.063, .051]
MPS-b8/code	0.5B	MPS	8‡	code	3	.0863 [.0407, .1271]*	−.040 [−.075, −.002]*	.047 [−.010, .103]
MPS-b8/MCQ	0.5B	MPS	8‡	MCQ	3	.0321 [−.0591, .1236]	−.016 [−.082, .048]	.016 [−.046, .076]
CUDA-b8/code	0.5B	CUDA	8	code	3	.0858 [.0425, .1297]*	−.027 [−.050, −.009]*	.059 [.014, .102]*
CUDA-b8/MCQ	0.5B	CUDA	8	MCQ	3	.0398 [−.0308, .1105]	−.007 [−.071, .055]	.033 [−.027, .092]
1.5B CUDA/code	1.5B	CUDA	16	code	3	.0142 [−.0158, .0413]	−.014 [−.039, .011]	.000 [−.023, .025]
1.5B CUDA/MCQ	1.5B	CUDA	16	MCQ	3	.0250 [−.0589, .1139]	−.021 [−.106, .063]	.004 [−.059, .070]

* joint seed/item interval excludes zero; † sixth seed added after five; ‡ rerunnable from pinned inputs (other rows replay scores); italic: exploratory 7B study. Size: Qwen2.5-Instruct; 1.5B rows halve presentations; 45+45 items per seed.

timate how often nonzero seed effects are positive, while mean intervals estimate their magnitude (Table 9). The 7B MPS study uses five code seeds and six MCQ seeds. Its sixth MCQ seed followed five unanimous signs. We retain all six and mark this outcome-dependent cohort as exploratory. In Supplement Table S10 and Table 10, rows name the orbit, interaction, and between-exposure effects directly. Supplement Table S11 gives plain-view and orbit AUROC under exact (E) and mixed (M) exposure, with batch size, optimizer steps per fit and presentations per member.

Optimizer budget. For code, the interaction in Table 8 is largest with batch 8 and 46 optimizer steps (.0863 on MPS and .0858 on CUDA) and smaller with batch 16 and 24 steps (.0419). This pattern is consistent with an optimizer-budget effect, but it does not recur consistently for MCQ

E/M identify exact/mixed training and Int. their interaction. Each n counts nonzero seed effects; p_E, p_M, p_I are their sign-test p-values. In backend labels, b8/b16 mean batch sizes 8/16.

Table 9: For each study, the table counts positive signs among nonzero training-seed effects and gives their p-values (7B MPS MCQ uses outcome-dependent sampling).

Configuration / task	E positive/n	p_E	M positive/n	p_M	Int. positive/n	p_I
0.5B MPS-b16/code	0/10	.0020	7/10	.3438	8/10	.1094
0.5B MPS-b16/mcq	5/10	1.0000	7/10	.3438	7/10	.3438
7B MPS-b16/code	1/5	.3750	3/5	1.0000	3/5	1.0000
7B MPS-b16/mcq	1/6	.2188	3/6	1.0000	5/6	.2188
0.5B MPS-b8/code	0/3	.2500	3/3	.2500	3/3	.2500
0.5B MPS-b8/mcq	1/3	1.0000	2/3	1.0000	2/3	1.0000
0.5B CUDA/code	0/3	.2500	3/3	.2500	3/3	.2500
0.5B CUDA/mcq	1/3	1.0000	3/3	.2500	3/3	.2500
1.5B CUDA/code	0/2	.5000	1/3	1.0000	2/3	1.0000
1.5B CUDA/mcq	1/3	1.0000	2/3	1.0000	2/3	1.0000

(.0321 on MPS and .0398 on CUDA with batch 8, against .0355 with batch 16). No contrast isolates hardware.

One analysis table, two uncertainty targets. We generate every synthetic contrast from one versioned analysis table, which records per-seed effects and two intervals for their mean. Training seeds are independent units, while item identities pair the four cells. Let n be the number of training seeds. The seed-only interval is a percentile bootstrap over n training seeds: we enumerate all n^n ordered resamples when $n \leq 6$ and otherwise use 20,000 Monte Carlo draws with empirical-quantile endpoints. With three seeds, each of the 27 resamples has mass $1/27 > 2.5\%$, so the interval is the range of the three seed effects under inverse cumulative distribution function (CDF, type 1) quantiles. The joint seed/item interval also resamples member/control identities within each selected seed. It uses 10,000 draws, with the ith synthetic cell drawn using seed $20260919 + 100(i-1)$, and is a three-cluster bootstrap in three-seed studies. Neither procedure has calibrated coverage with three seeds, so we do not choose between them by significance. Relative to endpoints displayed before this rule was fixed, 25 endpoints moved and 3 contrasts changed zero exclusion: the exact-view effect for 7B MPS with batch 16 on code, and the exact-view effect and interaction for 1.5B CUDA with batch 16 on code. No main-text claim rests on these. Table 10 lists both intervals for every contrast. Each study has 3 to 10 fits, 45+45 paired items and 10,000 joint draws.

Table 10: Each synthetic contrast is shown with both its seed-only and its joint seed/item 95% interval.

		95% interval				95% interval	
Effect	**Mean**	**seed-only**	**joint seed/item**	**Effect**	**Mean**	**seed-only**	**joint seed/item**
0.5B MPS, batch 16, code, 10 seeds				**0.5B MPS, batch 8, MCQ, 3 seeds**			
Exact orbit	−.0293	[−.0435, −.0163]	[−.0476, −.0127]*	Exact orbit	−.0161	[−.0346, .0168]	[−.0818, .0479]
Mixed orbit	.0125	[−.0087, .0312]	[−.0131, .0370]	Mixed orbit	.0160	[−.0128, .0494]	[−.0464, .0760]
Interaction	.0419	[.0115, .0721]	[.0072, .0780]*	Interaction	.0321	[−.0296, .0800]	[−.0591, .1236]
Plain exposure	−.0876	[−.1334, −.0423]	[−.1409, −.0335]*	Plain exposure	−.0469	[−.0889, .0030]	[−.1351, .0342]
Orbit exposure	−.0457	[−.1040, .0092]	[−.1077, .0161]	Orbit exposure	−.0148	[−.0844, .0489]	[−.1055, .0724]
0.5B MPS, batch 16, MCQ, 10 seeds				**0.5B CUDA, batch 8, code, 3 seeds**			
Exact orbit	−.0200	[−.0509, .0107]	[−.0684, .0279]	Exact orbit	−.0270	[−.0420, −.0178]	[−.0495, −.0089]*
Mixed orbit	.0155	[.0016, .0307]	[−.0218, .0533]	Mixed orbit	.0588	[.0415, .0765]	[.0145, .1019]*
Interaction	.0355	[.0063, .0653]	[−.0183, .0918]	Interaction	.0858	[.0760, .0978]	[.0425, .1297]*
Plain exposure	−.0550	[−.0885, −.0212]	[−.1075, −.0038]*	Plain exposure	−.0721	[−.1417, −.0049]	[−.1575, .0125]
Orbit exposure	−.0195	[−.0548, .0153]	[−.0715, .0307]	Orbit exposure	.0137	[−.0440, .0711]	[−.0537, .0807]
7B MPS, batch 16, code, 5 seeds				**0.5B CUDA, batch 8, MCQ, 3 seeds**			
Exact orbit	−.0092	[−.0191, −.0003]	[−.0258, .0062]	Exact orbit	−.0067	[−.0163, .0109]	[−.0708, .0553]
Mixed orbit	.0140	[−.0249, .0579]	[−.0294, .0651]	Mixed orbit	.0331	[.0277, .0410]	[−.0265, .0919]
Interaction	.0232	[−.0114, .0606]	[−.0188, .0717]	Interaction	.0398	[.0301, .0454]	[−.0308, .1105]
Plain exposure	−.1060	[−.1901, −.0133]	[−.1994, −.0069]*	Plain exposure	−.0499	[−.0835, −.0168]	[−.1210, .0258]
Orbit exposure	−.0828	[−.1541, .0052]	[−.1634, .0121]	Orbit exposure	−.0100	[−.0380, .0133]	[−.0752, .0538]
7B MPS, batch 16, MCQ, 6 seeds				**1.5B CUDA, batch 16, code, 3 seeds**			
Exact orbit	−.0370	[−.0520, −.0175]	[−.0807, .0082]	Exact orbit	−.0143	[−.0341, .0000]	[−.0387, .0107]
Mixed orbit	−.0051	[−.0467, .0269]	[−.0626, .0509]	Mixed orbit	−.0002	[−.0198, .0207]	[−.0234, .0253]
Interaction	.0319	[−.0290, .0774]	[−.0486, .1065]	Interaction	.0142	[−.0015, .0296]	[−.0158, .0413]
Plain exposure	−.0415	[−.0683, −.0145]	[−.0965, .0142]	Plain exposure	.0033	[−.0109, .0267]	[−.0449, .0517]
Orbit exposure	−.0096	[−.0673, .0374]	[−.0836, .0592]	Orbit exposure	.0174	[−.0123, .0410]	[−.0278, .0614]
0.5B MPS, batch 8, code, 3 seeds				**1.5B CUDA, batch 16, MCQ, 3 seeds**			
Exact orbit	−.0395	[−.0607, −.0015]	[−.0752, −.0023]*	Exact orbit	−.0207	[−.0751, .0202]	[−.1058, .0627]
Mixed orbit	.0467	[.0148, .0998]	[−.0102, .1030]	Mixed orbit	.0043	[−.0069, .0148]	[−.0591, .0696]
Interaction	.0863	[.0756, .1012]	[.0407, .1271]*	Interaction	.0250	[−.0272, .0800]	[−.0589, .1139]
Plain exposure	−.0634	[−.1773, .0128]	[−.1748, .0346]	Plain exposure	.0198	[−.0558, .0874]	[−.0788, .1131]
Orbit exposure	.0229	[−.1017, .0948]	[−.0907, .1228]	Orbit exposure	.0448	[.0005, .1096]	[−.0352, .1294]

* joint interval excludes zero. Exact/Mixed orbit: orbit minus plain view under the named exposure; Interaction: their difference; Plain/Orbit exposure: mixed-view minus exact-text training on the named score.

A.6 DIAGNOSTIC PREDICTION AND CALIBRATION ACCESS

These diagnostics ask whether measured dependence can predict an averaging gain on unseen data. The in-sample Gaussian uses the target run's labels, whereas its majority baseline uses other runs. Their in-sample comparison therefore gives the Gaussian model more information. For seed holdout, we estimate class means and residual covariance from the other two seeds in the same model/domain/exposure cell and exclude target source identities. The item-held-out version uses balanced random halves (200 splits per run). We average valid non-tie split accuracies within runs and bootstrap those run averages (29 run averages and 28 non-tie full-run targets). The seed- and item-held-out gaps to the majority sign are $-.143$ [$-.321$,.036] and $-.123$ [$-.226$,$-.035$]. Transfer across size or task gives .609 [.391,.783] and .478 [.261,.696], both including chance (Table 11), and configurations sharing a seed are dependent. Mean absolute prediction error is .0167 in sample and .0202 held out, against mean observed effect .0233. The errors are comparable to the effects being predicted, so Figure 3's in-sample agreement does not establish predictive performance. Item-held-out accuracy exceeds one half in 16/29 runs, with no accuracy exactly .5 (two-sided sign test $p = .71$). All 30 LDA folds lack canonical-task overlap, so purging leaves predictions unchanged, although shared-generator dependence remains. LDA standardizes scores using calibration-fold means and scales, uses the least-squares solver (`lsqr`), and applies automatic covariance shrinkage. View selection uses the same labels and chooses the highest mean AUROC on other seeds.

Table 11: Gaussian and majority-sign prediction accuracy by calibration access, with descriptive run-bootstrap 95% intervals from 20,000 draws.

Label access / target	Gaussian [95% CI]	Majority [95% CI]	Paired gap [95% CI]	Runs
Evaluation labels / in-sample	.7857 [.6429,.9286]	.7500 [.5714,.8929]	.0357 [-.1786,.2500]	28
Fit-half labels / item holdout	.5963 [.4895,.7016]	.7196 [.6252,.8092]	-.1233 [-.2260,-.0353]	29
Other seeds / seed holdout	.6071 [.4286,.7857]	.7500 [.5714,.8929]	-.1429 [-.3214,.0357]	28
Other size / transfer				
0.5B↔1.5B, same family	.6087 [.3913,.7826]	–	–	23
Other task domain (code↔MCQ) / transfer	.4783 [.2609,.6957]	–	–	23

Both axes are orbit-minus-original AUROC. Sign acc. is accuracy in predicting the difference's sign; MAE is mean absolute prediction error; $|\text{observed}|$ is mean absolute observed difference. MCQ denotes multiple-choice questions.

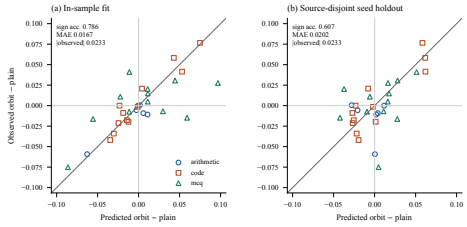


Figure 3: Gaussian AUROC-gain predictions and observed gains in sample (a) and under source-disjoint seed holdout (b).

Figure 3 shows identity and zero lines for 30 arithmetic and CUDA synthetic runs, of which 28 have nonzero signs. The held-out errors are comparable in magnitude to the effects.

