DATA2DELTA: GRADIENT GEOMETRY FOR BUDGETED PRETRAINING DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

When deduplication leaves more data than a pretraining budget permits, a selector must choose among examples that are not literal copies. We introduce Data2Delta, which measures redundancy in a fixed probe's gradient space: it clusters normalized gradient sketches and globally removes examples with low within-cluster ridge leverage. The cluster-average score is its effective dimension divided by its size, linking deletion allocation to directional redundancy without explicit cluster quotas. Two controlled Pythia-1B comparisons support the representation and membership choices: under the same selector, gradients improve endpoint perplexity (PPL) over dimension-matched BGE-M3 features; with each cluster's deletion count fixed, leverage-ranked membership improves PPL over random membership. End to end, Data2Delta lowers PPL by 0.0154 relative to uniform selection. Curve interpolation translates this gap into about 19% fewer effective prediction tokens to reach uniform selection's endpoint PPL. Low-NLL filtering achieves similar PPL, yet loses over 13 times as many source families in a separate natural-news evaluation at 10% removal. Training-pool provenance analysis also distinguishes the selections, while mean zero-shot accuracy is lower for Data2Delta than for the controls on the perturbed pool. These results support gradient geometry as a candidate-relative selection signal and distinguish its training-fit gains from source coverage and downstream accuracy.

1 INTRODUCTION

Deduplication answers whether a corpus contains repeated text. A fixed pretraining budget poses a further question: which examples should remain when removing copies is not enough? The remaining choices affect both how much of each group of related examples is retained and which members represent it. This question matters when a finite pool is reused over multiple training passes, a regime studied in data-constrained language modeling (Hernandez et al., 2022; Muennighoff et al., 2023). FineWeb and SemDeDup demonstrate the importance of deduplication and representative selection (Penedo et al., 2024; Abbas et al., 2023); budgeted pruning also requires a signal for the value of nonidentical examples.

Loss and semantic similarity offer different information. A probe's loss measures how well it predicts an example, but does not describe the example's relation to the other candidates. Semantic embeddings describe content, without directly measuring the pretraining model's parameter updates. Difficulty-based pruning can sacrifice coverage (Zheng et al., 2023), and factual knowledge depends on exposure to related documents (Kandpal et al., 2023). This motivates measuring redundancy in the model's update space: whether an example's gradient direction is already represented by the current candidates.

Data2Delta maps data to parameter-update directions for budgeted selection. It clusters normalized gradient sketches from a fixed probe, computes within-cluster ridge leverage, and removes the lowest-scoring examples in one global ranking. The geometric intuition is explicit: a cluster's mean leverage equals its effective dimension divided by its size (Proposition 1). Many examples spanning few effective directions therefore have a low average score. A single ranking sets both deletion counts and membership, without prescribing a quota for each cluster. Figure 1 illustrates the computation and its single-deletion geometry.

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

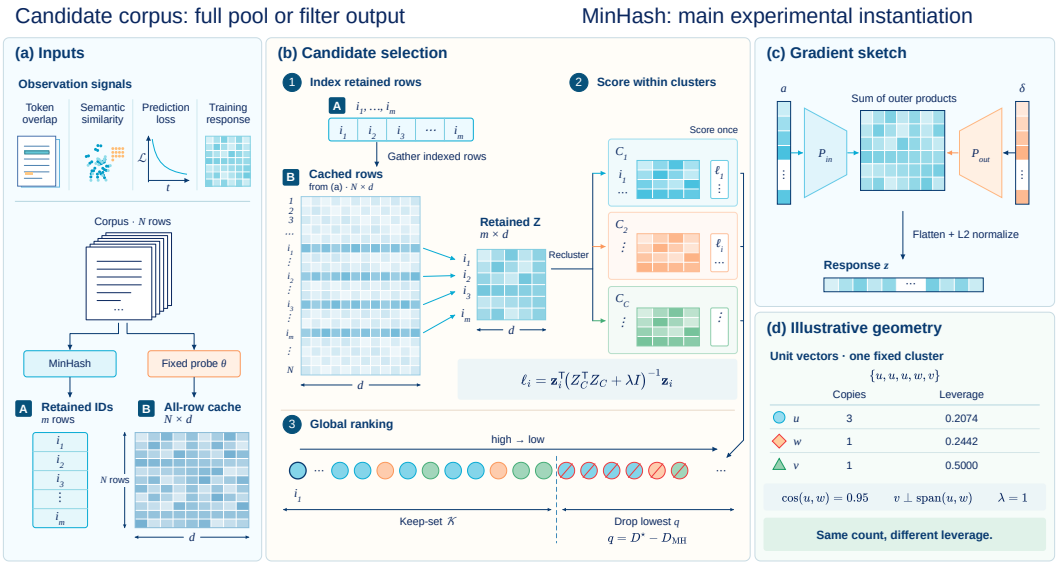


Figure 1: **Data2Delta for budgeted corpus selection.** (a) Candidates may be a full pool or filter output; the main experiment uses MinHash IDs to index a fixed probe's sketch cache. (b) Candidate sketches are clustered, scored once within clusters, and globally ranked to retain $|\mathcal{C}| - q$ rows. (c) Projected backward signals and activations form tokenwise outer products, summed before flattening and normalization. (d) Illustrative unit vectors $\{u, u, u, w, v\}$ share one cluster, with $u^\top w = 0.95$, $v \perp \mathrm{span}(u, w)$, and $\lambda = 1$. Panel (d) uses constructed vectors to illustrate the rule.

We test representation, selection, and evaluation separately. In controlled Pythia-1B continuation, gradients improve PPL over dimension-matched BGE-M3 under the same selector, and leverage-ranked membership improves PPL over random membership at fixed cluster deletion counts. End to end, Data2Delta lowers PPL relative to uniform selection, with an interpolated token equivalent of about 19%. Low-NLL filtering achieves similar fit. We compare their selections through training-pool provenance and natural-news source coverage: on the latter, low NLL loses 1,060 families at 10% removal, versus 79 for Data2Delta and 60 for BGE. Lower PPL coexists with lower mean zero-shot accuracy. Our contributions connect a selection rule to controlled tests and behavioral diagnostics:

- **Selection through gradient geometry.** A global ridge-leverage ranking couples cluster allocation and member choice; its scores order single-deletion changes in regularized gradient volume.
- **Controlled representation and membership comparisons.** Gradients improve PPL over dimension-matched BGE features, and leverage-ranked membership improves PPL over random membership at fixed deletion counts per cluster.
- **A distinction between evaluation goals.** Similar held-out fit can accompany different source coverage, while lower PPL need not improve zero-shot accuracy. Same-pool provenance counts and natural duplicate labels make the coverage comparison explicit.

2 Data Selection through Gradient Geometry

2.1 Selection Problem

Let $X = \{x_i\}_{i=1}^N$ be a corpus and let $\mathcal{C} \subseteq \{1, \ldots, N\}$ index the candidate examples. Given a deletion budget q, Data2Delta returns retained IDs satisfying

$$\mathcal{K} \subseteq \mathcal{C}, \qquad |\mathcal{K}| = |\mathcal{C}| - q, \qquad 0 \le q \le |\mathcal{C}|. \tag{1}$$

Data2Delta uses candidate-relative gradient geometry as a proxy for selecting data that improves held-out fit. Within each group, it favors directions less represented by the other candidates. Candidates may be the full corpus or an upstream filter's output, and no near-duplicate labels are required; the main experiment uses a shared MinHash residual (Section 3).

2.2 Projected gradient representations

A fixed probe relates examples through local loss gradients. For probe parameters θ, let $\mathcal{L}(x_i;\theta)$ be mean cross-entropy over valid prediction positions. For an ordinary SGD update with sufficiently small step size η, full gradients satisfy the first-order approximation also used by TracIn (Pruthi et al., 2020):

$$\mathcal{L}(y;\theta-\eta\nabla\mathcal{L}(x;\theta))\approx\mathcal{L}(y;\theta)-\eta\nabla\mathcal{L}(y;\theta)^\top\nabla\mathcal{L}(x;\theta). \tag{2}$$

Training on x thus changes the loss on y through their gradient inner product. This motivates using coverage of the direction of $\nabla\mathcal{L}(x;\theta)$ among other candidates as a selection proxy. For a parameter block $\mathcal{B}$ and a fixed linear projection Π, define

$$g_i=\nabla_{\theta_\mathcal{B}}\mathcal{L}(x_i;\theta),\qquad z_i=\frac{\Pi g_i}{\max(\|\Pi g_i\|_2,10^{-12})}. \tag{3}$$

Row normalization keeps the update direction and discards the gradient norm, which is itself used as a difficulty score (Paul et al., 2021); the score below measures directional redundancy, while it may still correlate with loss. Data2Delta uses these normalized sketches as geometric descriptors at the fixed probe.

We use the layer-8 MLP expansion weight of Pythia-1B, with zero-based layer indexing. Its gradient is a matrix $G_i\in\mathbb{R}^{8192\times2048}$. Two Rademacher projections produce

$$S_i=P_{\text{out}}G_iP_{\text{in}}^\top,\quad P_{\text{out}}\in\mathbb{R}^{91\times8192},\quad P_{\text{in}}\in\mathbb{R}^{90\times2048}. \tag{4}$$

Flattening S_i yields 8,190 features. In $G_i=\sum_t\delta_t a_t^\top$, a_t denotes the layer's input activation and δ_t the derivative of the example's mean loss with respect to the layer output at position t. We accumulate $(P_{\text{out}}\delta_t)(P_{\text{in}}a_t)^\top$ without materializing the full per-example gradient, as in structured projections such as LoGra (Choe et al., 2024). We use the first 256 tokens of each example and apply no Adam preconditioning.