A.7 FIXED-ORDER PREFIXES, VIEW GEOMETRY, SIMULATIONS, AND OPERATING POINTS

To measure what each additional view contributes, we add views in a fixed order: original prompt (plain), reordered operands (swap), a neutral operation (neutral), irrelevant text (distractor), layout change (format), then their combination (composed). We select no prefix retrospectively. Each added view costs one target evaluation and no reference call. Table 13 reports standardized class-mean shift (the member-minus-control mean difference relative to score variation), the added view's mean within-class correlation ρ with earlier views, and total cost, with no increment in the first row (all configurations in Supplement Table S1). The 0.5B CUDA mixed-exposure code analysis uses three seeds, 45 members and 45 controls per seed, and seed/item intervals from 2,000 draws.

The simulations separate three questions: what independent repetitions can add, how correlated views interact with unequal signals, and how difficulty differences can mimic membership evidence. Supplement Table S13 tests the correlated and heterogeneous shifts of Equation 2, and its Monte

LDA: linear discriminant analysis; Cal.: calibration; Sel.: selected view or exact detector, as headed. Headroom is one minus comparator AUROC. $O_{\mathrm{arith/rank}}$ averages scores/ranks. UB_{95}/UB_B denote pointwise/Bonferroni one-sided upper bounds; n/a means not applicable.

Table 12: Fixed orbit means versus label-calibrated aggregation and selection under seed or fit holdout.

Cell / exposure	Plain	Orbit	LDA	Sel. view	LDA − orbit [95% CI]	Selected − orbit [95% CI]	Cal. rows	Folds
0.5B CUDA/code/E	.8204	.7934	.8219	.8204	.0285 [.0025,.0489]	.0270 [.0178,.0420]	180	3
0.5B CUDA/code/M	.7483	.8071	.7969	.7867	-.0102 [-.0227,.0010]	-.0204 [-.0321,.0005]	180	3
0.5B CUDA/mcq/E	.6538	.6471	.6357	.6538	-.0114 [-.0331,.0015]	.0067 [-.0109,.0163]	180	3
0.5B CUDA/mcq/M	.6040	.6370	.6326	.6469	-.0044 [-.0148,.0040]	.0099 [-.0119,.0291]	180	3
1.5B CUDA/code/E	.6519	.6375	.6670	.6519	.0295 [-.0015,.0899]	.0143 [.0000,.0341]	180	3
1.5B CUDA/code/M	.6551	.6550	.6268	.6484	-.0281 [-.0509,-.0158]	-.0066 [-.0207,.0109]	180	3
1.5B CUDA/mcq/E	.5664	.5457	.5182	.5332	-.0275 [-.0652,.0153]	-.0125 [-.0202,-.0044]	180	3
1.5B CUDA/mcq/M	.5862	.5905	.6184	.5722	.0280 [-.0119,.0696]	-.0183 [-.0400,.0202]	180	3

Natural cell	Update	Selected	Headroom	W−sel. mean [UB_{95}/UB_B]	O_{arith}−sel. mean [UB_{95}/UB_B]	O_{rank}−sel. mean [UB_{95}/UB_B]
pythia/WikiMIA	weak	.713	.287	−.056 [−.026/.002]	−.003 [.005/.019]	−.006 [.008/.021]
pythia/WikiMIA	mid	.976	.024	−.057 [−.039/−.026]	−.014 [−.004/.004]	−.024 [−.012/−.003]
pythia/WikiMIA	strong	.998	.002	−.161 [−.135/−.112]	−.021 [−.009/−.002]	−.059 [−.042/−.029]
Qwen/WikiMIA	weak	.786	.214	−.039 [−.023/−.008]	−.002 [.006/.013]	−.007 [.003/.012]
Qwen/WikiMIA	mid	.984	.016	−.027 [−.014/−.006]	−.008 [.001/.007]	−.012 [−.002/.004]
Qwen/WikiMIA	strong	.995	.005	−.050 [−.036/−.026]	−.004 [.003/.010]	−.014 [−.006/.001]
pythia/arXiv	weak	.710	.290	−.069 [−.041/−.017]	−.007 [.009/.029]	−.015 [.004/.022]
pythia/arXiv	strong	1.000	.000	−.110 [−.090/−.073]	−.008 [−.001/.006]	−.035 [−.023/−.014]
Qwen/arXiv	weak	.762	.238	−.019 [.005/.027]	.002 [.021/.042]	−.001 [.020/.041]
Qwen/arXiv	strong	1.000	.000	−.045 [−.028/−.016]	−.006 [−.001/.000]	−.017 [−.009/−.004]

Natural-text extension: convention F; arithmetic orbit uses a 2× finite-score cap. Each contrast is mean [pointwise one-sided 95% upper bound / Bonferroni upper bound over the 30 comparisons, 30,000 draws], averaged within five source folds and three held-out fits. Headroom is one minus the selected exact AUROC, the largest gain any view could show. Upper bounds below margins .005: 23 (18), .010: 28 (21), .020: 29 (24) of all 30 comparisons, pointwise (Bonferroni), and .005: 11 (8), .010: 16 (9), .020: 17 (12) among the 18 with headroom of at least .01. Six exact-text candidates; the target fitted checkpoint and test-source scoring rows are excluded from selection within each model family, which is repeated within each bootstrap draw. Intervals condition on the observed fits and fixed score transforms.

Weak cell	Fits	W−sel. mean [UB_{95}/UB_B]	O_{arith}−sel. mean [UB_{95}/UB_B]	O_{rank}−sel. mean [UB_{95}/UB_B]
Qwen/WikiMIA	12 new	−.044 [−.038/−.033]	−.011 [−.009/−.007]	−.015 [−.012/−.010]
pythia/WikiMIA	12 new	−.051 [−.038/−.028]	−.002 [.001/.004]	−.007 [−.001/.004]
Qwen/arXiv	12 new	−.003 [.003/.008]	.005 [.010/.013]	.005 [.009/.013]
pythia/arXiv	12 new	−.045 [−.032/−.022]	−.002 [.005/.011]	−.008 [.000/.006]
Qwen/arXiv-96	3	−.020 [−.004/n/a]	−.003 [.007/n/a]	−.004 [.007/n/a]
pythia/arXiv-96	3	−.032 [−.006/n/a]	−.002 [.012/n/a]	−.006 [.009/n/a]

Fresh-fit extension (Appendix E.4): twelve fresh-assignment fits per weak cell, with fits as the unit, a one-sided 95% t bound and a Bonferroni bound over the 12 cell-by-contrast comparisons. The arXiv-96 rows use 96-word prefixes and three fits with conditional source/fit bounds, for which no Bonferroni bound applies (n/a).

Cell	Fits	Infilling exact	F exact	Min-k++ exact	Infilling $O-E$ [95% CI]	Infilling $O-W$ [95% CI]	O_{arith}−sel., 7 cand. mean [UB_{95}/UB_B]
Qwen/WikiMIA weak	12	.531	.796	.536	.001 [−.004, .007]	.000 [−.008, .008]	−.011 [−.009/−.007]
pythia/WikiMIA weak	12	.532	.720	.543	.000 [−.005, .006]	.002 [−.006, .010]	−.002 [.001/.004]
Qwen/arXiv weak	12	.513	.743	.539	.003 [−.001, .007]	−.003 [−.008, .002]	.005 [.010/.013]
pythia/arXiv weak	12	.519	.686	.535	−.001 [−.004, .002]	.004 [−.001, .008]	−.002 [.005/.011]
Qwen/WikiMIA strong	3	.967	.989	.988	−.041* [−.058, −.027]	.069* [.055, .083]	n/a
pythia/WikiMIA strong	3	.973	.986	.987	−.064* [−.086, −.043]	.091* [.072, .111]	n/a
Qwen/arXiv strong	3	.997	.996	.998	−.021* [−.034, −.011]	.042* [.029, .057]	n/a
pythia/arXiv strong	3	.997	.992	.998	−.024* [−.039, −.011]	.059* [.045, .075]	n/a
Qwen/arXiv-96 weak	3	.529	.751	.530	.003 [−.006, .011]	.000 [−.011, .011]	n/a
Qwen/arXiv-96 strong	3	.999	1.000	.999	−.003* [−.008, −.001]	.015* [.008, .024]	n/a
pythia/arXiv-96 weak	3	.532	.696	.541	.001 [−.009, .011]	.006 [−.008, .018]	n/a
pythia/arXiv-96 strong	3	.998	.999	1.000	−.002* [−.005, .000]	.019* [.010, .028]	n/a

Infilling Score extension (Appendix E.4): retrained replication fits scored with Infilling Score, implemented from the paper's equations with one future token for 32-word and five for 96-word inputs. Weak rows use twelve fits with fit-level intervals, and the last column adds Infilling as a seventh held-out candidate. Other rows use three fits with conditional source/fit intervals.

Table 13: Detection, dependence, and cost as views are added to the 0.5B CUDA mixed-exposure code cell.

Added view	Cost	AUC	TPR5	△AUC [95% CI]	Class-mean shift	Within-class ρ
plain	1	.7483	.2370	.0000 [.0000,.0000]	.969	–
swap	2	.7745	.1630	.0262 [-.0156,.0693]	.920	.515
neutral	3	.7965	.1556	.0221 [-.0115,.0543]	.896	.464
distractor	4	.7954	.1481	-.0012 [-.0196,.0186]	1.010	.586
format	5	.8033	.2296	.0079 [-.0217,.0385]	.794	.293
composed	6	.8071	.2222	.0038 [-.0082,.0171]	1.146	.621

Carlo error is distinct from training-seed uncertainty. In the tested aligned-shift settings the mean helps on average, by .0980 at zero correlation but only .0035 at correlation .9, while with heterogeneous shifts it loses $-.0308$ at correlation .9 and wins in a fraction .028 of replicates. With opposed shifts it loses everywhere. In the difficulty-confounding simulation, the rejection rate of Welch's two-sample mean test rises from 5.46% at zero shift to 37.32% at shift .50. The mean really differs between groups, so rejecting equal means is not a false rejection of that null. The misleading step would be to attribute the difference to membership. At 1%, 5% and 10% nominal FPR, 45 controls permit at most 0, 2 and 4 false positives, 60 controls 0, 3 and 6, and 300 natural controls 3, 15 and 30 (ties can lower the count). The raw-row audit (Supplement, Table S2) prints each authored seed's counts. Without an independent threshold-calibration split, every such column is an evaluation-ROC operating point rather than a calibrated deployment rate. For the temporal-label audit, propensities are predicted date-label probabilities from text embeddings, estimated without each evaluated source. We use them to reweight the score comparison. In Supplement Tables S14 to S16 the unit is a text pair and the denominator is 389 temporal controls, so their TPRs are evaluation-ROC points for date labels, not membership TPRs. The 389 controls permit at most 3, 19 and 38 false positives at 1, 5 and 10% FPR (fewer with ties), and a parenthesised rank counts the items at or above the threshold.

Table 14: The table reports evaluation-ROC TPR at 5% FPR for exact and orbit scores in synthetic cells and for F and likelihood in natural cells, with seed/item-paired intervals.

Model/domain	Exposure	Exact	Orbit	Orbit–exact 95% CI
0.5B/code	exact	.348	.274	$-.074$ [$-.163$, .067]
0.5B/code	mixed	.237	.222	$-.015$ [$-.170$, .230]
0.5B/MCQ	exact	.193	.067	$-.126$ [$-.259$, .096]
0.5B/MCQ	mixed	.074	.119	.044 [$-.081$, .141]
1.5B/code	exact	.126	.119	$-.007$ [$-.059$, .067]
1.5B/code	mixed	.148	.156	.007 [$-.074$, .052]
1.5B/MCQ	exact	.030	.081	.052 [$-.052$, .126]
1.5B/MCQ	mixed	.059	.096	.037 [$-.104$, .111]

Natural exact training	Update	F exact	LL exact	Arithmetic $O-E$ [95% CI]	Rank $O-E$ [95% CI]
pythia/WikiMIA	weak	.183	.054	$-.009$ [$-.042$, .049]	$-.012$ [$-.070$, .064]
pythia/WikiMIA	strong	.954	.998	$-.009$ [$-.019$, .005]	$-.150$ [$-.208$, $-.101$]
Qwen/WikiMIA	weak	.311	.049	.028 [$-.046$, .073]	.014 [$-.071$, .062]
Qwen/WikiMIA	strong	.974	.987	$-.011$ [$-.025$, .000]	$-.049$ [$-.086$, $-.025$]
pythia/arXiv	weak	.202	.090	$-.009$ [$-.041$, .060]	$-.007$ [$-.058$, .068]
pythia/arXiv	strong	.990	.999	$-.003$ [$-.011$, .000]	$-.100$ [$-.138$, $-.066$]
Qwen/arXiv	weak	.278	.092	.007 [$-.038$, .045]	.000 [$-.058$, .053]
Qwen/arXiv	strong	.991	1.000	$-.004$ [$-.012$, .000]	$-.044$ [$-.067$, $-.024$]

Natural-text extension: TPR at evaluation FPR $\leq 5\%$; 300 controls per fit permit at most 15 false positives, with whole ties. Three-fit/source conditional intervals; thresholds recomputed in each draw. Arithmetic F uses a $2\times$ cap. Counts and the 1% endpoint are retained in the artifact.