2.3 Within-cluster ridge leverage

For cluster c with n_c rows, stack unit representations in Z_c and let $M_c=Z_c^\top Z_c$. We score example i by

$$\ell_i=z_i^\top(M_c+\lambda I)^{-1}z_i,\qquad\lambda>0. \tag{5}$$

If $M_c=V\,\mathrm{diag}(\mu_1,\dots,\mu_d)V^\top$, then

$$\ell_i=\sum_{j=1}^{d}\frac{(v_j^\top z_i)^2}{\mu_j+\lambda}. \tag{6}$$

Directions with more second-moment energy receive less weight. In a cluster of r identical unit vectors, every copy scores $1/(r+\lambda)$. When a cluster has fewer rows than sketch coordinates, a score also reflects how much of z_i lies outside the span of the other rows (Appendix M).

Proposition 1. *Fix the representations and cluster assignments, and let $\Phi=\sum_c\log\det(M_c+\lambda I)$. (i) Deleting candidate i alone decreases Φ by $-\log(1-\ell_i)$, which increases with ℓ_i. (ii) Within cluster c, $\sum_{i\in c}\ell_i=d_\lambda(c)=\sum_j\mu_j/(\mu_j+\lambda)$, so the cluster's mean score is $d_\lambda(c)/n_c$.*

Part (i) follows from the matrix determinant lemma and part (ii) from a trace identity (Appendix M). The global ranking thus deletes first the rows whose removal alone costs the least regularized gradient volume, a standard diversity objective (Kulesza & Taskar, 2012), while the mean-score identity links the ranking to effective dimension per member; actual cluster deletion counts depend on the full score distribution. Because these decreases are not additive, one-shot deletion with initial scores has neither batch optimality nor the guarantees of randomized leverage sampling (Cohen et al., 2016). Data2Delta performs unconstrained budgeted pruning (Algorithm 1), so a source family or an entire cluster may be removed.

Algorithm 1 Data2Delta selection

Input: Cached sketches with unique example IDs; candidate IDs $\mathcal{C}$; removal budget q; clustering configuration; regularization $\lambda > 0$.
Output: Retained candidate IDs $\mathcal{K}$.
1: Gather sketches for $\mathcal{C}$ and normalize each row.
2: Cluster candidate rows with spherical k-means using the fixed configuration.
3: Compute each row's leverage score once within its cluster using Eq. (5).
4: Rank all candidate rows by increasing score, using a fixed tie rule.
5: **return** $\mathcal{K} = \mathcal{C} \setminus \{\text{first } q \text{ ranked IDs}\}$.

3 Experimental design

Evaluation protocol. Training tests use paired endpoint NLL with three replicates per contrast and Holm correction within prespecified comparison families. Curves, source-family counts, and the eight-task zero-shot average (Avg8) are descriptive unless a test is stated. We distinguish same-pool provenance coverage from NEWS-COPY selection without training; effective-token savings are interpolated within saved curves, not net compute savings. Appendix A gives the full statistical protocol.

Candidate corpus. We sample 600,000 original 512-token chunks without replacement from the Pythia preshuffled Pile stream (Gao et al., 2020; Biderman et al., 2023), from steps after the step-5,000 probe and continuation checkpoint (Appendix F). Distinct stream positions can still share near-duplicate content.

The main pool adds 59,814 perturbed copies (termed descendants in the provenance records), each with a unique original parent: 29,898 *word-shuffled* and 29,916 *template-wrapped* examples, giving 659,814 candidates. The two types stand in for scrambled, spam-like text and boilerplate-wrapped reposts. Word shuffling reorders roughly ten-word fragments, swaps adjacent words, and deletes words with probability about 35%; it does not preserve semantics. Template wrapping keeps the original verbatim and adds navigation, copyright, advertisement, or related boilerplate. A historical five-type pool adds markup insertion, character noise, and contiguous excerpts. Original and descendant identities provide provenance, not quality labels.

Main comparison: common candidate pool and budget. The first-stage filter uses 64-dimensional unigram MinHash signatures (Broder, 1997) and an estimated Jaccard threshold of 0.8. It removes 34,992 examples, leaving a common residual of 624,822, including 43,208 descendants. Each second-stage rule removes 37,901 more examples. All final sets contain 586,921 examples, for a total removal rate of 11.05%. Grouping each original with its injected descendants into a source family, the residual contains 26,723 extra members beyond one per represented family. At least 11,178 second-stage deletions therefore cannot reduce those extra members. The budget therefore requires pruning beyond source-copy removal.

MinHash retains IDs $\mathcal{K}_{\mathrm{MH}}$ and removes $D_{\mathrm{MH}} = N - |\mathcal{K}_{\mathrm{MH}}|$ examples; we set $\mathcal{C} = \mathcal{K}_{\mathrm{MH}}$ and, for a target total removal $D_\star$, use the second-stage budget $q = D_\star - D_{\mathrm{MH}}$. Table 1 compares loss-based and gradient-geometry alternatives at the same q; BGE and activation substitutions test representations within the same selector.

Selection configuration and allocation controls. The gradient variants share projection seed 0 and row normalization. Spherical k-means uses $K = 512$, seed 0, a 50,000-example fitting sample, and 25 iterations; ridge scoring uses $\lambda = 0.01$ and float64 arithmetic. We add independent uniform selection and a control preserving Data2Delta's deletion count in every cluster while selecting deletions uniformly within each cluster. Both use fixed masks across training seeds (Appendix B).

Training and statistical protocol. Each set is used to continue Pythia-1B from step 5,000 with a constant learning rate of 10^{-5}, AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, zero weight decay, and sequence length 512. Texts are joined with EOS, with cross-document attention enabled. Training and evaluation read 513-token windows, using the first 512 tokens as inputs and the last 512 as targets in an explicit cross-entropy loss. Every update consumes 1,048,576 effective prediction tokens; the

Table 1: **Selection baselines and geometry ablations on the common residual.** "Low NLL" retains low-loss examples. The gradient variants share the same cached sketches.

Rule	Selection decision
Selection baselines	
Shared-rank random	Remove the final q examples in the fixed rank that also breaks first-stage ties, i.e., what the pipeline does without a second-stage scorer.
Low NLL	Remove the q examples with the highest probe NLL.
Lexical + low NLL	Visit TF-IDF candidate edges and retain the lower-NLL endpoint.
Geometry ablations and main method	
Frozen scores	Rank residual examples by leverage computed with full-pool clusters and moments.
Fixed clusters	Keep full-pool cluster assignments and recompute moments using residual examples.
Data2Delta (main)	Recluster residual sketches and recompute leverage before global selection.

endpoint is about 2.5B tokens, or 8.3 passes over a retained set (exact counts in Appendix B). Held-out evaluation uses 2,000 sequences starting at stream step 12,000, totaling 1,024,000 prediction tokens. The initial checkpoint has PPL 13.656.

Seeds 300–302 and 400–402 form separate training-order groups on six fixed selected sets (discovery and confirmation in the records); representation and allocation comparisons use seeds 300–302. Tests are two-sided paired t tests, with three pairs per contrast. PPL intervals are pointwise paired 95% t intervals. The two seed groups retain separate comparison families and are not pooled (Appendices A–B).

Representation substitution. The primary representation test compares 1,024-dimensional gradients with newly encoded BGE-M3 features (Chen et al., 2024), using an audited text prefix based on the probe's first 256 tokens; normalization, clustering, ridge regularization, and budget are shared. Supplementary comparisons replace the 8,190-dimensional gradients with cached 1,024-dimensional BGE embeddings or same-probe layer-8 activations averaged over the 256-token window (Appendices A–B). A 1B-token original-only study rebuilds the residual without injected descendants and runs four arms with three paired selection/training replicates (seeds 500–502; Appendix B.2).

Natural duplicates without a first-stage filter. We apply selection directly to the validation and test splits of a fixed NEWS-COPY evaluation mirror (Silcock et al., 2023). The splits contain 4,988 and 14,211 texts in 2,609 and 9,282 labeled source families. Data2Delta, BGE-leverage, low NLL, and random deletion each remove 5%, 10%, or 20% of texts, with the same prespecified K and λ for the geometric methods. All 24 selections are descriptive, with one mask per configuration and no NEWS-COPY training (Appendix C).

4 What Gradient Geometry Contributes

4.1 Training Outcomes under Matched Budgets

Against independent uniform selection, the main Data2Delta configuration lowers PPL by 0.0154, or about 0.12%, and reaches uniform's endpoint with 19.0–19.6% fewer effective prediction tokens across seeds (Figure 2). Low NLL saves a similar 18.7–19.1%. The curves show why a small vertical PPL gap corresponds to a larger horizontal token difference: improvement per token is slow near the endpoint. Token equivalents use interpolation within measured curves and exclude selection cost.

The advantage over the shared-rank random control is also observed in another three training seeds: PPL differences are -0.0157 for seeds 300–302 and -0.0148 for seeds 400–402 (the latter Holm $p = 0.0193$). Every saved checkpoint of each paired trajectory favors Data2Delta over that control in both groups. Differences from low NLL, lexical-plus-NLL, frozen scores, and fixed clusters are not significant after correction in the second group (Appendix A).

Table 2: **Language-modeling fit and source coverage.** Descriptive summary; tests are in Table 3 and Appendix A. Perturbed-pool columns use seeds 300–302: ΔPPL is the same-seed endpoint difference from independent uniform selection, and tokens saved is the budget share saved in reaching uniform's endpoint PPL (interpolated; n.r.: not reached by every seed; Table 11). Families retained is from NEWS-COPY test at 10% removal, without training (random: fixed random rank). Feature times are standalone extraction microbenchmarks (Table 20), not full-configuration timings; the 1,024-d gradient entry excludes its additional projection.

	Perturbed-pool continuation			NEWS-COPY	
Selector	ΔPPL ($\times 10^{-3}$)	Tokens saved (%)	Avg8 (%)	Families retained (%)	Feature time (s)
Uniform random	0 (ref.)	—	44.27	91.34	—
Low NLL	−15.07	18.9	44.22	88.58	4.50
BGE-leverage, 1,024-d	−0.62	n.r.	44.34	99.35	3.97
Data2Delta, 1,024-d	−13.74	17.6	44.26	—	9.86
Data2Delta, 8,190-d	−15.41	19.4	44.21	99.15	9.86

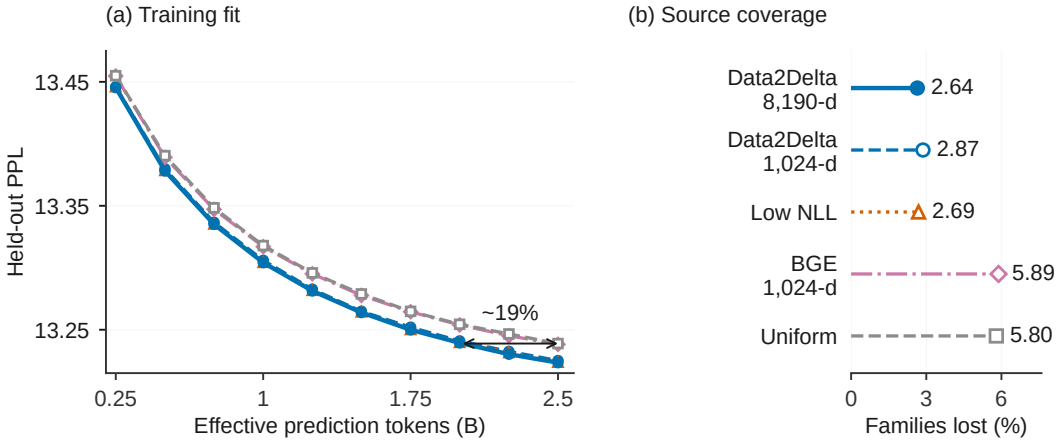


Figure 2: **Training curves and coverage on the same pool.** (a) Held-out PPL over effective prediction tokens, with means and sample SD over training seeds 300–302. Markers are saved evaluations; lines connect them. The arrow marks Data2Delta's interpolated token equivalent at uniform's endpoint (about 19%; Table 11). (b) Fully lost provenance families after second-stage selection on this same residual pool, using one fixed mask per rule. Coverage records survival of an original or any injected copy, irrespective of text quality. Full two-cohort PPL intervals are in Appendix A.

4.2 What the representation contributes

With dimension, clustering, regularization, and deletion budget held fixed, gradient-1024 gives endpoint PPL 13.225 versus 13.238 for freshly encoded BGE-1024. The paired difference is -0.0131 (95% interval $[-0.0147, -0.0116]$; Holm-adjusted NLL $p = 0.00148$). This is the primary representation comparison: the gradient representation yields lower endpoint PPL under the same selection rule. The saved curves imply 16.8% fewer effective prediction tokens to reach BGE's endpoint PPL (Table 11). The comparison substitutes the complete feature construction, with encoder and tokenization differences described in Appendix B.

Figure 3 also reports the 8,190-dimensional comparisons with cached BGE and same-probe activations; full values are in Appendix A.

4.3 What member selection contributes

Holding Data2Delta's deletion count fixed in every cluster and choosing deletions uniformly raises PPL by 0.00168 (Holm-adjusted NLL $p = 0.00229$; Table 3). The sets share 93.9% of deletions,

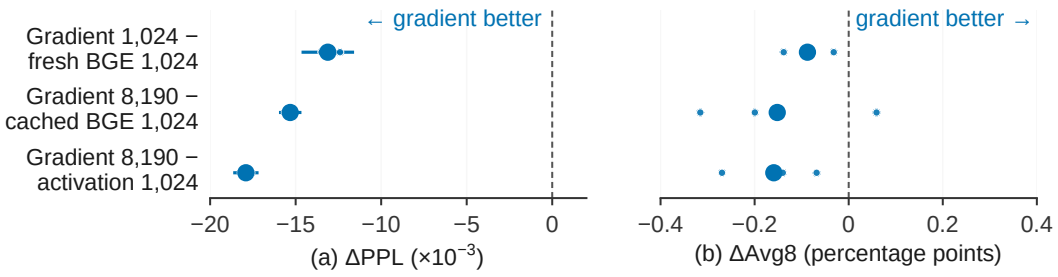


Figure 3: **Representation substitution within the same selector.** Each row is gradient minus its named comparator for (a) PPL and (b) Avg8 (percentage points), using seeds 300–302. Large points show means and small points the three paired differences; PPL lines are pointwise paired 95% t intervals, while Avg8 is descriptive. Arrows mark the direction that favors gradients: negative for PPL, positive for Avg8. Low NLL is included in Figure 2 and Appendix A.

Table 3: **Three controlled comparisons at 2.5B effective prediction tokens.** Main is 8,190-dimensional Data2Delta. Differences (first minus second) and pointwise paired 95% t intervals are in 10^{-3} PPL, using seeds 300–302 with main trajectories reused. Endpoint-NLL p-values are Holm-adjusted across the three contrasts.

Comparison	ΔPPL	95% interval	NLL Holm p
Main − independent uniform	−15.41	$[-16.19, -14.63]$	4.14×10^{-4}
Main − cluster-budget random	−1.68	$[-2.03, -1.34]$	0.00229
Gradient 1,024-d − BGE 1,024-d	−13.12	$[-14.66, -11.58]$	0.00148

largely from four entirely removed clusters. With allocation held fixed, this contrast estimates the effect of member choice at Data2Delta's inherited budgets (Appendix B).

4.4 Selection behavior beyond held-out fit

Coverage on the training pool. To connect held-out fit to the selected data, we first count provenance families on the same training pool. Each original and its injected copies form a family, and coverage records whether any member survives. At the same second-stage budget, Data2Delta loses 15,807 families, low NLL 16,100, independent uniform 34,670, and BGE-1024 35,202 (Figure 2b; Table 23). Thus Data2Delta and low NLL are close in both PPL and provenance coverage here, while retaining more source families than uniform or BGE. Every family, including each singleton original, receives equal weight.

Coverage on naturally duplicated news. We next evaluate natural source families on NEWS-COPY, using its duplicate labels and selecting directly from each split. If L of F labeled families lose every text, coverage is $(F - L)/F$. Of D deletions, $D - L$ remove extra family members; the two quantities are linked at a fixed budget (Appendix E).

At 10% test removal, Data2Delta retains 99.15% of source families, compared with 88.58% for low NLL and 99.35% for BGE-leverage. Low NLL loses 1,060 families, including 1,052 singletons, versus 79 for Data2Delta—over a 13-fold difference in lost families. Across both splits and all tested budgets, Data2Delta preserves more families than low NLL, and BGE achieves the highest coverage (Figure 4). The different ordering on NEWS-COPY, evaluated without training, shows how selection behavior varies across corpora.

4.5 Scope and allocation diagnostics

Downstream accuracy. At the shared initialization, Avg8 is 43.777%; Data2Delta reaches 44.160% in the second seed group, but remains 0.107–0.186 percentage points below the five controls. Thus continuation improves its mean accuracy from initialization, while its PPL advantage

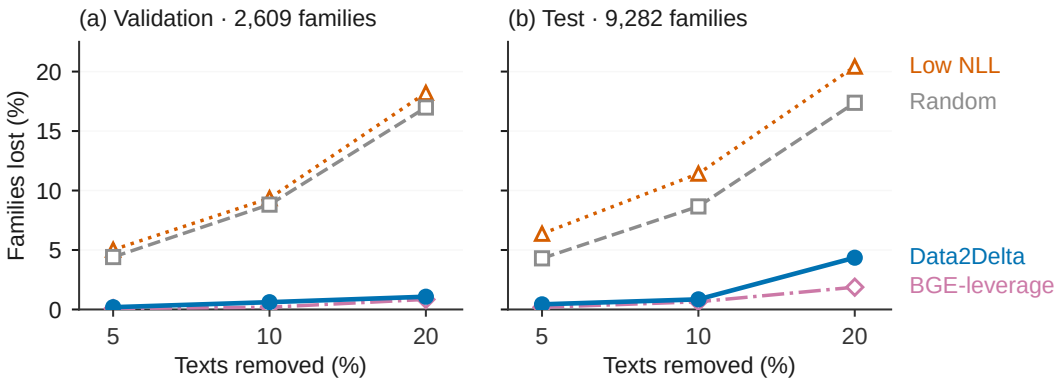


Figure 4: **Source coverage on natural news duplicates.** Fully lost families (% of all labeled families) in (a) validation and (b) test; lower is better. Each marker is one fixed selection at a prespecified budget; lines guide the eye, with no seed intervals. Table 14 gives counts for 10% test removal; Appendix C reports all budgets.

over the random control does not yield a relative Avg8 advantage. Fresh BGE-1024 also has 0.088 points higher Avg8 than gradient-1024. Taskwise differences favor Data2Delta on HellaSwag but favor all five controls on ARC-Challenge (Appendix A); these outcomes are exploratory.

Original-only continuation. Without injected copies, all four selectors' mean PPLs lie within 0.00392 at 1B effective prediction tokens. Data2Delta's differences are -0.00391 versus low NLL, -0.00194 versus BGE, and -0.00050 versus independent uniform; none passes the prespecified Holm correction over three paired selection/training replicates. The effect relative to random is substantially smaller in this original-only setting. This study compares pruned sets; the absolute effect relative to an unpruned arm remains unmeasured (Appendix B.2).

Concentrated allocation. Five of 512 clusters receive 92.1% of the main method's deletions. Four are deleted entirely: three contain 82–91% word-shuffled examples, and one contains 5,487 template-wrapped copies. Across the model-informed rules, the common PPL gain over random coincides with deleting many more word-shuffled copies. Gradient-1024 deletes 20,819 such copies, compared with 1,523 for BGE-1024 (Table 24). These composition differences associate the fit gains with corruption removal.

The fifth cluster is a stress case for unconstrained allocation: Data2Delta deletes 4,873 of its 4,877 members, including textually diverse originals, whereas low NLL and the fixed-geometry variants delete at most 98. A sampled text audit found no common textual attribute (Appendix D.3). Residual reclustering changes allocation substantially without a detectable PPL benefit over those variants. Here, low leverage coexists with textual diversity, illustrating the distinction between gradient redundancy and text-level interchangeability. Per-cluster caps and iterative rescoring are natural untested ways to constrain this behavior.