B Natural intervention grids, realized dose, and support decomposition

The Qwen WikiMIA strong-update study has six trajectories and 12 checkpoints. The Qwen weak-update and pythia WikiMIA studies have 18 fits each, and the arXiv replication has 24 (12 per model). All 764 sources fit every view without truncation. The strong-update study uses 76/152 steps for two/four passes, and other fits use 38 per pass. Added fits per model are three exact and three mixed fits at two passes of 5×10^{-6} (76 steps), three dose-matched mixed fits at 5×10^{-5} (226 steps, 1,800 presentations), and 54 paraphrase scorings across both models (Appendix A.2). Supplement Tables S4 to S7 give full grids, and S3 gives per-row support decompositions.

Bootstrap and pooling. Each draw resamples shared sources within strata, using the same multiplicities across fits and scores. Fit weights follow multinomial$(3, 1/3)/3$: three draws give equal selection probability to each fit, and counts are divided by three. We average source-weighted metrics over fits and report percentile intervals (2.5/97.5 and 5/95, seed 20260919). Imputation constants, ranks and comparator choices remain fixed across draws, so intervals condition on this preprocessing.

For the two-sided bootstrap tail fraction p, let $P(\cdot)$ here denote the empirical fraction of resampled contrasts satisfying its argument; it is distinct from EZ-MIA's positive gain P. Holm uses $p = \min\{1, 2\min[P(\text{draw} \leq 0), P(\text{draw} \geq 0)]\}$ over the 2,000 draws, with j indexing the ordered tests

and step j tested at $.05/(9-j)$. A stored zero means no crossing among the 2,000 draws, which reflects the Monte Carlo resolution rather than a calibrated tail bound, and seven primary contrasts have no crossing. The eighth (Qwen WikiMIA strong, exact-text) has $p = .0070$, tested last at .05.

Fit-level tests check the ranking without relying on these p-values. Across the twelve fresh fits of each weak cell, EZ(F) minus likelihood is .162 to .267 AUROC, with one-sided 95% lower bounds of .151 to .245 and every one-sided t test at $p < 10^{-9}$. Across six strong fits per setting (three original and three fresh), it is $-.005$ to $-.015$, and every 95% t interval lies below zero. These fits come from frozen protocols, so the reversal holds under the published rule at the fit level, although the strong-update deficit is small because both scores approach 1.

Realized dose and the detector ranking. Table 15 reports, per configuration, optimizer steps, active tokens (padding excluded), the loss change (mean over the last five steps minus the first five, negative when the loss falls) and the relative weight-update norm. With fitted and base parameter vectors θ_{fit} and θ_{base}, and $\|\cdot\|_2$ the Euclidean norm, this is $\|\theta_{\text{fit}} - \theta_{\text{base}}\|_2 / \|\theta_{\text{base}}\|_2$. We compare with the bfloat16 (16-bit floating-point) cast of the base, accumulating over all saved tensors in float32. We compute these norms post hoc from retained Qwen weak-update and pythia WikiMIA checkpoints, and at save time for arXiv. Using the float16 pythia-1.4b release instead would add a rounding floor of 1.72×10^{-3}. At 38 steps, the Qwen WikiMIA one-pass study's positive loss changes mostly reflect batch composition, even at 5×10^{-7} when weights barely move. The 12 strong-update Qwen fits lack retained weights and hence update norms. In Table 15, Exp. abbreviates exposure and Act. tokens excludes padding. The Qwen one-pass study at $5{\times}10^{-5}$ (8.19×10^{-3}) is the retained Qwen reference for a mid one-pass update. Table 2 orders the corresponding detector contrasts by realized update: the sign follows update strength within each model, but the magnitude need not. We order Qwen's weight-missing two-pass cell by its shared rate, doubled steps and lower loss, and the ranking switches after the one-pass norm 8.19×10^{-3}. Pythia's ranking switches between 6.64 and 8.56×10^{-3}. At 5×10^{-7} all scores are near chance (Table S5). We exclude this setting from the .01 comparison: mixed-training width sits .012 above it (EZ-MIA at .506, 90% interval [−.018,.042]), the only point estimate above $+.01$ across views, detectors and settings under M in the retained lift audit.

Table 15: The table reports the realized training dose of each configuration: steps, active training tokens, the first and last step losses, the change between the means of the first and last five steps, and the relative weight update (three-fit mean and range). Qwen/WikiMIA multi-pass retains step, token and loss records but no weights, so its norms are unavailable and shown as a dash. Definitions are in the text.

Study	Base	Passes×rate	Exp.	Fits	Steps	Act. tokens	First loss	Last loss	Δ, 5-step means	$\frac{\|\Delta\theta\|}{\|\theta\|} \cdot 10^3$	Range
Qwen/WikiMIA, one-pass	Qwen2.5-0.5B	1×5e-7	E	3	38	14400	2.698	2.536	.112	0.010	[0.01,0.01]
Qwen/WikiMIA, one-pass	Qwen2.5-0.5B	1×5e-7	M	3	38	14616	2.763	2.604	.114	0.010	[0.01,0.01]
Qwen/WikiMIA, one-pass	Qwen2.5-0.5B	1×5e-6	E	3	38	14400	2.698	2.433	.025	0.303	[0.30,0.31]
Qwen/WikiMIA, one-pass	Qwen2.5-0.5B	1×5e-6	M	3	38	14616	2.763	2.494	.018	0.310	[0.31,0.31]
Qwen/WikiMIA, one-pass	Qwen2.5-0.5B	1×5e-5	E	3	38	14400	2.698	2.428	.021	8.188	[8.05,8.41]
Qwen/WikiMIA, one-pass	Qwen2.5-0.5B	1×5e-5	M	3	38	14616	2.763	2.475	.005	8.260	[8.11,8.46]
Qwen/WikiMIA, added two-pass	Qwen2.5-0.5B	2×5e-6	E	3	76	28800	2.698	2.169	-.184	0.492	[0.49,0.50]
Qwen/WikiMIA, added two-pass	Qwen2.5-0.5B	2×5e-6	M	3	76	29232	2.763	2.257	-.197	0.502	[0.50,0.51]
Qwen/WikiMIA, added two-pass	Qwen2.5-0.5B	2×5e-5	M_d	3	226	87687	2.677	.297	-2.393	15.535	[15.15,15.89]
Qwen/WikiMIA multi-pass	Qwen2.5-0.5B	2×5e-5	E	3	76	28800	2.698	1.186	-1.257	–	–
Qwen/WikiMIA multi-pass	Qwen2.5-0.5B	2×5e-5	M	3	76	29232	2.763	1.298	-1.216	–	–
Qwen/WikiMIA multi-pass	Qwen2.5-0.5B	4×5e-5	E	3	152	57600	2.698	.439	-2.131	–	–
Qwen/WikiMIA multi-pass	Qwen2.5-0.5B	4×5e-5	M	3	152	58460	2.763	.497	-2.106	–	–
pythia/WikiMIA	pythia-1.4b	1×5e-6	E	3	38	12931	2.808	2.405	-.066	0.298	[0.30,0.30]
pythia/WikiMIA	pythia-1.4b	1×5e-6	M	3	38	13275	2.878	2.451	-.091	0.307	[0.30,0.31]
pythia/WikiMIA	pythia-1.4b	1×5e-5	E	3	38	12931	2.808	2.505	-.100	6.645	[6.54,6.76]
pythia/WikiMIA	pythia-1.4b	1×5e-5	M	3	38	13275	2.878	2.477	-.167	6.692	[6.50,6.83]
pythia/WikiMIA	pythia-1.4b	2×5e-6	E	3	76	25861	2.808	2.273	-.230	0.455	[0.45,0.46]
pythia/WikiMIA	pythia-1.4b	2×5e-6	M	3	76	26536	2.878	2.357	-.243	0.471	[0.47,0.47]
pythia/WikiMIA	pythia-1.4b	2×5e-5	E	3	76	25861	2.808	1.120	-1.499	8.555	[8.36,8.70]
pythia/WikiMIA	pythia-1.4b	2×5e-5	M	3	76	26536	2.878	1.248	-1.447	8.626	[8.49,8.77]
pythia/WikiMIA	pythia-1.4b	2×5e-5	M_d	3	226	79596	2.836	.304	-2.585	11.369	[11.21,11.65]

Z: incorrectly predicted tokens; P, N: total gain/loss there. For each view and fit, a, b are the maximum/median positive released scores pooled without labels. First means tied above all finite ratios; U/M/E1/F are the released, median-imputed, maximum-imputed and published rules.

Table 16: Scores assigned by each EZ-MIA convention across error-zone cases.

Case	U	M	E1	F
$\|Z\| = 0$	0	b	a	First
$\|Z\| = 1$, improved	0	b	a	First
$\|Z\| = 1$, worsened	0	b	a	0
$\|Z\| \geq 2, P > 0, N = 0$	0	a	a	First
$\|Z\| \geq 2, N > 0$	P/N	P/N	P/N	P/N
Nonempty $P = N = 0$	0	b or a	a	Unspecified*

Here, a and b are the maximum and median positive released scores within one view and fit, pooled over all items without labels, and "First" ties flagged items above all finite ratios. *There is no retained case, and our F implementation returns zero. Appendices B and E.1 give the fallbacks, arithmetic cap and timing.

continued

Study	Base	Passes×rate	Exp.	Fits	Steps	Act. tokens	First loss	Last loss	△, 5-step means	$\frac{\|\triangle\theta\|}{\|\theta\|} \cdot 10^3$	Range
arXiv replication	pythia-1.4b	1×5e-6	E	3	38	12669	3.472	3.413	-.025	0.307	[0.31,0.31]
arXiv replication	pythia-1.4b	1×5e-6	M	3	38	13022	3.563	3.445	-.050	0.319	[0.32,0.32]
arXiv replication	pythia-1.4b	2×5e-5	E	3	76	25338	3.472	1.672	-1.935	9.891	[9.89,9.90]
arXiv replication	pythia-1.4b	2×5e-5	M	3	76	26056	3.563	1.768	-1.878	9.816	[9.79,9.85]
arXiv replication	Qwen2.5-0.5B	1×5e-6	E	3	38	12443	3.655	3.443	-.022	0.316	[0.32,0.32]
arXiv replication	Qwen2.5-0.5B	1×5e-6	M	3	38	12667	3.749	3.500	-.042	0.324	[0.32,0.33]
arXiv replication	Qwen2.5-0.5B	2×5e-5	E	3	76	24886	3.655	1.786	-1.828	12.986	[12.95,13.06]
arXiv replication	Qwen2.5-0.5B	2×5e-5	M	3	76	25351	3.749	1.862	-1.783	13.118	[13.07,13.17]

Scoring conventions. Table 16 collects the edge cases introduced in Section 2. U reproduces the released code, whereas F implements the published edge-case recommendations. For positive N, U adds 10^{-12} to the denominator for numerical stability; we suppress this offset in the ratio notation and retain it in code-parity checks. M and E1 isolate the role of short error zones. The paragraphs below explain the finite substitutes used when averaging scores ranked first and the timing of each rule.

Zero-score causes, parity, and conventions. The released rule assigns zero in three observed disjoint cases: a short error zone ($|Z| < 2$), at least two errors with zero negative mass and positive gain, or zero positive mass with some worsening (a genuine zero ratio). Zero negative mass means no error token worsened, allowing unchanged tokens. Across the strong natural settings, 3,683 exact-view member scores are zero: 409 (11%) are short zones and 3,262 have zero negative mass. Our scalar reproduces the pinned upstream scoring code (commit `aa0b6b7`, SHA-256 `f86f9587...`) on 129,600 item-view scores from 72 checkpoints: maximum relative difference 4.3×10^{-7}, with all 10,863 zeros reproduced.

Conventions M/E1/F follow Table 16. M uses the median b for fewer than two errors and the maximum a otherwise, and M and E1 leave zeros unchanged when a view and fit has no positive released score. Because a and b pool all items of a view and fit, the imputations depend on the evaluation cohort. F arithmetic means replace "First" by twice the largest finite ratio in that view and fit (ten times in the sensitivity), rank means use average ranks, and all preprocessing stays fixed across bootstrap draws. No retained nonempty zone has all changes exactly zero, so F can leave that unobserved case unspecified. D restricts to items positive under both exact and orbit scores. U, M and the common-support subset D were specified before their corresponding computations. E1 was added with the additional fits and F after those contrasts had been seen. The convention replay reproduces the earlier U/M/E1 values before adding F.

The one-error distinction matters: of the 7,200 exact-view observations in the four strong primary cells, 198 have one error and 19 of these worsened, all members. Thus F and E1 differ beyond top-score ties. Tables 17 and 18 retain all contrasts and support counts. Their second row per cell gives the number n_D of common-support items and mean member error-zone length. D retains a resolved reversal in 5 of 8 cells. At the prespecified weak update M, E1 and F coincide. Within the primary family, Holm retains 5 of eight contrasts under E1 and 5 under F. Supplementary grids retain per-fit AUROCs and zero decompositions.

EZ−LL: EZ-MIA minus likelihood AUROC. In the Train column E/M mean exact/mixed training, not a view or scoring rule; M_d means dose-matched mixed training. $O - E$ and $O - W$ compare the orbit with original and width views. D restricts to items positive under both scores.

Table 17: Sensitivity of the main contrasts to scoring conventions and aggregation rules.