Interventions reinforce the distinction between shared structure and family identity. Removing a copy's own original changes its deletion rate by 0.51 (template) and 0.21 (word-shuffled) percentage points. A shared visible template on unrelated originals raises residual deletion from 1.75% to 30.97%, while suffixes outside the 256-token window leave the mask unchanged (Appendix D). Shared-graph diagnostics also separate injected recall from extra-member removal (Appendix I).

5 Related Work

Deduplication and representative selection. SemDeDup prunes near neighbors within embedding clusters; D4 combines deduplication and diversification (Abbas et al., 2023; Tirumala et al., 2023). Prototype and density-based methods likewise use geometry to guide pruning (Sorscher et al., 2022; Kamal et al., 2024). Coverage-centric and diversity–difficulty approaches address the limitations of difficulty-only selection (Zheng et al., 2023; Maharana et al., 2024). Data2Delta uses

gradient geometry and a global ridge-leverage ranking instead of a semantic-neighbor threshold or explicit cluster quota.

Model-aware signals. CRAIG and GRAD-MATCH approximate full or validation gradients with weighted subsets (Mirzasoleiman et al., 2020; Killamsetty et al., 2021). LESS and TRAK use projected gradients for targeted selection or attribution, and LoGra provides efficient factored projections (Xia et al., 2024; Park et al., 2023; Choe et al., 2024). Loss-based pruning and model-aware pretraining selection offer single-example or learned value signals (Marion et al., 2023; Ankner et al., 2025; Yu et al., 2024; Engstrom et al., 2024). Data2Delta scores directions relative to candidate clusters, without a target set; our representation substitutions test this choice under the same selector.

Set geometry and allocation. DPP objectives and greedy feature-spectrum selection connect diversity to set geometry (Kulesza & Taskar, 2012; Chen et al., 2018; Fan et al., 2025). Quad combines influence-based quality with cluster-level diversity (Zhang et al., 2025). Data2Delta's score has a single-deletion log-determinant interpretation and a cluster-average effective-dimension identity. Its one-shot global ranking is distinct from greedy rescoring, which remains an untested alternative here. Appendix N gives the broader context, including data mixtures, feature construction, and repeated-data training.

6 Discussion and Limitations

The central finding is that candidate-relative gradient geometry provides a useful signal for budgeted pretraining selection. The matched-selector comparison supports the gradient representation over semantic features, and the fixed-budget control supports leverage-ranked member choice. Together, these tests support the representation and membership components of Data2Delta in the evaluated continuation setting.

Low NLL is both a strong training baseline and a diagnostic contrast: similar held-out PPL accompanies different source-retention behavior on natural news. The smaller original-only effect, lower mean Avg8, and concentrated deletion of diverse text identify distinct limits of the measured fit gains. They motivate evaluating retained sources and downstream accuracy alongside PPL.

Our evidence covers Pythia-1B continuation from one checkpoint, about eight passes over each selected set, and three replicates per comparison. Main-pool gains coincide with removing synthetic corruptions; original-only differences are smaller and nonsignificant. Main-pool training uses fixed masks, whereas original-only replicates jointly vary selection and training order. Priorities are same-gradient-space nearest-neighbor and greedy diversity baselines, broader model/checkpoint tests, and joint training and coverage evaluation on natural corpora.

Computational scope. Warm extraction takes 9.86 seconds per 1,024 texts for gradients on one A800, versus 4.50 for NLL and 3.97/4.89 for BGE-M3 at two maximum-window settings. These exclude loading, clustering, ridge scoring, I/O, and training; about 2.7 gradient-extraction GPU-hours per million 256-token windows must be included in any net-compute comparison (Appendix D).

7 Conclusion

Data2Delta turns gradient geometry into a budgeted selection rule, linking candidate-relative scores to effective dimension and using one ranking for allocation and membership. Controlled comparisons show PPL gains over semantic features, uniform selection, and random membership at fixed cluster budgets. Source coverage and downstream results distinguish these fit gains from broader data-value judgments. Gradient geometry offers a candidate-relative notion of redundancy for pretraining selection, grounded in how each example's direction is represented by the other candidates.

AI USE STATEMENT

The authors developed the research idea and experimental design. Generative AI tools assisted with experimental implementation, result analysis, figure preparation, literature search, and manuscript drafting, translation, and editing. The authors take responsibility for the final manuscript and research artifacts.

REPRODUCIBILITY STATEMENT

Section 2 specifies the selection rule for a supplied candidate set; Section 3 gives its main MinHash-residual instantiation, data, training, and statistical protocol. Appendix F records implementation details and cost definitions. Appendices A–B document 66 distinct continuation trajectories across the main, representation/allocation, and original-only studies, with cohort-specific replication and separate statistical comparison families. Each planned paired contrast uses three replicates; reused main trajectories are counted once. Token-matching and cluster-allocation summaries (Tables 11 and 19) are derived from saved curves and masks without new runs. Appendix C records 24 natural-data masks, Appendix G retains the discovery diagnostics, and Appendix M gives the derivations. The archived manuscript workspace contains the figures, bibliography, frozen plotting summaries, and provenance records. Selection code, fixed masks, and analysis scripts will be released with the paper. Historical full-pool protocols and their reproducibility scope are documented separately in Appendices F and K.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication, 2023. URL https://arxiv.org/abs/2303.09540.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max M Marion, Matthew Leavitt, and Mansheej Paul. Perplexed by Perplexity: Perplexity-Based Data Pruning With Small Reference Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 100615–100634, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fa30825df7ceaba452d5533538ea29c2-Paper-Conference.pdf.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

A.Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, SEQUEN-97, pp. 21–29. IEEE Comput. Soc, 1997. doi: 10.1109/sequen.1997.666900. URL http://dx.doi.org/10.1109/SEQUEN.1997.666900.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast Greedy MAP Inference for Determinantal Point Process to Improve Recommendation Diversity. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,

2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/dbbf603ff0e99629dda5d75b6f75f966-Paper.pdf.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions, 2024. URL https://arxiv.org/abs/2405.13954.

Michael B. Cohen, Cameron Musco, and Christopher Musco. Input Sparsity Time Low-Rank Approximation via Ridge Leverage Score Sampling, 2016. URL https://arxiv.org/abs/1511.07263.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: Model-Aware Dataset Selection with Datamodels, 2024. URL https://arxiv.org/abs/2401.12926.

Ziqing Fan, Siyuan Du, Shengchao Hu, Pingjie Wang, Li Shen, Ya Zhang, Dacheng Tao, and Yanfeng Wang. Combatting Dimensional Collapse in LLM Pre-Training Data via Submodular File Selection. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 6655–6679, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/136729ae4b0fee25a0d28077442506da-Paper-Conference.pdf.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling Laws and Interpretability of Learning from Repeated Data, 2022. URL https://arxiv.org/abs/2205.10487.

Amro Kamal, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 52013–52035, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e4bc9cdebfbcc43b0800cf99ffec5ee1-Paper-Conference.pdf.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large Language Models Struggle to Learn Long-Tail Knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Joshua Kazdan, Noam Levi, Rylan Schaeffer, Jessica Chudnovsky, Abhay Puri, Bo He, Mehmet Donmez, Sanmi Koyejo, and David Donoho. Scale Dependent Data Duplication, 2026. URL https://arxiv.org/abs/2603.06603.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRADMATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Alex Kulesza and Ben Taskar. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, December 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL http://dx.doi.org/10.1561/2200000044.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In search of the next generation of training sets for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 14200–14282. Curran Associates, Inc., 2024. doi: 10.52202/079017-0455. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/19e4ea30dded58259665db375885e412-Paper-Datasets_and_Benchmarks_Track.pdf.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. $\mathbb{D}^2$ Pruning: Message Passing for Balancing Diversity & Difficulty in Data Pruning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 55740–55759, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f646a4c970799f789e031b63161d018d-Paper-Conference.pdf.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale, 2023. URL https://arxiv.org/abs/2309.04564.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for Data-efficient Training of Machine Learning Models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960. PMLR, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling Data-Constrained Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376. Curran Associates, Inc., 2023. doi: 10.52202/075280-2191. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf.

Nihal V. Nayak, Paula Rodriguez-Diaz, Neha Hulkund, Sara Beery, and David Alvarez-Melis. A Critical Look at Targeted Instruction Selection: Disentangling What Matters (and What Doesn't), 2026. URL https://arxiv.org/abs/2602.14696.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ac56f8fe9eea3e4a365f29f0f1957c55-Paper.pdf.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.

Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, and Melissa Dell. Noise-Robust De-Duplication at Scale. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bAz2DBS35i.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19523–19536. Curran Associates, Inc., 2022. doi: 10.52202/068431-1419. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7b75da9b61eda40fa35453ee5d077df6-Paper-Conference.pdf.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving LLM Pretraining via Document De-Duplication and Diversification. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53983–53995. Curran Associates, Inc., 2023. doi: 10.52202/075280-2348. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a8f8cbd7f7a5fb2c837e578c75e5b615-Paper-Datasets_and_Benchmarks.pdf.

Hai Wang, Chenhao Wang, Qifeng Cai, Yixiu Liu, Miao Peng, Nuo Chen, Yuanlin Tu, Chengcheng Xu, and Feng Zhang. Scalable Frequency- and Length-Aware Subdocument Deduplication for Large Language Model Pretraining, 2026. URL https://arxiv.org/abs/2608.03089.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. QuRating: Selecting High-Quality Data for Training Language Models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52915–52971. PMLR, 2024. URL https://proceedings.mlr.press/v235/wettig24a.html.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. In A. Oh, T. Naumann, A. Globerson,

K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818. Curran Associates, Inc., 2023a. doi: 10.52202/075280-3059. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dcba6be91359358c2355cd920da3fcbd-Paper-Conference.pdf.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data Selection for Language Models via Importance Resampling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34201–34227. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1482. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6b9aa8f418bde2840d5f4ab7a02f663b-Paper-Conference.pdf.

Zichun Yu, Spandan Das, and Chenyan Xiong. MATES: Model-Aware Data Selection for Efficient Pretraining with Data Influence Models, 2024. URL https://arxiv.org/abs/2406.06046.

Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, Ye Yuan, Guoren Wang, and Conghui He. Harnessing Diversity for Important Data Selection in Pretraining Large Language Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 72980–73003, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b588d9b67932b459ea66ff6e2804c6b3-Paper-Conference.pdf.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric Coreset Selection for High Pruning Rates. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QwKvL6wC8Yi.

A Complete training outcomes and statistical protocol

A.1 Evidence inventory and analysis units

The study contains 66 distinct continuation trajectories and 66 endpoint Avg8 evaluations: 54 on the perturbed residual and 12 on original-only data. The six-rule comparison uses two groups of three training seeds (18 runs per group); BGE and activation substitutions add six runs, and independent-random, cluster-budget, and dimension-matched controls add twelve. The original-only comparison has four rules with three replicates each. Shared Data2Delta trajectories are counted once. Every planned paired contrast uses three replicates; neither checkpoints nor downstream tasks increase that number.

The perturbed-pool groups share initialization, training configuration, evaluation sample, and effective prediction-token counts. Training seeds vary order on fixed selected data. Seeds 300–302 are labeled discovery and seeds 400–402 confirmation in the records; representation and allocation comparisons reuse the first group and are not independent confirmations. The original-only replicates jointly vary selection and training order. Per-run values, hashes, and configuration checks are retained in the reproducibility bundle. Appendix B describes representation/allocation controls and the original-only study; Appendix G gives composition and trajectory diagnostics.

The confirmation differences from low NLL and fixed-geometry variants remain unresolved; these runs do not isolate a benefit of reclustering.

Table 4: **All endpoint groups.** Mean ± sample SD, $n = 3$ in every cell. Extension Data2Delta is the discovery Data2Delta group, not three new runs.

Cohort	Rule	PPL	Avg8 (%)
Discovery	Shared-rank random	13.2391 ± 0.0001	44.2629 ± 0.0504
Discovery	Low NLL	13.2238 ± 0.0004	44.2225 ± 0.0716
Discovery	Lexical + low NLL	13.2245 ± 0.0001	44.2680 ± 0.0696
Discovery	Frozen scores	13.2242 ± 0.0001	44.2862 ± 0.0814
Discovery	Fixed clusters	13.2239 ± 0.0002	44.2288 ± 0.0508
Discovery	Data2Delta	13.2234 ± 0.0002	44.2081 ± 0.0258
Confirmation	Shared-rank random	13.2383 ± 0.0007	44.2730 ± 0.0838
Confirmation	Low NLL	13.2238 ± 0.0002	44.2864 ± 0.1037
Confirmation	Lexical + low NLL	13.2246 ± 0.0006	44.3459 ± 0.0997
Confirmation	Frozen scores	13.2238 ± 0.0003	44.3126 ± 0.0661
Confirmation	Fixed clusters	13.2237 ± 0.0004	44.2664 ± 0.0480
Confirmation	Data2Delta	13.2235 ± 0.0010	44.1597 ± 0.0765
Extension	BGE-leverage	13.2388 ± 0.0005	44.3600 ± 0.1728
Extension	Activation-leverage	13.2414 ± 0.0005	44.3673 ± 0.0910

A.2 Statistical protocol and complete comparisons

Endpoint NLL is the prespecified inferential outcome. Within each cohort, Data2Delta is paired with a control using the same training seed, and a two-sided paired t test is applied. Holm adjustment covers five discovery tests, five confirmation tests, and two representation-extension tests separately. PPL intervals are pointwise 95% paired t intervals, not simultaneous intervals. Discovery and confirmation are not pooled to replace the separate confirmation result.

Each comparison has three pairs and two degrees of freedom. Such a sample does not permit a reliable check of normality of the paired differences. Exact two-sided sign-flip sensitivity checks have minimum attainable $p = 0.25$ with three pairs and do not reject any comparison at 0.05. These checks also require symmetry or exchangeability; they do not remove all assumptions. The manuscript's significance statements refer to the prespecified parametric test. A nonsignificant difference is not equivalence or noninferiority, for which no margin was specified.

The small discovery advantages over lexical-plus-NLL and frozen scores (Holm $p \approx 0.044$) are not confirmed by the second seed group. Table 5 lists both groups. Figure 5 displays both groups on a broken axis at a common scale; no seed groups are pooled.

Across all three seeds and ten checkpoints per seed, Data2Delta has lower PPL than shared-rank random in both cohorts and than BGE- and activation-leverage in the original representation extension; consistency against the other controls varies by comparison and cohort.

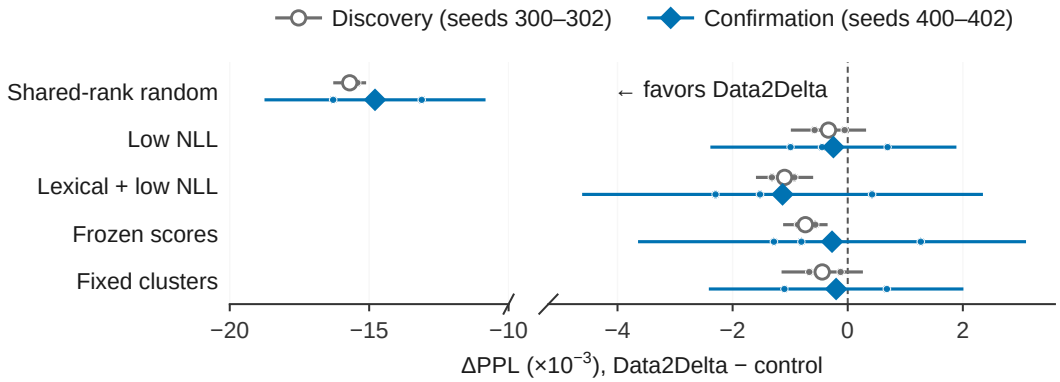


Figure 5: **Endpoint PPL comparisons across two seed groups.** Data2Delta minus control, in 10^{-3} PPL; negative values favor Data2Delta. Open gray circles denote seeds 300–302 and filled blue diamonds seeds 400–402. Lines show pointwise paired 95% t intervals, and small dots the three seed differences. The x-axis is broken between -10 and -5.2 at a common scale. Both groups use the same selected sets and separate five-comparison Holm families.

Table 5: **All paired training comparisons.** Differences are Data2Delta minus control. PPL intervals are pointwise paired 95% t intervals. Tests use endpoint NLL; Holm families have five, five, and two comparisons for discovery, confirmation, and extension.

Cohort	Control	ΔPPL	95% interval	NLL Holm p
Discovery	Shared-rank random	−0.01570	[−0.01629, −0.01511]	0.00038
Discovery	Low NLL	−0.00033	[−0.00099, +0.00032]	0.2286
Discovery	Lexical + low NLL	−0.00110	[−0.00159, −0.00060]	0.0436
Discovery	Frozen scores	−0.00074	[−0.00112, −0.00035]	0.0437
Discovery	Fixed clusters	−0.00044	[−0.00115, +0.00026]	0.2286
Confirmation	Shared-rank random	−0.01479	[−0.01875, −0.01083]	0.0193
Confirmation	Low NLL	−0.00025	[−0.00239, +0.00189]	1.0000
Confirmation	Lexical + low NLL	−0.00113	[−0.00461, +0.00235]	1.0000
Confirmation	Frozen scores	−0.00027	[−0.00364, +0.00310]	1.0000
Confirmation	Fixed clusters	−0.00020	[−0.00241, +0.00201]	1.0000
Extension	BGE-leverage	−0.01532	[−0.01597, −0.01466]	0.00018
Extension	Activation-leverage	−0.01791	[−0.01865, −0.01717]	0.00018

A.3 Downstream evaluation

At the common step-5,000 initialization, full-evaluation Avg8 is 43.7765%; the Data2Delta confirmation mean is 44.1597%, a descriptive increase of 0.3831 percentage points. The five control means range from 44.2664% to 44.3459%, leaving Data2Delta 0.1067–0.1862 points lower. This initialization-to-endpoint change describes continuation from a shared model, not a paired selection-method effect. In particular, the Data2Delta–shared-rank contrast is -0.1133 points, with pointwise 95% interval $[-0.184, -0.043]$ and exploratory Holm $p = 0.1017$.

Avg8 is the unweighted mean of eight zero-shot task metrics on a common accuracy scale, evaluated with lm-eval 0.4.12 and no example limit. ARC-Challenge (1,172 examples), OpenBookQA (500), and HellaSwag (10,042) use normalized accuracy. LAMBADA (5,153), ARC-Easy (2,376), PIQA (1,838), SciQ (1,000), and WinoGrande (1,267) use accuracy. Task hashes, metric names, and effective evaluation sample counts match across the original 42 endpoints reported in this section. Values in Table 6 are percentage-point differences. Their uncertainty covers the three training orders, not independently sampled tasks or questions.

The original evaluation summaries supply unadjusted Avg8 tests. We additionally report an exploratory Holm correction across the five, five, and two comparisons in each cohort. All 12 mean differences across the three cohorts are negative, although the comparisons share Data2Delta runs and are not independent evidence of degradation. The same-seed confirmation difference against the Shared-rank random control is -0.1133 percentage points; its pointwise interval excludes zero before multiple-comparison adjustment. These results do not establish preserved downstream ability.

Table 6: **Paired Avg8 comparisons.** Differences and pointwise 95% intervals are in percentage points; negative values are unfavorable to Data2Delta. Holm adjustments are post hoc, exploratory, and separate within each cohort.