		EZ−LL on the exact text unless stated (mean, 95% CI) by scoring convention			
		Released	Imputed	Short-zone-first	Published
Train	Update	U	M	E1	F
Qwen2.5-0.5B, WikiMIA					
E	weak	.271 [.207, .329]*	.274 [.215, .330]*	.274 [.215, .330]*	.274 [.215, .330]*
	strong	−.320 [−.365, −.278]*	−.009 [−.017, −.002]*	−.004 [−.012, .002]	−.005 [−.013, .001]
M	weak	.270 [.210, .324]'	.274 [.213, .326]'	.274 [.213, .326]'	.274 [.213, .326]*
	strong	−.256 [−.300, −.210]'	−.009 [−.017, −.002]'	−.006 [−.014, .000]	−.007 [−.015, −.001]*
M_d	strong, EZ−LL	−.709 [−.772, −.638]*	−.216 [−.256, −.180]*	−.032 [−.055, −.014]*	−.141 [−.191, −.093]*
	$O - W$, arithmetic	.114 [.078, .156]*	.040 [.019, .064]*	.004 [−.000, .013]	.017 [.006, .032]*
	$O - W$, ranks	−.033 [−.076, .004]	−.012 [−.040, .017]	−.006 [−.016, .003]	−.017 [−.038, .003]
	$O - E$, arithmetic	.093 [.058, .134]'	.100 [.076, .126]'	.031 [.014, .051]'	.138 [.091, .187]*
	$O - E$, ranks	−.054 [−.104, −.014]*	.047 [.025, .070]*	.020 [.006, .036]*	.104 [.064, .145]*
Qwen2.5-0.5B, arXiv					
E	weak	.213 [.141, .281]*	.219 [.151, .286]*	.219 [.151, .286]*	.219 [.151, .286]*
	strong	−.676 [−.721, −.628]*	−.027 [−.038, −.017]*	−.001 [−.006, .000]	−.007 [−.016, −.000]*
M	weak	.204 [.133, .273]'	.212 [.143, .278]'	.212 [.143, .278]'	.212 [.143, .278]*
	strong	−.503 [−.561, −.442]'	−.018 [−.032, −.008]'	−.008 [−.020, −.000]'	−.018 [−.038, −.004]*
pythia-1.4b, WikiMIA					
E	weak	.180 [.132, .226]*	.186 [.142, .229]*	.186 [.142, .229]*	.186 [.142, .229]*
	strong	−.631 [−.675, −.583]*	−.039 [−.053, −.027]*	−.010 [−.019, −.003]*	−.018 [−.034, −.007]*
M	weak	.191 [.146, .234]'	.197 [.153, .240]'	.197 [.153, .240]'	.197 [.153, .240]*
	strong	−.495 [−.546, −.446]'	−.020 [−.031, −.012]'	−.009 [−.020, −.002]'	−.011 [−.021, −.003]*
M_d	strong, EZ−LL	−.827 [−.870, −.780]*	−.292 [−.333, −.253]*	−.018 [−.030, −.007]*	−.116 [−.150, −.085]*
	$O - W$, arithmetic	.169 [.126, .212]*	.064 [.035, .093]*	.006 [.000, .015]*	.028 [.015, .044]*
	$O - W$, ranks	−.021 [−.059, .012]	−.017 [−.053, .020]	−.001 [−.008, .008]	−.012 [−.031, .008]
	$O - E$, arithmetic	.128 [.077, .178]'	.114 [.082, .144]'	.016 [.006, .028]'	.108 [.079, .140]*
	$O - E$, ranks	−.062 [−.108, −.026]*	.033 [.001, .063]*	.010 [.002, .019]*	.068 [.040, .103]*
pythia-1.4b, arXiv					
E	weak	.164 [.110, .218]*	.172 [.117, .226]*	.172 [.117, .226]*	.172 [.117, .226]*
	strong	−.690 [−.747, −.634]*	−.041 [−.056, −.029]*	−.002 [−.006, .002]	−.007 [−.019, .000]
M	weak	.175 [.115, .234]'	.175 [.115, .234]'	.175 [.115, .234]'	.175 [.115, .234]*
	strong	−.551 [−.610, −.495]'	−.014 [−.023, −.008]'	−.004 [−.010, −.000]'	−.005 [−.012, −.001]*

* interval excludes zero. U: the pinned released implementation (zero for fewer than two error tokens or zero negative mass); M and E1: this paper's imputations (Section 2); F: the published edge cases implemented separately (error-free sequences and zero-denominator ratios ranked first, one-error sequences by their ratio). Train E/M: exact-text/mixed-view; M_d: mixed-view with the exact-text dose held equal, with the averaging increment $O - W$ and the orbit gain $O - E$ under arithmetic averaging (F: ranked-first constant at twice the largest finite ratio) and under the mean of within-view ranks; weak: one pass at 5×10^{-6}; strong: two passes at 5×10^{-5} (Qwen2.5-0.5B on WikiMIA: the two-pass checkpoint); the positive-support subset (D): Table 18.

Aggregation of ranked-first scores. Arithmetic means require a finite stand-in for infinity (Table 17). Setting this constant to twice or ten times the largest finite ratio leaves every E1 and F value of $O - W$ in all 30 cells unchanged at three decimals, and the largest change is 7×10^{-5}. The largest finite ratios already dominate the mean. We report a within-view rank mean, with ties averaged, beside arithmetic aggregation. Ranks preserve single-view AUROC but rescale both finite ratios and imputed constants, so this comparison cannot isolate imputation alone. The largest change from arithmetic to rank aggregation of $O - W$ over all 30 cells is .081 under M and .047 under E1. Under F with rank aggregation the four strong exact-text cells give $O - E$ −.040 to −.010 and the four mixed-view strong cells $O - W$ −.002 to .007 (largest 95% upper bound .015). Table 18 uses ranks for F and includes four-pass sensitivities. Table 17 compares exact-text EZ−LL with the dose-matched $O - W$ and $O - E$ contrasts. U denotes the released rule, M and E1 the imputed rules, and F the published edge cases. We report means and 95% paired-draw intervals using seed 20260919.

Source-fold detector selection. We form five folds by cycling source IDs sorted within strata by the SHA-256 hash of a fixed salt and ID (Table 12). For each target fit and fold, the other two fits select the exact detector with the highest mean AUROC among F, likelihood, reference gain, Min-k++, zlib, and Min-k%. Selection excludes every test-fold source ID. Thus, all 150 target/fold combinations exclude both the target sources and target fit from calibration. We average AUROCs over five folds and then three fits without pooling scores from different selected detectors. Each of 2,000 source-resampling draws reselects the comparator on the two observed calibration fits. The draw resamples evaluation fits while holding calibration identities, folds, and score transformations fixed. These 30 comparisons lie outside the original 20-row family. Their specification was frozen after earlier results were seen, so this reanalysis adds no training replication.

Earlier evaluation-selected comparison. The evaluation-selected envelope is the highest exact-text AUROC among the candidate detectors, chosen using evaluation labels. The earlier M analysis against this envelope gives the upper bounds no greater than the reference margin across 20 comparisons. All upper bounds are at most $+.01$, but only 2 meet a 90% equivalence criterion, 12 are inferior beyond $-.01$, and weak/mid differences span $-.014$ to .000 (Appendix B).

Comparisons against the retrospective margin. The 20 prespecified rows are the width view and the orbit of EZ-MIA under M after exact-text training at the weak, mid and strong WikiMIA cells and the weak and strong arXiv cells of both models. The comparator is selected among EZ-MIA (M), likelihood, Min-k++ and reference gain. We call a comparison below the margin when the 95th percentile of view-minus-best-exact AUROC is at most $+.01$. This criterion holds for 20 of 20 rows and for 20 rows at the 97.5th percentile. A 90% interval entirely inside $\pm.01$ defines equivalence (2 rows: Qwen2.5-0.5B WikiMIA weak orbit; Qwen2.5-0.5B arXiv weak orbit), while a 95th percentile below $-.01$ defines inferiority (12 rows). The 6 of 48 all-cell rows that fail the upper bound are Qwen2.5-0.5B WikiMIA 1×5e-07 exact-text width; Qwen2.5-0.5B WikiMIA 1×5e-07 exact-text orbit; Qwen2.5-0.5B WikiMIA 1×5e-07 mixed-text width; Qwen2.5-0.5B WikiMIA 1×5e-07 mixed-text orbit; Qwen2.5-0.5B WikiMIA 1×5e-06 mixed-text orbit; pythia-1.4b arXiv 1×5e-06 mixed-text orbit. None was prespecified, and all arise at near-chance or weak updates.

Across all four detectors, 174 of 192 exploratory rows fall below the margin. Among them, 63 are equivalent and 86 inferior. The remaining 18 rows fail the bound and remain undetermined, with the last involving near-chance comparators. Every original EZ-MIA row in Table 19 has a 95% upper bound below $+.01$. Under M, all resolved within-detector lifts follow mixed-view training at the strong update and, for the orbit, the mid update. Width adds .008 on Qwen2.5-0.5B WikiMIA, .013 on Qwen2.5-0.5B arXiv, and .011 on pythia-1.4b arXiv. The orbit adds .008 to .015 in all four strong cells and .005 on both models at the mid update (Table 2). Each remains below the best exact-text detector in its cell, and none follows exact-text training.

Combining view-level significance tests. Combining view-level evidence through Fisher's method would require calibrated view-level p-values, which these membership scores do not provide.

Earlier held-out comparator checks. Replacing evaluation-label selection with selection on the other two fits, among the same four candidates, agrees in 26 of 30 prespecified folds, and the disagreements concern strong cells with leading scores within .002. This preserves all 20 prespecified verdicts and moves the 95th percentiles by at most .004. Purging the held-out fit's sources from the selection rows (at least 124 items per fold) picks the in-sample detector in 18 folds and costs one verdict, the pythia-1.4b WikiMIA weak orbit (.016, 95th percentile .045), after one fold selected reference gain (.657 against EZ-MIA's .712 in sample). Under F, with F in the comparator set, 20 of 20 prespecified rows meet the .01 bound (largest 95th percentile .007). In the full grid, D is the common positive-score subset and n_D its size. Error-zone length counts incorrect tokens. E/M in training labels mean exact/mixed exposure.

Table 18: For every natural and added cell, the table reports exact-text EZ−LL and $O - W$ under M, E1, F (rank orbit) and the positive-support subset D. Mean [95% CI], paired draw seed 20260919, with conventions in Table 17.

		Mean, 95% CI, by convention			
Setting, train	Contrast	Imputed M	Short-zone-first E1	Published F rank orbit	subset (D)
Qwen/WikiMIA multi-pass: Qwen2.5-0.5B, WikiMIA					
2 pass, E	EZ−LL	−.009 [−.017, −.002]*	−.004 [−.012, .002]	−.005 [−.013, .001]	−.002 [−.009, .005]
n_D 500/505/506; tok 7.7	$O - W$	.039 [.027, .053]*	.046 [.032, .061]*	.035 [.024, .048]*	.033 [.022, .046]*
2 pass, M	EZ−LL	−.009 [−.017, −.002]'	−.006 [−.014, .000]	−.007 [−.015, −.001]'	−.009 [−.020, −.001]*
n_D 520/532/523	$O - W$	.001 [−.001, .002]	.001 [−.001, .003]	−.002 [−.007, .001]	.089 [.060, .123]*
4 pass, E	EZ−LL	−.181 [−.210, −.153]*	−.025 [−.038, −.014]*	−.077 [−.102, −.055]*	−.043 [−.071, −.022]*
n_D 411/393/394; tok 2.1	$O - W$	.104 [.080, .129]*	.233 [.196, .271]*	.130 [.103, .160]*	.066 [.040, .095]*
4 pass, M	EZ−LL	−.130 [−.156, −.104]'	−.052 [−.070, −.035]'	−.130 [−.158, −.102]'	−.037 [−.062, −.017]*
n_D 445/457/466	$O - W$	.023 [.013, .034]'	.004 [.000, .009]'	−.011 [−.027, .005]	.273 [.198, .349]*
Qwen/WikiMIA, one-pass: Qwen2.5-0.5B, WikiMIA					
1x5e-07, E	EZ−LL	.035 [−.031, .097]	.035 [−.031, .097]	.035 [−.031, .097]	.035 [−.031, .097]
n_D 600/600/600	$O - W$	.014 [−.014, .041]	.014 [−.014, .041]	.013 [−.014, .040]	.014 [−.014, .041]
1x5e-07, M	EZ−LL	.019 [−.056, .090]	.019 [−.056, .090]	.019 [−.056, .090]	.019 [−.056, .090]
n_D 600/600/600	$O - W$	−.011 [−.038, .015]	−.011 [−.038, .015]	−.006 [−.032, .019]	−.011 [−.038, .015]
1x5e-06, E	EZ−LL	.274 [.215, .330]*	.274 [.215, .330]*	.274 [.215, .330]*	.274 [.214, .330]*
n_D 599/598/600	$O - W$	.036 [.022, .050]*	.036 [.022, .050]*	.032 [.018, .045]*	.036 [.022, .050]*
1x5e-06, M	EZ−LL	.274 [.213, .326]'	.274 [.213, .326]'	.274 [.213, .326]'	.273 [.212, .327]*
n_D 600/599/598	$O - W$	.028 [.014, .042]'	.028 [.014, .042]'	.030 [.014, .044]'	.030 [.015, .045]*
1x5e-05, E	EZ−LL	.084 [.057, .113]*	.084 [.057, .113]*	.084 [.057, .113]*	.084 [.055, .115]*
n_D 582/575/569	$O - W$	.019 [.011, .026]*	.019 [.011, .026]*	.014 [.009, .020]*	.020 [.012, .028]*
1x5e-05, M	EZ−LL	.089 [.063, .116]'	.089 [.063, .116]'	.089 [.063, .116]'	.086 [.060, .114]*

continued

Setting, train	Contrast	Mean, 95% CI, by convention: Imputed M	Short-zone-first E1	Published F rank orbit	subset (D)
n_D 589/583/582	$O - W$	.001 [−.002, .005]	.001 [−.002, .005]	.000 [−.003, .004]	.025 [.011, .041]*
pythia/WikiMIA: pythia-1.4b, WikiMIA					
1x5e-06, E	EZ−LL	.186 [.142, .229]*	.186 [.142, .229]*	.186 [.142, .229]*	.187 [.142, .230]*
n_D 597/595/598	$O - W$	.054 [.025, .086]*	.054 [.025, .086]*	.049 [.028, .072]*	.055 [.027, .088]*
1x5e-06, M	EZ−LL	.197 [.153, .240]'	.197 [.153, .240]'	.197 [.153, .240]'	.199 [.155, .242]*
n_D 594/596/598	$O - W$	.015 [.002, .027]'	.015 [.002, .027]'	.017 [.004, .031]'	.018 [.002, .033]*
1x5e-05, E	EZ−LL	.041 [.018, .066]*	.041 [.018, .066]*	.041 [.018, .066]*	.041 [.014, .071]*
n_D 568/550/539	$O - W$	.044 [.028, .062]*	.044 [.028, .062]*	.033 [.023, .045]*	.045 [.029, .064]*
1x5e-05, M	EZ−LL	.053 [.033, .072]'	.053 [.033, .072]'	.053 [.033, .072]'	.053 [.032, .073]*
n_D 569/572/570	$O - W$	.005 [.001, .010]'	.005 [.001, .010]'	.004 [−.000, .009]	.052 [.031, .078]*
2x5e-05, E	EZ−LL	−.039 [−.053, −.027]*	−.010 [−.019, −.003]*	−.018 [−.034, −.007]*	−.022 [−.039, −.009]*
n_D 418/409/408; tok 4.8	$O - W$	.111 [.088, .136]*	.150 [.122, .181]*	.103 [.084, .126]*	.123 [.088, .163]*
2x5e-05, M	EZ−LL	−.020 [−.031, −.012]'	−.009 [−.020, −.002]'	−.011 [−.021, −.003]'	−.017 [−.033, −.006]*
n_D 459/443/459	$O - W$	.008 [.003, .015]'	.009 [.002, .016]'	.007 [−.001, .015]	.229 [.162, .300]*
arXiv replication: Qwen2.5-0.5B, arXiv					
1x5e-06, E	EZ−LL	.219 [.151, .286]*	.219 [.151, .286]*	.219 [.151, .286]*	.219 [.150, .287]*
n_D 599/598/597	$O - W$	.020 [.006, .033]*	.020 [.006, .033]*	.018 [.006, .030]*	.020 [.006, .033]*
1x5e-06, M	EZ−LL	.212 [.143, .278]'	.212 [.143, .278]'	.212 [.143, .278]'	.210 [.140, .279]*
n_D 598/598/597	$O - W$	.010 [−.006, .027]	.010 [−.006, .027]	.012 [−.002, .027]	.016 [−.001, .033]
2x5e-05, E	EZ−LL	−.027 [−.038, −.017]*	−.001 [−.006, .000]	−.007 [−.016, −.000]*	−.004 [−.017, .000]
n_D 393/393/407; tok 6.2	$O - W$	.024 [.012, .039]*	.044 [.025, .065]*	.029 [.016, .045]*	.021 [.005, .048]*
2x5e-05, M	EZ−LL	−.018 [−.032, −.008]'	−.008 [−.020, −.000]'	−.018 [−.038, −.004]'	−.007 [−.021, −.000]*
n_D 434/456/461	$O - W$	.001 [−.003, .004]	.000 [−.000, .002]	−.002 [−.006, .000]	.267 [.207, .333]*
arXiv replication: pythia-1.4b, arXiv					
1x5e-06, E	EZ−LL	.172 [.117, .226]*	.172 [.117, .226]*	.172 [.117, .226]*	.172 [.117, .226]*
n_D 597/597/599	$O - W$	.061 [.040, .084]*	.061 [.040, .084]*	.055 [.036, .074]*	.061 [.040, .084]*
1x5e-06, M	EZ−LL	.175 [.115, .234]'	.175 [.115, .234]'	.175 [.115, .234]'	.175 [.115, .234]*
n_D 600/600/600	$O - W$	.017 [.002, .032]'	.017 [.002, .032]'	.018 [.003, .032]'	.017 [.002, .032]*
2x5e-05, E	EZ−LL	−.041 [−.056, −.029]*	−.002 [−.006, .002]	−.007 [−.019, .000]	−.005 [−.020, .004]
n_D 405/380/395; tok 5.4	$O - W$	.062 [.048, .078]*	.108 [.086, .134]*	.075 [.060, .093]*	.049 [.029, .072]*
2x5e-05, M	EZ−LL	−.014 [−.023, −.008]'	−.004 [−.010, −.000]'	−.005 [−.012, −.001]'	−.004 [−.010, −.000]*
n_D 447/437/422	$O - W$	.001 [−.001, .003]	.001 [.000, .002]'	.000 [−.003, .003]	.226 [.171, .281]*
Final stage: Qwen2.5-0.5B, WikiMIA					
2x5e-06, E	EZ−LL	.329 [.270, .384]*	.329 [.270, .384]*	.329 [.270, .384]*	.328 [.268, .384]*
n_D 596/595/600	$O - W$	.044 [.030, .059]*	.044 [.030, .059]*	.039 [.027, .051]*	.044 [.030, .059]*
2x5e-06, M	EZ−LL	.327 [.269, .381]'	.327 [.269, .381]'	.327 [.269, .381]'	.326 [.269, .381]*
n_D 599/598/600	$O - W$	.032 [.020, .045]'	.032 [.020, .045]'	.033 [.020, .046]'	.035 [.021, .048]*
Final stage: pythia-1.4b, WikiMIA					
2x5e-06, E	EZ−LL	.250 [.203, .296]*	.250 [.203, .296]*	.250 [.203, .297]*	.252 [.205, .298]*
n_D 596/597/598	$O - W$	.069 [.042, .098]*	.069 [.042, .098]*	.061 [.040, .082]*	.069 [.042, .098]*
2x5e-06, M	EZ−LL	.253 [.207, .298]'	.253 [.207, .298]'	.253 [.207, .298]'	.254 [.209, .299]*
n_D 598/596/597	$O - W$	.017 [.003, .031]'	.017 [.003, .031]'	.020 [.005, .036]'	.019 [.002, .033]*
Final stage: Qwen2.5-0.5B, WikiMIA					
2x5e-05, M_d	EZ−LL	−.216 [−.256, −.180]*	−.032 [−.055, −.014]*	−.141 [−.191, −.093]*	−.007 [−.016, −.002]*
n_D 372/375/401; tok 1.7	$O - W$	.040 [.019, .064]*	.004 [−.000, .013]	−.017 [−.038, .003]	.314 [.213, .414]*
Final stage: pythia-1.4b, WikiMIA					
2x5e-05, M_d	EZ−LL	−.292 [−.333, −.253]*	−.018 [−.030, −.007]*	−.116 [−.150, −.085]*	−.013 [−.027, −.003]*
n_D 348/337/360; tok 1.2	$O - W$	.064 [.035, .093]*	.006 [.000, .015]*	−.012 [−.031, .008]	.571 [.470, .675]*

* interval excludes zero. Train E/M: exact-text/mixed-view; M_d: mixed-view with the exact-text dose held equal; the released rule (U) is in Table 17 and Supplement Table S9; the second row of each cell gives n_D, the items per seed in the D subset (600 otherwise), and tok, the mean error-zone tokens per member on the exact view.

Boundary tokens after changing line breaks. A position is a *boundary* token when its decoded target token contains a newline or is the first token after one (classified on the full prepared token sequence). Every other position is a *content* token. Retained-token reconstruction reproduces all scores. In Qwen WikiMIA exact-text training, width introduces 4.0 boundary positions per member text (sentence view 1.7). Its member boundary mass share is 85.9%, compared with 76.0% for sentence breaks and 94.3 to 96.8% for width in the other three model/corpus cells. Control shares are 21.9 to 45.6%. Mean boundary mass is 6.92 nats for members and 6.71 for controls. Mean content mass is 1.13 and 19.67, respectively. After mixed exposure the width/sentence shares fall to 15.9%/ 3.3%. Dose-matched width shares are 11.0%/30.9% (Qwen/pythia). Section 4.1 reports both the content-only rescore and the distinct denominator-only intervention. The former removes boundary positions from the error zone. The latter retains positive mass and the error-count guard and replaces only negative mass. Replacing the erased denominator by zero uses U's zero-score fallback. This fallback differs from the limit as a positive replacement tends to zero. Replacing the erased mass by common constants .1, 1 and 10 nats yields AUROC ranges .969 to .991, .961 to .984 and .886 to .954 across the four cells, so the recovery does not depend on the single one-nat choice. Changing line breaks can also add or remove content-token errors. Thus, content-only rescoring does not recover the exact-text context. The paraphrase view contains no boundary positions (0.0%) and retains 88 member zeros. From two to four passes, short-zone members rise from 28 to 410 of 900 and contribute 410 of the 598 four-pass zeros. We retain all source-level transitions.

Exact accounting. Here q and r again denote member and control zero fractions. The support identity separates positive/zero pairs and assigns half credit to zero ties (Section 4.1). A_{++} also gives half credit to ties between positive scores and is undefined when there are no positive pairs. Such pairs contribute zero to overall AUROC. Zero-member/positive-control pairs contribute nothing; positive-member/zero-control pairs contribute one, and zero ties contribute one half. These four cases establish the support identity and its upper bound. Let C_{+0} be positive-member/zero-control credit and C_{00} zero-tie credit: $C_{+0} = (1-q)r$ and $C_{00} = qr/2$. Write $b = C_{+0} + C_{00}$ for their sum and $p = (1-q)(1-r)$ for the weight of positive pairs. These local b, p differ from the imputation constant and a statistical p-value. Subscripts 0 and 1 identify the compared views, and δ denotes the change from view 0 to view 1. When both views have positive pairs, expanding the difference of the products pA_{++} gives $\delta A = \delta b + \delta p\,(A_{++,1} + A_{++,0})/2 + \delta A_{++}\,(p_1 + p_0)/2$. If either view has $p = 0$, its positive-pair contribution is zero but A_{++} is undefined. Our analysis reports the total change and leaves the support/ranking split unavailable; assigning an arbitrary value to A_{++} would make that split depend on the convention. The decomposition is descriptive rather than causal. We check it against an independently recomputed rank AUROC (Supplement, Table S3).

The published EZ-MIA evaluation uses 128-token sequences and reports .95 to .99 AUROC under full fine-tuning (Ilić et al., 2026). These AUROCs alone do not establish how often its fallback occurs: the bound applies only if those cases receive zero, as under U. Under that rule, $q \leq (1 - A)/(1 - r/2)$, so negligible control zeros give an approximate $1 - A$ ceiling, or 0.7 to 5% using the unrounded published AUROCs. This conditional ceiling concerns frequency, not the effect on detector rankings. Separately, the paper's Appendices E.5–E.6 report no zero-error or zero-loss cases among 1,000 members and 1,000 nonmembers in that check. We have not rerun the upstream evaluation.

Dose and aggregation details. At two weak passes the fixed comparator matters: the EZ-MIA orbit adds .044 to .069 beyond width but only .022 to .029 beyond sentence breaks (Supplement, Table S9). Of the 900 exact member observations per model in the dose-matched fits, 640 (Qwen) and 746 (pythia) score zero under U, and 471 and 601 of those have fewer than two error tokens. Under F, the dose-matched $O - W$ is .017 [.006,.032] and .028 [.015,.044], and doubling or multiplying the ranked-first cap by ten barely changes it, whereas rank averaging lowers it as reported in Section 4.3. Because ranks rescale finite scores too, this does not isolate imputation. Tables 17 and 18 give all convention/aggregation and four-pass checks.

Detector comparison grid. Table 19 gives the full detector comparison behind Section 4.3. Supplement Table S9 gives every natural-cell $W - E$, $O - W$, and $O - E$, paired as in Table 2, with evaluation-selected comparators excluding released-rule EZ-MIA. Section 4.3 explains the weak-update exceptions and the held-out comparison.

Retained-context Tag&Tab. Following Antebi et al. (2025), let K be the number of keyword spans per sentence. We select the $K = 4$ lowest-entropy word spans per sentence together with named entities. We score each keyword by its first tokenizer piece, average within each sentence, and then weight nonempty sentences equally. The RC suffix marks local adaptations: spaCy `en_core_web_sm` segmentation, wordfreq entropy, a 7 to 40-word sentence filter with whole-document fallback, full retained contexts, and omission of keywords at the unscored initial token, with no later-subtoken substitution or full-text fallback. These adaptations differ from the released pretraining experiment.