Cohort	Control	ΔAvg8	95% interval	Raw p	Holm p
Discovery	Shared-rank random	-0.055	$[-0.241, +0.131]$	0.3320	1.0000
Discovery	Low NLL	-0.014	$[-0.167, +0.138]$	0.7224	1.0000
Discovery	Lexical + low NLL	-0.060	$[-0.257, +0.137]$	0.3202	1.0000
Discovery	Frozen scores	-0.078	$[-0.344, +0.188]$	0.3340	1.0000
Discovery	Fixed clusters	-0.021	$[-0.205, +0.163]$	0.6757	1.0000
Confirmation	Shared-rank random	-0.113	$[-0.184, -0.043]$	0.0203	0.1017
Confirmation	Low NLL	-0.127	$[-0.564, +0.311]$	0.3388	0.3388
Confirmation	Lexical + low NLL	-0.186	$[-0.476, +0.104]$	0.1099	0.3097
Confirmation	Frozen scores	-0.153	$[-0.383, +0.077]$	0.1032	0.3097
Confirmation	Fixed clusters	-0.107	$[-0.200, -0.013]$	0.0391	0.1563
Extension	BGE-leverage	-0.152	$[-0.628, +0.325]$	0.3037	0.3037
Extension	Activation-leverage	-0.159	$[-0.412, +0.094]$	0.1138	0.2275

Tables 7 and 8 separate the confirmation results by task. Relative to the Shared-rank random control, the largest negative mean differences occur on WinoGrande (-0.421 percentage points), SciQ (-0.233), and ARC-Challenge (-0.228). ARC-Easy, OpenBookQA, and HellaSwag have positive mean differences, and PIQA has zero mean difference. Thus the aggregate decrease does not describe a decrease on every task. ARC-Challenge is lower for Data2Delta against all five controls, whereas HellaSwag is higher against all five. These are exploratory taskwise descriptions, without per-task significance claims or a retrospectively chosen noninferiority margin.

Table 7: **Confirmation endpoint scores by task.** Percent accuracy, mean ± sample SD over training seeds 400–402 ($n = 3$), with fixed selected sets. Normalized accuracy is used for ARC-Challenge, OpenBookQA, and HellaSwag. Shared-rank random is not independent uniform residual sampling.

Task	Data2Delta	Shared-rank random	Low NLL	Lex.+NLL	Frozen	Fixed clusters
LAMBADA	36.99±0.06	37.14±0.11	37.00±0.02	37.07±0.12	37.07±0.09	37.02±0.13
ARC-Easy	45.41±0.15	45.38±0.11	45.33±0.08	45.45±0.07	45.55±0.06	45.50±0.08
ARC-Challenge	23.01±0.10	23.24±0.13	23.58±0.10	23.55±0.23	23.46±0.09	23.49±0.32
PIQA	62.62±0.14	62.62±0.14	62.57±0.05	62.88±0.19	62.79±0.20	62.69±0.23
SciQ	74.80±0.10	75.03±0.15	74.80±0.10	74.80±0.00	74.73±0.06	75.07±0.21
OpenBookQA	26.80±0.40	26.73±0.12	26.93±0.23	26.60±0.00	26.87±0.23	26.80±0.35
HellaSwag	31.89±0.05	31.87±0.07	31.86±0.07	31.86±0.01	31.81±0.09	31.84±0.04
WinoGrande	51.75±0.51	52.17±0.60	52.22±0.43	52.57±0.57	52.22±0.30	51.72±0.16
Avg8	44.16±0.08	44.27±0.08	44.29±0.10	44.35±0.10	44.31±0.07	44.27±0.05

Discovery Data2Delta has mean endpoint Avg8 44.2081% under the same full evaluation. The earlier two-example-per-task smoke evaluation remains excluded. Machine-readable task values and source hashes accompany the new control bundle (Appendix B).

A.4 Representation and restart details

The activation representation uses the input to layer 8's `dense_h_to_4h` linear module, the same weight block used for the gradient sketch. It averages valid prediction positions and applies a fixed random projection to 1,024 coordinates. Gradients and activations use the first 256 tokens with the same probe tokenizer. BGE-M3 yields 1,024 coordinates. Its residual cache predates this comparison and does not fully document the original input-window setting. Natural NEWS-COPY BGE

Table 8: **Paired confirmation differences by task.** Data2Delta minus each control, in percentage points (mean ± sample SD of three same-seed differences). Positive values favor Data2Delta. These exploratory taskwise summaries carry no per-task significance or noninferiority claim; the eight tasks are not independent training replicates.

Task	Shared-rank random	Low NLL	Lex.+NLL	Frozen	Fixed clusters
LAMBADA	-0.14 ± 0.10	-0.01 ± 0.04	-0.07 ± 0.09	-0.07 ± 0.11	-0.03 ± 0.19
ARC-Easy	$+0.03\pm0.06$	$+0.08\pm0.17$	-0.04 ± 0.07	-0.14 ± 0.17	-0.08 ± 0.17
ARC-Challenge	-0.23 ± 0.05	-0.57 ± 0.10	-0.54 ± 0.18	-0.46 ± 0.18	-0.48 ± 0.27
PIQA	$+0.00\pm0.05$	$+0.05\pm0.09$	-0.25 ± 0.33	-0.16 ± 0.11	-0.07 ± 0.36
SciQ	-0.23 ± 0.23	$+0.00\pm0.10$	$+0.00\pm0.10$	$+0.07\pm0.12$	-0.27 ± 0.12
OpenBookQA	$+0.07\pm0.31$	-0.13 ± 0.61	$+0.20\pm0.40$	-0.07 ± 0.46	$+0.00\pm0.53$
HellaSwag	$+0.02\pm0.06$	$+0.03\pm0.07$	$+0.03\pm0.06$	$+0.08\pm0.14$	$+0.05\pm0.07$
WinoGrande	-0.42 ± 0.28	-0.47 ± 0.93	-0.82 ± 0.73	-0.47 ± 0.24	$+0.03\pm0.43$
Avg8	-0.11 ± 0.03	-0.13 ± 0.18	-0.19 ± 0.12	-0.15 ± 0.09	-0.11 ± 0.04

features, in contrast, were newly extracted with a 256-token maximum using the BGE tokenizer. Gradient features have 8,190 coordinates. Shared K, λ, normalization, and selection budget therefore do not make these representations dimension- or information-matched. The new 1,024-dimensional gradient and freshly encoded BGE comparison matches feature dimension and records the input-prefix construction (Appendix B); the 256-dimensional variant remains a selection-only diagnostic.

Among the original 42 trajectories, four formal confirmation runs resumed from saved training state: lexical-plus-NLL seed 400 at update 477, Data2Delta seeds 400 and 401 at update 954, and lexical-plus-NLL seed 402 at update 954. The other 38 runs record no such restart. Seven nodes passed a ten-update, eight-rank bf16 continuation smoke test: model weights, AdamW state, random states, and sampler counters were restored exactly before continuation, and subsequent discrepancies satisfied the declared engineering tolerances. This checks the restoration implementation on the small smoke model; it is neither a production noise bound nor an uninterrupted same-seed control for the four formal trajectories. Their restart effects remain unestimated. All runs remain included, and the larger Data2Delta confirmation SD is not attributed to a single cause.

B Representation, allocation, and original-only comparisons

B.1 Comparison protocol and provenance

Twelve residual trajectories evaluate independent uniform sampling, a cluster-budget-matched random selection, 1,024-dimensional gradients, and freshly encoded 1,024-dimensional BGE-M3 features. Each arm uses one fixed selected set and training seeds 300–302. They share the original residual (624,822 examples), retention budget (586,921 examples), step-5,000 initialization, and endpoint of 2,499,805,184 effective prediction tokens. The main Data2Delta trajectories are reused from discovery; these controls share its seed group rather than supplying another independent replication.

Independent uniform sampling uses a fresh selection seed, separate from the MinHash tie-breaking rank. It removes 2,604 injected descendants, compared with the uniform expectation of approximately 2,621 and the shared-rank control's 1,786. The cluster-budget control preserves every main-method cluster's deletion count exactly and randomizes which rows are deleted within each cluster. Its within-cluster randomization seed is 1; this is not a new clustering seed. It shares 35,576 of 37,901 deletions with Data2Delta (93.87%). Four fully deleted clusters account for 30,030 deletions (79.23% of the budget), which remain fixed under this control (Appendix D.3). With cluster sizes n_c and deletion counts q_c, uniform sampling within each cluster gives expected overlap $\sum_c q_c^2/n_c = 35{,}594.76$, or 93.92%. This expectation concerns deletion overlap, not PPL. Its three training seeds vary order on this one mask.

The 1,024-dimensional gradients apply an additional Rademacher projection to the cached 8,190-dimensional sketches (NumPy random seed 0; entries $\pm1/\sqrt{1{,}024}$), followed by row ℓ_2 normalization. Fresh BGE extraction records the encoder, tokenizer, input pool, and per-row text-prefix policy. The source prefix is constructed from the gradient probe's first-256-token text window, with

character-boundary handling recorded in the extraction manifest; BGE encodes this supplied prefix without truncation. This makes input construction traceable, without equating token sequences from different tokenizers. Both representations have 1,024 coordinates and use row normalization, $K = 512$, $\lambda = 0.01$, the same clustering procedure, and the same deletion budget. Encoder and representation construction still differ.

The frozen plan declares three paired endpoint-NLL comparisons: main Data2Delta versus independent uniform sampling, main Data2Delta versus cluster-budget-matched random selection, and gradient-1024 versus fresh BGE-1024. Two-sided paired t tests use the three matched training seeds, with one Holm correction over this three-comparison family. PPL intervals are pointwise paired 95% t intervals. Checkpoints and Avg8 are descriptive secondary outcomes. These tests remain separate from the original five-, five-, and two-comparison families.

Table 9: **Residual extension endpoints.** Means $\pm$ sample SD over training-order seeds 300–302 ($n = 3$), at 2,499,805,184 effective prediction tokens on fixed selected sets. Data2Delta uses the original main trajectories; the four control arms are newly trained. Avg8 is descriptive. Cluster-budget random preserves the main method's deletion count in every cluster and randomizes membership within each cluster.

Rule	PPL ↓	Avg8 (%) ↑
Data2Delta (8,190-d)	13.22344 ± 0.00025	44.208 ± 0.026
Gradient (1,024-d)	13.22511 ± 0.00055	44.257 ± 0.044
Cluster-budget random	13.22513 ± 0.00036	44.240 ± 0.036
BGE-M3 (1,024-d)	13.23823 ± 0.00067	44.345 ± 0.019
Independent uniform	13.23885 ± 0.00037	44.266 ± 0.029

Table 10: **Complete predeclared Residual extension family.** Main denotes 8,190-dimensional Data2Delta. Differences and pointwise paired 95% intervals are in 10^{-3} PPL. Tests use endpoint NLL with three paired seeds ($df = 2$); Holm adjustment covers these three contrasts. Negative differences favor the first method. This extension reuses discovery seeds and does not constitute an independent confirmation.