RC exact AUROC is .51 to .53 at weak updates and .969 to .995 at strong updates. Adding it to the four-detector evaluation-selected envelope makes it best in 0 of 10 prespecified cells. Across 24 cells, $|O - E| \leq .004$, while $K \in \{2, 8\}$ and no-entity variants change exact AUROC by at most .030. The weak Qwen paraphrase score is .510, trailing the selected exact comparator by .276 ($\mathrm{UB}_{95} = -.212$). We specified the corrected rule before rescoring and archived the original outputs. Hand-worked first-subtoken/sentence tests and unchanged legacy AUROCs check the implementation. RC remains outside the six-candidate source-fold selector.

SPV-MIA on refits. We refit the eight exact-training primary cells on H100 using the recorded recipes, seeds, and data orders because the retained checkpoints could not be transferred and the strong-update Qwen weights were missing. All SPV-MIA comparisons are therefore internal to

E: original; V: the single transformed view named in the panel; O: orbit AUROC. UB_{95} is the one-sided 95% upper bound against the best exact detector selected on evaluation labels. $*$ denotes zero exclusion; ‡/§ mark prespecified/exploratory bounds below .01. RC denotes retained context; superscript r denotes a refit.

Table 19: Full transformation and averaging grid against the evaluation-selected exact-text comparator.

	AUROC			Increments (mean, 95% CI)			vs. best exact	
Detector	E	V	O	$V-E$	$O-V$	$O-E$	**mean**	**UB_{95}**
A. Exact-text training — strong update — width view								
Likelihood	.995	.992	.994	−.003 [−.005, −.001]*	.002 [.000, .003]*	−.001*	−.001	.000§
zlib	.995	.992	.995	−.003 [−.006, −.002]*	.002 [.001, .004]*	−.001*	−.001	.001§
Min-k% 20%	.995	.989	.993	−.006 [−.009, −.003]'	.004 [.002, .007]*	−.002'	−.002	.000§
Min-k++	.993	.824	.955	−.169 [−.201, −.139]'	.131 [.107, .153]*	−.038'	−.040	−.023§
$\text{Tag\&Tab}_{\text{rc}}$	.969	.961	.968	−.008 [−.015, −.003]*	.007 [.003, .012]*	−.001	−.027	−.017§
reference-model detectors								
Reference gain	.995	.995	.994	.000 [−.002, .002]	.000 [−.001, .001]	.000	−.001	.003§
SPV-MIA$^{\text{r}}$	.998	.998	.998	.000 [−.002, .001]	.000 [−.001, .002]	.000	.002	.005§
EZ-MIA	.986	.945	.985	−.041 [−.055, −.028]'	.039 [.027, .053]*	−.001	−.011	−.003‡
EZ-MIA, other cells								
pythia, WikiMIA	.959	.836	.947	−.123 [−.150, −.098]*	.111 [.088, .136]*	−.012*	−.051	−.035‡
Qwen, arXiv	.973	.954	.978	−.019 [−.037, −.004]*	.024 [.012, .039]*	.005	−.022	−.012‡
pythia, arXiv	.958	.889	.952	−.069 [−.087, −.052]'	.062 [.048, .078]*	−.007	−.048	−.032‡
B. Mixed-view training — strong update — width view								
Likelihood	.995	.995	.995	.000 [−.001, .001]	.000 [−.000, .001]	.000	.000	.004§
zlib	.995	.995	.995	.000 [−.001, .001]	.000 [−.000, .001]	.000	.000	.004§
Min-k% 20%	.994	.993	.994	−.001 [−.002, .001]	.001 [−.000, .002]	.000	−.001	.003§
Min-k++	.992	.975	.990	−.016 [−.026, −.009]'	.015 [.008, .023]*	−.002	−.005	.001§
$\text{Tag\&Tab}_{\text{rc}}$	.966	.967	.967	.001 [−.004, .006]	.000 [−.004, .003]	.001	−.028	−.017§
reference-model detectors								
Reference gain	.995	.996	.996	.001 [−.001, .002]	.000 [−.001, .001]	.001*	.001	.001§
EZ-MIA	.986	.994	.994	.008 [.002, .016]*	.001 [−.001, .002]	.008*	−.001	.003§
EZ-MIA, other cells								
pythia, WikiMIA	.977	.984	.992	.006 [−.005, .020]	**.008** [.003, .015]*	.015*	−.006	−.001§
Qwen, arXiv	.982	.995	.996	.013 [.005, .025]*	.001 [−.003, .004]	.014*	−.004	−.001§
pythia, arXiv	.985	.996	.998	.011 [.005, .018]*	.001 [−.001, .003]	.012'	−.002	.000§
exact-dose-matched training (M_{d})								
EZ-MIA	.784	.843	.883	.060 [.026, .095]*	**.040** [.019, .064]*	.100*	−.117	−.095§
Likelihood	1.000	1.000	1.000	.000 [−.000, .000]	.000 [−.000, .000]	.000	.000	.000§
EZ-MIA, pythia	.708	.758	.822	.050 [.016, .083]*	**.064** [.035, .093]*	.114'	−.178	−.149§
EZ-MIA (F, arith.)	.859	.980	.997	.121 [.066, .175]*	**.017** [.006, .032]*	.138'	−.003	.000§
EZ-MIA (F, ranks)	.859	.980	.962	.121 [.066, .175]*	−.017 [−.038, .003]	.104*	−.038	−.023§
pythia (F, arith.)	.884	.964	.992	.080 [.047, .115]*	**.028** [.015, .044]*	.108*	−.008	−.001§
pythia (F, ranks)	.884	.964	.952	.080 [.047, .115]*	−.012 [−.031, .008]	.068'	−.048	−.031§
C. Exact-text training — weak update — released paraphrase								
EZ-MIA	.786	.762	.782	−.024 [−.048, −.001]*	.020 [.007, .034]*	−.004	−.024	−.001‡
Likelihood	.512	.505	.509	−.007 [−.016, .002]	.003 [−.001, .008]	−.003	−.281	−.217§
zlib	.510	.502	.506	−.007 [−.018, .002]	.004 [−.001, .009]	−.004	−.284	−.223§
Min-k% 20%	.526	.519	.522	−.007 [−.019, .005]	.003 [−.003, .009]	−.004	−.267	−.208§
Min-k++	.533	.522	.529	−.011 [−.028, .005]	.006 [−.002, .015]	−.004	−.264	−.200§
$\text{Tag\&Tab}_{\text{rc}}$	.513	.510	.512	−.003 [−.015, .009]	.002 [−.004, .008]	−.001	−.276	−.212§
Reference gain	.723	.715	.721	−.009 [−.018, .001]	.006 [.001, .011]*	−.003	−.071	−.036§
EZ-MIA, pythia	.712	.697	.712	−.015 [−.035, .003]	.015 [.003, .028]*	.000	−.015	.003‡

Qwen2.5-0.5B/WikiMIA unless named. Strong: two passes at 5×10^{-5}; weak: one pass at 5×10^{-6}. V is width in A/B and paraphrase in C; O is the orbit; M_{d} is dose-matched mixed exposure. EZ-MIA uses M unless F is named. The exact-text comparator is evaluation-selected; Table 12 gives held-out selection. $\text{Tag\&Tab}_{\text{rc}}$ is the retained-context adaptation. SPV-MIA$^{\text{r}}$ and its comparator are computed within the H100 refit and are not paired with retained rows. $*$: interval excludes zero; bold: orbit beats E and V; ‡/§: prespecified/exploratory $\text{UB}_{95} < .01$.

these refits. Their exact-text likelihood and EZ-MIA (M) AUROCs differ from the retained values by at most .001 and .010, respectively. The selected comparator agrees in 8 of 8 cells, and 0 refits are flagged.

For our adaptation of Fu et al. (2024), the target generates 600 samples of 128 tokens from held-out pool prefixes. We fine-tune the base model on these samples for one pass to obtain the reference model. In this formula only, subscripts t, r identify target and reference models, not token position or control zero fraction. Let x be a view, $\tilde{x}$ its rewritten neighbours, and $p = e^{-\text{loss}}$, where loss is mean token negative log likelihood. A bar denotes an average over ten neighbours produced by the t5-base text-infilling model. We score $[p_t(x) - \bar{p}_t(\tilde{x})] - [p_r(x) - \bar{p}_r(\tilde{x})]$. The neighbour count and the 600-by-128-token, one-pass reference recipe are local choices.

SPV-MIA reaches .76 to .79 at the weak update and .997 to 1.000 at the strong update. It is the best exact detector in 5 cells, trails the four-candidate refit comparator by at most .011, and leads it by up

to .059. Width changes SPV-MIA by at most .071, and the orbit adds at most .023, with a resolved gain in 1 weak cell. Consequently, 9 of 16 width/orbit rows meet the .01 bound, and the largest upper bound is .105 (Supplement, Table S9).

Adding SPV-MIA's own exact score to the evaluation-selected refit envelope changes the comparison. With the envelope fixed across draws, 11 of 16 rows meet the .01 bound, 5 do not, and weak-cell orbit upper bounds span .027 to .035. The orbit resolves a gain over this envelope in 1 cell. Our frozen addendum and dated erratum record all local adaptations. Sequence loss averages active tokens with padding masked. The 600 reference samples, twice the member count, come from 164 held-out pool prefixes. Reference training matches the weak target duration but is shorter than strong target training, and neighbours are shared across seeds. Likelihood agreement checks the target-training recipe. The separately generated reference-model pipeline still lacks an end-to-end parity check.

Prospective replication and larger model. The frozen predictions confirm the ranking reversal and zero collapse, but not the 1.5B width prediction or the SPV-MIA contrast pattern. We specified six criteria before training any new fit. We test them on twelve Qwen2.5-0.5B fits with new membership assignments across both pools and both weak and strong exact-text training, and on six Qwen2.5-1.5B fits using the original WikiMIA assignments. Because the predictions use different criteria, Table 20 reports each outcome separately, and n/a indicates that no 1.5B WikiMIA prediction was specified.

The ranking reversal, strong-update likelihood lead, and released-rule zero collapse hold in every applicable cell. Under M, EZ-MIA leads likelihood at the weak update: .790 versus .527 on WikiMIA, .751 versus .515 on arXiv, and .796 versus .507 at 1.5B. Likelihood leads after the strong update. Under the released rule, the width increments are .287, .649, and .259, with 293, 624, and 313 zero scores among 900 member observations. The frozen threshold requires at least 100 such zeros.

The width-sensitivity criterion requires a width gain above .15 under U and gains within $\pm.05$ under M and F. Both new 0.5B assignments satisfy this criterion, but the 1.5B test does not: its M and F increments are $-.078$ and $-.079$, compared with .259 under U, approximately 3.3 times their magnitude.

The SPV-MIA increment criterion freezes the earlier qualitative pattern of a positive $O - E$ estimate with an unresolved $O - W$, rather than a minimum-effect criterion. It fails in both new weak cells because $O - W$ is resolved positive: .035 [.018, .053] and .024 [.006, .043]. In both cases, the orbit recovers a loss induced by the width view. The corresponding $O - E$ estimates are .003 [$-.010$, .015] and .008 [$-.005$, .021]. They are positive but unresolved, so the new fits neither resolve an own-exact gain nor establish that the effects differ. The exact-score agreement criterion, which requires SPV-MIA exact AUROC to lie within .05 of EZ-MIA, holds in both cells.

SPV component parity. The parity test covers neighbour fills, masked losses, and score arithmetic under shared masks. For 80 texts, upstream code (commit ee9ebaa) and ours share upstream-masked inputs: 80/80 neighbours and span counts match. Batch-one upstream and batched masked losses differ by at most 1e-5 over 440 sequences. Score arithmetic matches on 1800 rows. Independent masking, reference generation, and sample count remain unchecked local adaptations. A dated addendum records the executable paths and the MPS fallback used.

C Released paraphrases and a neighbourhood comparator

The paraphrase study asks whether changing wording yields gains similar to those from changing line breaks. We fixed its protocol after seeing the formatting results but before scoring the WikiMIA paraphrases. It reuses six Qwen WikiMIA two-pass checkpoints, scores, and assignments without further training. All 764 pairs fit within 128 tokens (maximum 117), and three pairs are identical.

A natural-language-inference (NLI) model estimates whether either text contradicts the other; bidirectional means checking both text orders. A fixed maximum-probability cutoff of .5 retains 752 unique sources for the automated sensitivity in Table 21, which we report alongside the all-pair analysis. The screen estimates fidelity; without a blinded human audit, it cannot establish fidelity for individual pairs.

Table 20: Frozen prospective criteria and outcomes, weak-update SPV-MIA results, and replication-protocol predictions.