Comparison	ΔPPL	95% interval	Raw p	Holm p
Main − independent uniform	-15.41	$[-16.19, -14.63]$	1.38×10^{-4}	4.14×10^{-4}
Main − cluster-budget random	-1.68	$[-2.03, -1.34]$	0.00229	0.00229
Gradient 1,024-d − BGE 1,024-d	-13.12	$[-14.66, -11.58]$	7.39×10^{-4}	0.00148

All three residual comparisons favor the gradient method in each seed and pass the prespecified adjusted NLL test. The cluster-budget control reaches PPL 13.22513, compared with 13.23885 for independent uniform sampling and 13.22344 for the full method. The full method has 0.00168 lower PPL than the budget-matched random control. This is a comparison of within-cluster choices on two substantially overlapping sets. The control inherits leverage-derived deletion counts, so these results do not independently estimate the training benefit of cluster allocation or support a causal percentage decomposition.

Avg8 is reported as a mean and sample SD over the three training orders. The three new mean paired contrasts are -0.0583, -0.0318, and -0.0877 percentage points, respectively, in the comparison order above. These descriptive outcomes do not show a downstream gain from the lower PPL. The complete initial checkpoint evaluation has Avg8 43.7765%, uses no task-example limit, and matches the endpoint metric definitions and evaluation counts.

Effective-token equivalents. Table 11 converts endpoint gaps into effective prediction tokens using the saved ten-point curves of the seed-300–302 arms, which share initialization, schedule, and evaluation sample. For each seed, it locates where a selector's curve first reaches the reference's 2.5B-token endpoint PPL, interpolating linearly between neighboring checkpoints. Data2Delta and low NLL save similar fractions relative to independent uniform selection, and fresh BGE-1024 does not reach uniform's endpoint in one seed. The discovery low-NLL and shared-rank curves come

from the original six-rule cohort, while the other curves come from the extension. These values are descriptive, involve no new test, and exclude selection cost; under about eight passes over each selected set, they are not single-pass or net compute savings.

Table 11: **Effective-token savings to match a reference endpoint.** For each training seed, the first selector's saved PPL curve is linearly interpolated to find where it reaches the reference's endpoint PPL at 2.5B effective prediction tokens; values are the saved share of that budget. Evaluations are about 250M tokens apart. These are descriptive effective-token equivalents under the repeated-data continuation, not net compute savings; selection cost is excluded.

Selector	Reference	Seed 300	Seed 301	Seed 302	Mean
Data2Delta, 8,190-d	Independent uniform	19.5	19.6	19.0	19.4
Low NLL	Independent uniform	19.1	18.7	18.8	18.9
Data2Delta, 1,024-d	Independent uniform	17.6	18.7	16.4	17.6
BGE-leverage, 1,024-d	Independent uniform	1.4	1.8	—	—
Data2Delta, 1,024-d	BGE-leverage, 1,024-d	16.3	17.0	17.0	16.8
Data2Delta, 8,190-d	Cluster-budget random	2.1	2.8	2.6	2.5
Data2Delta, 8,190-d	Low NLL	0.5	0.9	0.1	0.5

B.2 Original-only continuation

The original-only study omits the synthetic-injection step. Vocabulary, signatures, candidate graph, and MinHash filtering are rebuilt using only the 600,000 original chunks, yielding a common residual of 598,210 examples. All four rules retain 540,000 examples and continue the same step-5,000 model for 954 updates (1,000,341,504 effective prediction tokens). Saved original-only curves contain evaluations at updates 238, 477, 715, and 954. This is continued pretraining on original chunks, not random-initialization training or a guarantee that the source corpus contains no noise.

The frozen design pairs selection replicates 0–2 with training seeds 500–502. Gradient and BGE selections refit with the corresponding replicate seed; uniform sampling also changes its selection seed, while the low-NLL rule is deterministic. Thus the replicates jointly vary selection and training order where applicable, without separately estimating their variance components. The planned endpoint-NLL family contains Data2Delta versus low NLL, independent uniform sampling, and fresh BGE, with Holm correction across all three comparisons.

All twelve trajectories and endpoint Avg8 evaluations are complete. Data2Delta has the lowest mean PPL (13.30265), followed by independent uniform (13.30315), BGE (13.30459), and low NLL (13.30657; Table 12). Its PPL differences from low NLL and BGE are negative in all three replicates, and those from independent uniform in two of three. The corresponding mean differences are -0.00391, -0.00194, and -0.00050. None passes the prespecified three-comparison Holm test (Table 13). In particular, the low-NLL contrast has raw NLL $p = 0.0212$ and Holm $p = 0.0636$: its pointwise PPL interval excludes zero, but is not adjusted for the comparison family. These outcomes are analyzed within the 1B extension, separately from the 2.5B residual study.

Table 12: **Original-only continuation endpoints.** Means $\pm$ sample SD over seeds 500, 501, 502, at 1,000,341,504 effective prediction tokens; four methods use the same rebuilt first-stage residual and retain 540,000 rows. The source pool contains no additional synthetic injection. Avg8 is descriptive.

Rule	n	PPL $\downarrow$	Avg8 (%) $\uparrow$
Data2Delta	3	13.30265 ± 0.00084	44.169 ± 0.121
Low NLL	3	13.30657 ± 0.00026	44.320 ± 0.136
Independent uniform	3	13.30315 ± 0.00052	44.239 ± 0.055
BGE-M3 (1,024-d)	3	13.30459 ± 0.00211	44.158 ± 0.072

Mean Avg8 is 44.169% for Data2Delta, 44.320% for low NLL, 44.239% for independent uniform, and 44.158% for BGE. Data2Delta minus control is -0.1505, -0.0704, and $+0.0114$ percentage points in that order. These descriptive differences do not establish a downstream improvement or noninferiority. The seed-500 Data2Delta run resumed from update 95 with optimizer, per-rank

Table 13: **Original-only paired endpoint comparisons.** Data2Delta minus control, in 10^{-3} PPL. Intervals are pointwise paired 95% t intervals over the three selection/training replicates (seeds 500–502), not multiplicity-adjusted intervals. Tests use endpoint NLL; Holm adjustment covers all three planned contrasts within this family. Negative favors Data2Delta.

Control	ΔPPL	95% interval	Raw p	Holm p
Low NLL	-3.91	$[-6.41, -1.42]$	0.02119	0.06358
Independent uniform	-0.50	$[-2.54, +1.55]$	0.40646	0.40646
BGE-M3 (1,024-d)	-1.94	$[-5.16, +1.28]$	0.12209	0.24417

random-number-generator, and sampler states restored; its endpoint uses the same 954-update protocol.

C NATURAL DUPLICATE EVALUATION

We use the third-party NEWS-COPY mirror `chenghao/NEWS-COPY-eval` on Hugging Face, at revision `8886f0e7b5c7f59690641c0c007f3a285d8a97a3`. The validation split has 4,988 texts, 2,609 source families, 2,379 extra members, and 28,547 within-family positive pairs. The test split has 14,211 texts, 9,282 source families, 4,929 extra members, and 54,996 within-family positive pairs. Split sizes and positive-pair counts agree with the original evaluation specification. Prepared data and their manifests have recorded hashes.

Each rule operates directly on the entire split, with no preceding MinHash stage. The geometric rules use $K = 512$ and $\lambda = 0.01$ at the suite's predefined 5%, 10%, and 20% removal rates. No test labels tune these settings. The low-NLL rule retains lower probe losses; the natural random rule uses a fixed random rank without first-stage conditioning. Thus the shared-rank issue in the continuation experiment does not apply to this natural random rule.

All 24 selected masks were saved; their counts were recovered after a summary-serialization failure without changing the masks or selection settings. Reported counts pass both the identity $E = D - L$ and an independent count of the reduction in extra members. Hashes and recovery details are retained in the reproducibility bundle.

Source coverage counts singleton families as well as repeated families. At 10% test removal, Data2Delta loses 78 repeated families and one singleton, while low NLL loses eight repeated families and 1,052 singletons. A lost singleton contributes one to L and zero to extra-member removal. This explains much of the coverage difference, but it does not imply that every singleton is equally useful for language-model training. All rows below describe one selected set, with no selection-seed confidence intervals.

Table 14: **Source losses on NEWS-COPY test at 10% removal.** All rules delete 1,421 of 14,211 texts from 9,282 source families. Lost families have no surviving member; singleton and duplicate-family losses sum to the total. Retained-family percentages use all 9,282 families as the denominator.

Rule	Total lost families ↓	Lost singletons	Lost duplicate families	Families retained (%) ↑
Data2Delta	79	1	78	99.15
BGE-leverage	60	0	60	99.35
Low NLL	1,060	1,052	8	88.58
Random	804	798	6	91.34

Table 15: **All natural validation selections.** Counts are descriptive, with one saved mask per row. Lost singletons are a subset of L; E/I uses the initial extra-member count.

Removal	Rule	D	$L \downarrow$	$E \uparrow$	Lost singletons	E/I (%)
5%	Data2Delta	249	5	244	0	10.26
5%	BGE-leverage	249	2	247	0	10.38
5%	Low NLL	249	131	118	129	4.96
5%	Random	249	115	134	115	5.63
10%	Data2Delta	499	16	483	0	20.30
10%	BGE-leverage	499	5	494	0	20.77
10%	Low NLL	499	244	255	238	10.72
10%	Random	499	230	269	230	11.31
20%	Data2Delta	998	28	970	0	40.77
20%	BGE-leverage	998	22	976	1	41.03
20%	Low NLL	998	475	523	459	21.98
20%	Random	998	443	555	437	23.33

Table 16: **All natural test selections.** Counts are descriptive, with one saved mask per row. Lost singletons are a subset of L; E/I uses the initial extra-member count.