Prediction	Frozen criterion	0.5B WikiMIA	0.5B arXiv	1.5B WikiMIA
Ranking reversal	Weak EZ(M)−LL: positive 95% interval	Pass	Pass	Pass
Likelihood lead	Strong LL−EZ(M): positive point	Pass	Pass	Pass
Width sensitivity	U $W - E > .15$ and $\lvert W - E \rvert < .05$ under M/F	Pass	Pass	Fail: $-.078/-.079$
Zero collapse	Strong exact member zeros under U ≥ 100	Pass: 293	Pass: 624	Pass: 313
SPV increment	Weak SPV $O - E > 0$ with $O - W$ unresolved	Fail: $O - W > 0$	Fail: $O - W > 0$	n/a
SPV exact agreement	Weak SPV exact within .05 of EZ(M) exact	Pass	Pass	n/a

SPV-MIA cohort	Cell (weak)	E	W	O	$O - E$ [95% CI]	$O - W$ [95% CI]
original refit	Qwen/WikiMIA	.778	.756	.793	.015 [−.003, .036]	.037* [.018, .055]
original refit	pythia/WikiMIA	.774	.727	.787	.013 [−.004, .030]	.060* [.020, .096]
original refit	Qwen/arXiv	.789	.793	.812	.023* [.008, .037]	.019 [−.005, .041]
original refit	pythia/arXiv	.761	.690	.771	.010 [−.011, .031]	.081* [.054, .110]
prospective	Qwen/WikiMIA	.808	.775	.811	.003 [−.010, .015]	.035* [.018, .053]
prospective	Qwen/arXiv	.790	.774	.798	.008 [−.005, .021]	.024* [.006, .043]
prospective	Qwen 1.5B/WikiMIA	.773	.760	.782	.009 [−.005, .023]	.023* [.005, .042]

SPV-MIA scores on the four original weak refit cells and the three prospective weak cohorts, all under adapted local reference generation and neighbour construction. Component parity (neighbour fills, masked losses and score arithmetic under shared masks) holds, whereas the independently generated reference pipeline is not checked end to end. Intervals are three-fit, source-paired and unadjusted, and $*$ marks an interval excluding zero.

Prediction	Frozen criterion (first replication protocol)	Cells passing
Boundary erasure	Width AUROC falls when boundary loss is erased	4 of 4
Constant replacement	One-nat replacement above .95, interval over erasure excludes zero	4 of 4
Exact stabilization	Exact-text $P/(N+1)$ above .93	4 of 4
96-word reversal	Weak EZ(F)−LL positive and strong negative	0 of 2 (strong: Qwen .000, pythia .001)
96-word stabilization	Strong exact-text $P/(N+1)$ above U	2 of 2

Predictions frozen in the first replication protocol before its fits (Appendix E.4). The first three criteria are tested in four fresh strong cohorts and the final two on 96-word arXiv prefixes. The 96-word ranking-reversal criterion fails because EZ(F) and likelihood both reach the ceiling after strong updates.

For EZ-MIA, paraphrases contribute more through transformation than through averaging, but less than width on the same checkpoints. The transformation increment is .205 [.160,.252], compared with .2705 [.2260,.3179] for width after exact-text training, whereas the averaging increment is .009 [.001,.020] (Supplement, Table S8). Cosine similarity measures alignment between text-embedding vectors. The additional fixed grid crosses text-embedding cosine-similarity thresholds .80/.90/.95 with bidirectional-contradiction ceilings .10/.30/.50. The nine screens retain 680 to 750 sources and give weak exact-training $O - E$ between $-.0040$ and $.0015$, while residual likelihood dependence remains high (m_{eff} from 1.012 to 1.022, Table 21). These unvalidated cutoffs alter the cohort but do not identify a causal effect of fidelity. For the two weak exact-training cells, we compute M imputation before filtering. We estimate likelihood dependence after subtracting the class means separately within each fit. Operating-point and screen specifications were frozen before calculation.

Neighbourhood comparator. The retained-row analogue of the variant-based attack is exact log likelihood minus the mean neighbour log likelihood. Where all views are retained, the neighbour set contains the released paraphrase and both formatting views (NB-3: four target evaluations and no reference model). Elsewhere, it contains only the two formatting views (NB-2: three evaluations). We fix the score orientation before calculation and perform no new inference or training, so this design differs from the original neighbour-generation pipeline.

After two strong passes with exact-text training, NB-3 reaches .8525. It exceeds released-rule EZ-MIA on exact text but trails the EZ-MIA orbit (Table 22). After mixed-view training, its AUROC is .783, and neither the contrast with exact EZ-MIA nor the contrast with orbit EZ-MIA is resolved. The neighbourhood score trails likelihood by .14 to .21 after two strong passes and by .20 to .28 after one pass at 5×10^{-5} using two formatting neighbours. After one weak pass, both neighbourhood variants are near chance (.5179).

D Additional study protocols

The simulations isolate assumptions about dependence and difficulty; the temporal-label audit tests whether date groups differ in their text. Neither analysis replaces randomized natural-text inclusion.

R denotes the released paraphrase. NLI is natural-language inference; its maximum contradiction probability is checked in both text orders. Cosine measures embedding-vector alignment; min/max are lower/upper screening cutoffs. m_{eff} is an effective view count from within-class likelihood correlations.

Table 21: Paraphrase orbit gains across fidelity screens under U and M.

WikiMIA paraphrase / exp. / detector	Exact	Para.	Orbit	$O - E$ [95% CI]	$O - R$ [95% CI]
E/EZ-MIA	.6730	.8799	.8878	.2147 [.1709,.2585]	.0079 [-.0007,.0180]
E/Log likelihood	.9951	.9855	.9928	-.0023 [-.0039,-.0010]	.0073 [.0041,.0114]
E/Min-k++	.9942	.9388	.9820	-.0122 [-.0202,-.0058]	.0432 [.0322,.0558]
E/Reference gain	.9946	.9939	.9946	.0000 [-.0008,.0013]	.0007 [-.0001,.0015]
M/EZ-MIA	.7368	.9010	.9113	.1745 [.1349,.2175]	.0103 [.0021,.0195]
M/Log likelihood	.9949	.9846	.9924	-.0025 [-.0042,-.0012]	.0078 [.0046,.0119]
M/Min-k++	.9923	.9416	.9806	-.0117 [-.0183,-.0061]	.0390 [.0287,.0501]
M/Reference gain	.9949	.9941	.9948	-.0001 [-.0009,.0012]	.0007 [.0000,.0014]

Cosine min / NLI max	Sources	Qwen E/O	Qwen $O - E$ [95% CI]	m_{eff}	pythia E/O	pythia $O - E$ [95% CI]	m_{eff}
all	764	.786/.782	−.004 [−.015, .006]	1.014	.712/.712	.000 [−.008, .009]	1.020
.80 / .10	725	.784/.780	−.003 [−.016, .008]	1.014	.705/.706	.001 [−.007, .010]	1.021
.80 / .30	745	.787/.783	−.004 [−.015, .007]	1.014	.709/.710	.001 [−.007, .010]	1.021
.80 / .50	750	.787/.783	−.004 [−.015, .007]	1.014	.711/.712	.001 [−.007, .010]	1.021
.90 / .10	718	.784/.781	−.003 [−.015, .008]	1.014	.703/.705	.001 [−.007, .010]	1.022
.90 / .30	738	.787/.783	−.004 [−.015, .006]	1.014	.708/.709	.001 [−.007, .010]	1.021
.90 / .50	743	.787/.783	−.004 [−.015, .006]	1.014	.709/.710	.001 [−.007, .010]	1.021
.95 / .10	680	.779/.777	−.002 [−.013, .008]	1.012	.701/.703	.002 [−.007, .011]	1.020
.95 / .30	695	.781/.778	−.003 [−.013, .007]	1.012	.705/.706	.001 [−.007, .010]	1.020
.95 / .50	700	.781/.778	−.003 [−.014, .007]	1.012	.707/.708	.001 [−.008, .010]	1.020

Fixed grid: weak exact training, convention M; all nine screens retained. NLI: maximum bidirectional contradiction. Three-fit, source-paired conditional intervals; preprocessing fixed before filtering. Descriptive $m_{\text{eff}} = 2/(1+\rho)$ averages per-fit within-class residual likelihood correlations. All-pair analysis remains primary.

Table 22: Mean AUROC and cost for the adapted retained-neighbour detector and exact-text comparators.

Study / exp.	Cost	NB	EZ exact	EZ orbit	Ref. gain	LL exact	NB − EZ exact	NB − EZ orbit
Qwen/WikiMIA, two-pass/E	4 calls	.8525	.6747	.9390	.9947	.9951	.1779 [.1285,.2259]	-.0864 [-.1211,-.0567]
Qwen/WikiMIA, two-pass/M	4 calls	.7834	.7387	.8304	.9950	.9949	.0447 [-.0124,.1024]	-.0471 [-.0973,.0013]
Qwen/WikiMIA, one-pass 5e-06/E	3 calls	.5179	.7828	.7819	.7233	.5119	-.2649 [-.3215,-.2108]	-.2639 [-.3227,-.2058]
Qwen/WikiMIA, one-pass 5e-06/M	3 calls	.5090	.7813	.7880	.7173	.5110	-.2723 [-.3251,-.2203]	-.2790 [-.3306,-.2291]
Qwen/WikiMIA, one-pass 5e-05/E	3 calls	.6914	.8957	.9703	.9838	.8940	-.2043 [-.2494,-.1617]	-.2789 [-.3147,-.2440]
Qwen/WikiMIA, one-pass 5e-05/M	3 calls	.6031	.9269	.9544	.9828	.8879	-.3238 [-.3745,-.2691]	-.3513 [-.3977,-.3020]

Simulations and replication. The independent-view simulation contains 60 members and 60 controls. We set the independent-view base probability to .55, vary the member increment over $\delta \in \{0, .05, .10, .20\}$ and the view count over $m \in \{1, 3, 5, 10\}$, and use 199 permutations and 1,000 replicates per cell. The difficulty-confounding simulation instead uses 5,000 replicates per shift, five views with correlation .9, and 60 items per group. Its conditional membership effect is zero, while the mean difficulty shift ranges from 0 to .5. The 1.5B synthetic replication uses two rather than four repeats, with two epochs, 45 members, 45 controls, and three seeds. We also run two separate single-seed continued supervised fine-tuning (SFT) checks.

Temporal-label audit. We score 776 WikiMIA pairs (387/389 date labels) (Shi et al., 2024) with Qwen2.5 at 0.5B, 1.5B, and 7B parameters and SmolLM2 at 135M, 360M, and 1.7B. For loss, zlib, Min-k, and Min-k++, we use 2,000 paired resamples without tuning on labels. Screening excludes pairs with cosine similarity below .8 or contradiction probability above .5.

Logistic calibration fits a classifier to the score features on four folds and evaluates on the fifth. Repeated five times, it combines the exact and paraphrase scores, their difference (the temporal audit's specificity feature), the orbit mean, and log token count. Token-length matching permutes labels within ten bins. Paired stratified bootstraps then compare out-of-fold, exact, and orbit AUROCs. To assess semantic balance, the frozen encoders MiniLM, BGE-small-en-v1.5 and E5-small-v2 represent all exact texts. Predicted date-label probabilities, called propensity scores, determine the weights. We estimate them on other folds and apply clipped, stabilized inverse-probability weights to each exact and orbit Min-k++ AUROC.

E AUXILIARY RESULT DETAILS

Combining detector scores. Logistic calibration of the score features yields temporal-label AUROC of .469 to .539 across six models, below the orbit mean. Shrinkage reaches .506 to .565 on WikiMIA Min-k++, remains below the mean on all three SmolLM2 models, and loses in three of four mixed cells (Table 12). Length-matched SmolLM2-1.7B retains .558 AUROC over 387 pairs, whereas four models remain near chance.

View dependence and training exposure. Let λ_i be eigenvalue i of the view-score correlation matrix, with all eigenvalues nonnegative. The effective rank summarizes how many directions account for score variation. It is $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$. It differs from the effective view count used to summarize averaging variance. Exact-exposure code has an effective rank of 1.08. Mixed exposure raises the rank in five of six seed and domain pairs, an association rather than a causal estimate. The 1.5B exposure effects are $+.0174$ for code and $+.0448$ for MCQ, with both joint intervals crossing zero. The continued-training checks change orbit AUROC from .615 to .593 for code and from .622 to .550 for MCQ. As a single-seed diagnostic, this result does not establish unlearning.

Textual differences between date groups. MiniLM, BGE, and E5 predict the temporal labels at AUROCs of .973, .989, and .990, respectively. Yet propensity weighting changes Min-k++ by at most .017, and every orbit interval crosses zero (Supplement Table S14).

E.1 Analysis timing and comparison families

The timeline distinguishes tests specified before training from analyses chosen after results were available. U and M preceded their contrasts, but the first Qwen WikiMIA results preceded the eight-contrast family, Holm adjustment and the retrospective .01 margin. E1 accompanied the additional fits, and F was specified afterward.

In the timeline, SD means standard deviation and R1 to R4 identify four retrospective checks, not paraphrase views. A frozen specification is one recorded before the stated computation or training runs.

Analysis	What preceded the evidence?	Inference status
Eight M EZ−LL contrasts	Six pythia/arXiv contrasts and the weak-positive, strong-negative sign rule frozen before fits, with the initial two Qwen results already seen	Holm within eight, bootstrap validity still conditional
Twenty M view contrasts	Family and .01 margin followed initial results, comparator chosen on evaluation labels	Pointwise upper bounds, separate two-sided 90% equivalence checks
Boundary and 30 F selector contrasts	Specification frozen before reanalysis but after earlier results, no new fits	Pointwise conditional bounds, fixed preprocessing
SPV refits	Local pipeline specified before scoring, interpretation informed by earlier experiments	Exploratory, unadjusted within-refit contrasts
Prospective prediction set	Six criteria frozen before all 18 new-assignment and 1.5B fits	Criterion-specific outcomes, not one replication rate
Bonferroni sensitivity	30-contrast family, 30,000 draws, seed and tail probability fixed after the earlier 28/30 result	Retrospective multiplicity check, no coverage claim for new runs
First replication protocol (72 fits)	Estimands, bounds, Bonferroni family and sample size fixed before training (twelve fits give a Bonferroni half-width near .0095 when the per-fit SD is .012)	Fit-level t bounds, confirmatory for the weak cells
Second replication protocol (108 fits)	Mid and weak twelve-fit cells, five paraphrases, the training mixture, their estimands and four predictions fixed before any paraphrase or fit	Fit-level t bounds, three-fit conditional bounds for the mixture
Retrospective checks R1 to R4	Zero causes, low-FPR rates, disjoint caps and fit-level reversal tests fixed in the second protocol before computation	Descriptive or fit-level, on retained rows

E.2 Multiplicity, per-fit effects and selector stability

For the retrospective multiplicity sensitivity, we rerun the ten cells, five folds, six candidates, and F preprocessing with 30,000 draws (seed 20260922). Let j index a contrast, d_j^* its bootstrap values, and Q their empirical quantile function. Its upper bound is U_j, distinct from scoring rule U. We take $U_j = Q_{1-.05/30}(d_j^*)$, which gives simultaneous coverage only if each component bound is

calibrated. With three fits, we interpret these as approximate bounds conditional on the observed training runs.

Twenty-one of 30 bounds fall below .01, against 28 of 30 pointwise bounds. All nine exceptions are weak-update comparisons, namely both orbits in every model and corpus setting plus Qwen arXiv width. The largest bound is .0421 for Qwen arXiv arithmetic averaging. A bound crossing .01 leaves the contrast unresolved rather than demonstrating a benefit, and the two original pointwise exceptions remain .0209 and .0196.

Selection frequencies (Supplement Table S17) show that F is chosen in at least 99.8% of weak WikiMIA draws, compared with 88.7% and 90.3% in the weak arXiv cells. In strong cells, several near-ceiling detectors divide the selections. The selected detector varies across resamples, especially near the ceiling. Every selection still excludes the target fit and all test-fold sources. Per-fit effects (Supplement Table S18) show how the observed contrasts vary across training runs; they add no interval claim. The 12 constant-replacement effects are all positive. In each fit, both dose-matched cells have positive arithmetic but negative rank $O - W$. SPV-MIA's Qwen arXiv weak gain over its own exact score is positive in all three refits, whereas the other SPV-MIA cells vary.

E.3 Zero scores by error-zone size

Among the 3,600 member observations from the four strong exact-training boundary cohorts (three fits each), those with 0, 1, 2 to 4, 5 to 9 and 10 or more error tokens number 86, 198, 1,096, 1,672 and 548, of which 86, 198, 787, 859 and 147 score zero under U. Nearly all 3,600 control observations have ten or more errors (3,578, the other 22 have five to nine), and only 4 control observations score zero. Every empty or one-error zone therefore maps to zero, but zero scores also occur with ten or more errors. Short-zone handling alone cannot explain the pattern. Because the models tokenize 32-word inputs differently and training changes the error-zone size, these conditional counts describe the retained cohorts and do not identify a passage-length effect or predict long-document behavior.

E.4 Fresh-assignment and longer-passage replication

Before training, we froze a replication protocol, run on an H100 GPU, for 72 exact-training fits: twelve fresh-assignment fits per weak cell, three fresh strong fits per model and corpus setting with the boundary split, and 96-word arXiv prefixes from 701 qualifying abstracts. Every fresh strong cohort reproduces the denominator intervention and stabilization (Table 1C). The weak-update extension uses twelve fits per cell and treats fits as the unit for the held-out comparisons in Table 12. Its one-sided t bounds and Bonferroni bounds across twelve comparisons are separate from the original three-fit bootstrap. At 96 words, U still assigns zero to 52 to 71 of 300 members per fit because negative mass is zero. EZ(F) and likelihood both reach the ceiling after strong updates, so their predicted ranking reversal fails (Table 20). The retraining used to evaluate Infilling Score reproduces 67 parent row files byte for byte (Table 12).

E.5 Second replication: mid updates, paraphrases, and a training mixture

Before paraphrase generation or training, we froze the target quantities, bounds, and predictions. All fits use exact-text training and new membership-assignment seeds; manifests record the recipe and NVIDIA GB10 or Apple GPU.

Paraphrases. For each of the 764 sources of both pools, llama3.1-8B generated five paraphrases in separate calls and seeds, at temperature .9, which controls sampling variability. A fixed prompt requests the same facts and approximate word count. Accepted texts have 20 to 48 words and fit both tokenizers' padded length. Separate generation calls do not imply independent scores: all five versions share a source passage. The screen of Appendix C does not filter the primary analysis, and we report a screened sensitivity that keeps sources whose five paraphrases all have cosine similarity at least .80 and bidirectional contradiction at most .50.

Cells. Mid updates use one pass at 5×10^{-5} and weak updates one pass at 5×10^{-6}, with twelve fits per model and pool at each strength (96 fits). Every fit scores the three layout views and the five paraphrases. The mixture study embeds the 300 members of each WikiMIA assignment among

5,700 background passages, the first 32 words of Wikipedia passages from the Natural Questions (NQ) collection (Kwiatkowski et al., 2019) in the Benchmarking-IR (BEIR) information-retrieval benchmark (Thakur et al., 2021) that share no eight-word sequence with the pool, and trains one pass over all 6,000 passages at both rates (three fits per model and rate, 12 fits).

Target quantities. Mid cells repeat the held-out comparison of Table 12 with fits as the unit. Let k count paraphrases. The paraphrase orbit O_k reports AUROC after combining the exact text and first k paraphrases. We evaluate both arithmetic and rank means under F; displayed F contrasts use ranks and LL uses raw scores. The primary paraphrase contrast is O_5 minus the held-out selected exact detector, with Bonferroni bounds over 16 comparisons. Secondary contrasts are $O_k - E$ for $k = 1, \ldots, 5$, the orbit of all eight views, O_{all}, minus the best single view selected on other fits, and the effective number of views. Let $\bar{\rho}$ be the mean pairwise correlation of likelihood residuals over the six texts, after subtracting each class mean within each fit. Then $m_{\text{eff}} = 6/(1 + 5\bar{\rho})$ summarizes dependence after standardizing those residuals to unit variance. It describes variance reduction for standardized scores, not the actual raw-score average or the membership signal of individual views. The mixture cells use the three-fit conditional bootstrap of the core studies.

Results. Tables 23 to 26 cover every cell. Restricting to the 570 WikiMIA and 560 arXiv sources whose five paraphrases pass the screen changes the contrasts little (Table 24, last column). Computing the arithmetic cap from the other fits instead of the evaluated cohort leaves every twelve-fit weak contrast unchanged. Orbit ranks move in only 7 of 48 fits, and those changes stay within a class, preserving member–control rankings. The frozen predictions have mixed outcomes. For the paraphrase study, the predictions are that the five-paraphrase average stays below the held-out detector plus .01 in every cell (holds) and likelihood has fewer than two effective views in every cell (holds). For the mixture, they are that EZ-MIA leads likelihood after weak updates (holds) and all orbit bounds fall below .01 (fails).

EZ: EZ-MIA; U/F: released/published rules; LL: likelihood; Gain: reference gain; rank 1 is best among six detectors. Sel. denotes the held-out exact detector. W is width; O_{arith} and O_{rank} average raw scores and within-view ranks. Brackets give one-sided 95% t/Bonferroni upper bounds over 12 comparisons.

Table 23: Mid-update AUROCs, EZ-MIA detector ranks, and held-out contrasts across twelve fresh fits per cell.

Mid cell	EZ (U)	EZ (F)	LL	Gain	Rank U/F	$W-$sel.	$O_{\text{arith}}-$sel.	$O_{\text{rank}}-$sel.
Qwen/WikiMIA	.880	.971	.904	.979	5.8/2.0	−.027 [−.024/−.021]	−.010 [−.008/−.007]	−.014 [−.012/−.011]
pythia/WikiMIA	.791	.970	.930	.981	6.0/2.0	−.057 [−.052/−.048]	−.013 [−.010/−.007]	−.024 [−.021/−.018]
Qwen/arXiv	.768	.995	.954	.995	6.0/1.3	−.009 [−.007/−.006]	−.001 [−.001/.000]	−.003 [−.002/−.001]
pythia/arXiv	.784	.994	.955	.991	6.0/1.1	−.017 [−.014/−.012]	.000 [.000/.000]	−.004 [−.003/−.002]

E: original-text AUROC; O_k: original plus k paraphrases; O_{all}: three layout views plus five paraphrases. F uses rank means; LL uses arithmetic means. Sel./best view are chosen on other fits. m_{eff} summarizes likelihood dependence over six texts. Brackets are pointwise/Bonferroni upper bounds (16 comparisons); Screened is the mean O_5-sel. difference after fidelity filtering. Differences are fit-level means.

Table 24: Separately generated paraphrases: averaging effects and dependence across twelve fresh fits per cell.

Cell	Update	$O_1 - E$ (F)	$O_5 - E$ (F)	$O_5 - E$ (LL)	O_5-sel.	$O_{\text{all}}-$best view	m_{eff}	Screened
Qwen/WikiMIA	weak	−.063	−.121	−.008	−.120 [−.106/−.094]	−.071 [−.060/−.051]	1.41	−.117
pythia/WikiMIA	weak	−.050	−.101	−.014	−.102 [−.090/−.080]	−.062 [−.052/−.043]	1.60	−.093
Qwen/arXiv	weak	−.055	−.117	−.009	−.116 [−.110/−.104]	−.063 [−.058/−.054]	1.64	−.119
pythia/arXiv	weak	−.052	−.094	−.009	−.096 [−.088/−.080]	−.062 [−.056/−.051]	1.70	−.099
Qwen/WikiMIA	mid	−.053	−.089	−.170	−.097 [−.091/−.085]	−.044 [−.040/−.036]	1.50	−.095
pythia/WikiMIA	mid	−.051	−.093	−.153	−.103 [−.092/−.083]	−.050 [−.043/−.037]	1.69	−.100
Qwen/arXiv	mid	−.051	−.085	−.152	−.084 [−.079/−.074]	−.041 [−.039/−.036]	1.77	−.080
pythia/arXiv	mid	−.055	−.100	−.161	−.100 [−.096/−.092]	−.053 [−.050/−.047]	1.86	−.094

EZ(F/U): EZ-MIA under published/released rules; LL: likelihood; Gain: reference gain; zeros count members under U across three fits. W: width; $O_{\mathrm{arith/rank}}$: arithmetic/rank orbit; sel.: held-out exact detector. EZ(F)−LL brackets are conditional 95% intervals; orbit brackets are one-sided 95% upper bounds.

Table 25: Detection when members form 5% of a 6,000-passage training mixture, with three fits per cell.

Mixture cell	Update	EZ (U)	EZ (F)	LL	Gain	EZ (F)−LL	Zeros	O_{arith}−sel.	O_{rank}−sel.
Qwen/WikiMIA	weak	.676	.677	.494	.672	.183 [.130, .235]	1	.000 [.019]	−.005 [.017]
Qwen/WikiMIA	mid	.948	.985	.857	.985	.128 [.103, .155]	33	.000 [.004]	−.002 [.002]
pythia/WikiMIA	weak	.683	.687	.499	.657	.188 [.146, .232]	3	.003 [.018]	−.001 [.021]
pythia/WikiMIA	mid	.861	.982	.909	.988	.073 [.055, .093]	110	−.007 [.000]	−.016 [−.008]

FPR: fraction of controls flagged; TPR: fraction of members detected. EZ(F) and LL are exact-text EZ-MIA and likelihood; O_{arith} and O_{rank} combine layout scores or ranks. Brackets give 95% t intervals over twelve fits, with 300 controls per fit.

Table 26: True-positive rates at 1% and 5% FPR in the twelve-fit weak cells.

Weak cell	FPR	EZ (F) TPR	LL TPR	EZ (F)−LL	O_{arith}−EZ (F)	O_{rank}−EZ (F)
Qwen/WikiMIA	1%	.148	.015	.133 [.102, .164]	.006 [−.014, .026]	−.003 [−.033, .027]
Qwen/WikiMIA	5%	.329	.057	.272 [.232, .311]	.004 [−.011, .020]	−.003 [−.025, .019]
pythia/WikiMIA	1%	.066	.014	.052 [.028, .076]	−.001 [−.012, .010]	−.001 [−.019, .018]
pythia/WikiMIA	5%	.182	.056	.126 [.103, .150]	−.002 [−.019, .016]	−.006 [−.025, .014]
Qwen/arXiv	1%	.095	.013	.082 [.056, .108]	.017 [.002, .031]	.021 [.004, .039]
Qwen/arXiv	5%	.249	.052	.197 [.171, .223]	.002 [−.016, .021]	.005 [−.017, .027]
pythia/arXiv	1%	.066	.016	.050 [.030, .071]	.004 [−.003, .012]	.002 [−.011, .016]
pythia/arXiv	5%	.197	.053	.144 [.121, .167]	.005 [−.008, .019]	−.013 [−.038, .011]