Removal	Rule	D	$L \downarrow$	$E \uparrow$	Lost singletons	E/I (%)
5%	Data2Delta	711	40	671	0	13.61
5%	BGE-leverage	711	17	694	0	14.08
5%	Low NLL	711	591	120	589	2.43
5%	Random	711	399	312	397	6.33
10%	Data2Delta	1,421	79	1,342	1	27.23
10%	BGE-leverage	1,421	60	1,361	0	27.61
10%	Low NLL	1,421	1,060	361	1052	7.32
10%	Random	1,421	804	617	798	12.52
20%	Data2Delta	2,842	404	2,438	256	49.46
20%	BGE-leverage	2,842	174	2,668	4	54.13
20%	Low NLL	2,842	1,895	947	1852	19.21
20%	Random	2,842	1,613	1,229	1589	24.93

D MECHANISM DIAGNOSTICS, STABILITY, AND MEASURED COST

D.1 OWN-PARENT AND SHARED-TEMPLATE INTERVENTIONS

Controlled own-parent interventions use seeds 11, 12, and 13 to compare a descendant's deletion rate when its own original is present or removed. Candidate counts and deletion budgets are matched, and clustering is refitted on each intervention pool. S denotes the main two-type pool; R denotes the five-type pool. These interventions differ from the observational MinHash parent-availability split retained in Appendix G. The ranges in Table 17 span the three intervention effects; they are not 95% confidence intervals or independent per-document tests.

In the S residual, the mean present-minus-absent deletion-rate differences are $+0.51$ percentage points for template wrapping and $+0.21$ for word shuffling. Effects need not be equally small in other pools: R's residual excerpt effect is $+5.83$ points, and its template effect is $+20.75$ points with only 53 eligible targets. Intervention pools share a common background, which limits treating them as independent corpus replications.

The shared-template test replaces 5,000 unrelated originals instead of adding copies. A visible prefix-plus-suffix template raises their full-pool deletion rate from 3.12% to 95.26%. Among 4,970 eligible residual examples, the rate rises from 1.75% to 30.97%. In a suffix-only control, all added suffixes lie outside the first 256 probe tokens; the deletion mask remains unchanged. This establishes a dependence on the feature-visible input. It is not evidence that suffixes in general are harmless or that the method only recognizes artifacts. A shared template is itself a shared component, and the experiment does not isolate its training value.

Table 17: **Controlled own-parent removal.** S is the main two-type pool and R the five-type pool. Here, n counts eligible descendants per intervention and $\Delta = \text{Present} - \text{Absent}$. Rates average three intervention seeds; the final column is the observed range of paired differences in percentage points, not a confidence interval.

Pool	Domain	Type	n	Present (%)	Absent (%)	Δ pp	Range
S	Full	Template	2500	85.64	84.39	+1.25	[-0.60,+2.24]
S	Full	Word shuffle	2500	90.11	88.75	+1.36	[+0.04,+3.52]
S	Residual	Template	2500	68.31	67.80	+0.51	[+0.20,+1.12]
S	Residual	Word shuffle	2500	85.31	85.09	+0.21	[-0.32,+1.00]
R	Full	Template	1000	66.90	66.20	+0.70	[-18.70,+21.30]
R	Full	Markup	1000	82.63	81.27	+1.37	[-1.20,+3.90]
R	Full	Character noise	1000	62.53	61.63	+0.90	[-0.40,+2.40]
R	Full	Word shuffle	1000	85.67	85.30	+0.37	[-1.10,+1.90]
R	Full	Excerpt	1000	14.10	8.77	+5.33	[+4.90,+6.00]
R	Residual	Template	53	31.45	10.69	+20.75	[+7.55,+37.74]
R	Residual	Markup	1000	77.97	79.20	-1.23	[-5.80,+1.30]
R	Residual	Character noise	1000	59.13	54.83	+4.30	[-0.60,+11.90]
R	Residual	Word shuffle	1000	84.57	84.20	+0.37	[-1.70,+2.50]
R	Residual	Excerpt	1000	12.53	6.70	+5.83	[+5.20,+7.00]

D.2 Selection seeds and cluster allocation

With alternative clustering/fitting seeds 1, 2, and 3, Data2Delta's deletion overlap with seed 0 is 60.50%, 62.30%, and 66.30%, respectively. The denominator is the deletion-set size (37,901); these are not Jaccard similarities. Injected recall ranges from 73.76% to 77.37% across the four seeds. Confirmation nevertheless uses the original fixed mask, so these diagnostics do not establish robustness of training outcomes to reselecting the data.

Holding Data2Delta's deletion count in every cluster fixed but deleting randomly within each cluster gives 74.72%, 74.74%, and 74.70% recall in three selections, compared with 76.16% for Data2Delta. This preserves cluster-level allocation while replacing within-cluster leverage ranking. The within-cluster random selection with randomization seed 1 has now been trained with three training-order seeds (Appendix B). It retains exactly the main method's per-cluster deletion counts; those counts themselves are determined by leverage. The two other within-cluster random masks remain selection-only diagnostics. The identity $\sum_{i \in c} \ell_i = d_\lambda(c)$ gives a cluster-average score $d_\lambda(c)/n_c$, so cluster size and effective dimension can affect a global ranking (Appendix M).

Table 18: **Selection sensitivity.** Injected-descendant recall (%) across both stages, with fixed budgets and features. Apart from the main configuration ($K = 512$, $\lambda = 0.01$), this grid reports selection-only diagnostics.

K	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$
64	64.56	64.69	65.25
512	76.16	76.16	76.50
4096	75.30	75.37	75.86

Reducing gradient dimension to 1,024 or 256 yields recall of 75.45% and 72.97%, respectively, under the same deletion budget. The 1,024-dimensional variant now has three continuation trajectories and a dimension-matched BGE comparison (Appendix B). The 256-dimensional variant and the remaining sensitivity grid still concern selection statistics only. Higher-precision feature extraction and checkpoint/model-scale transfer remain untested.

D.3 Cluster allocation of the main mask

Data2Delta's 37,901 second-stage deletions touch 86 of its 512 residual clusters, and five clusters receive 34,903 (92.1%; Table 19). C1–C3 are 91.0%, 85.5%, and 81.5% word-shuffled and are

deleted entirely, including 3,361 original chunks that share these clusters. C4 contains only template-wrapped descendants and is also deleted entirely. Low NLL and the fixed-geometry variants delete most of C1–C3, but fewer C4 members.

C5 contains 4,874 original chunks and three template-wrapped descendants; Data2Delta deletes 4,873 of them. Frozen scores, fixed clusters, and low NLL delete 48, 98, and 44 of these rows, and gradient-1024, which clusters its own features, deletes 1,228. Because the cluster-budget control inherits Data2Delta's per-cluster counts, it also deletes 4,873. A random sample of 40 members spans biomedical and physical-science prose, court opinions, patents, C/C++ headers, shell scripts, HTML, LaTeX mathematics, news, reviews, and chat logs. Their density of code-like symbols matches that of residual originals (0.0314 versus 0.0312 in samples of 500), and the four surviving members are European Parliament proceedings in Danish, Swedish, and Hungarian. We did not identify a shared textual attribute. Despite their textual diversity, these rows receive low within-cluster leverage scores relative to the other residual candidates. Data2Delta and the frozen-score and fixed-cluster selectors differ markedly in C5 retention, while their confirmation endpoint PPL differences remain statistically unresolved.

Table 19: **Where Data2Delta's second-stage budget goes.** The five residual clusters with the most Data2Delta deletions (C1–C5, ordered by deletion count) and all remaining clusters. Composition columns count candidates by provenance; deletion columns count each rule's deletions among those rows. Frozen and fixed-cluster rows are defined by Data2Delta's residual clusters, not by the clusters those rules used. All rules delete 37,901 rows.

		Composition			Deleted by				
Cluster	Size	Shuffled	Template	Original	Data2Delta	Frozen	Fixed	Low NLL	Grad-1024
C1	9,033	8,220	0	813	9,033	8,561	8,582	8,332	7,878
C2	7,977	6,819	0	1,158	7,977	6,371	6,509	6,331	6,208
C3	7,533	6,143	0	1,390	7,533	6,608	6,794	6,616	6,310
C4	5,487	0	5,487	0	5,487	2,882	1,338	166	2,465
C5	4,877	0	3	4,874	4,873	48	98	44	1,228
Other 507	589,915	5,580	10,956	573,379	2,998	13,431	14,580	16,412	13,812

D.4 Same-text feature-extraction microbenchmark

Table 20 measures warm scoring of the same 1,024 texts three times. Gradient, NLL, and activation extraction use the first 256 probe tokens; BGE uses its own tokenizer at maximum lengths 256 and 8,192. The latter is a configured maximum, not the actual length of every text. GPU measurements use one A800, batch size eight, and exclude model loading. The gradient timing covers the 8,190-dimensional sketch; its additional projection to 1,024 dimensions is performed separately and excluded. BGE timing uses the raw-text sample with its own truncation policy, whereas the representation comparison encodes the supplied probe-token prefixes. The CPU MinHash row measures local vocabulary/signatures/candidate union-find on this batch, not an extrapolation to the full corpus.

Table 20: **Warm timing on the same 1,024 texts.** Three timing repetitions; GPU methods use one A800 and batch size eight, including their own tokenizers and excluding model loading. These are extraction timings, not end-to-end corpus selection or training costs.

Method	Max tokens	Mean (s)	SD (s)	Texts/s
Gradient sketch	256	9.862	0.009	103.8
Probe NLL	256	4.500	0.156	227.6
Probe activation	256	4.351	0.000	235.3
BGE-M3	256	3.968	0.050	258.1
BGE-M3	8192	4.889	0.045	209.4
MinHash (CPU)	—	0.256	0.002	3996.3

These results do not support a speed advantage for the current gradient configuration over BGE-M3. Clustering, ridge scoring, corpus I/O, and continuation costs must be added for an end-to-end comparison; the saved forward computation of low-NLL selection has not been exchanged for additional training under a total-compute budget. Arithmetic cost estimates and historical equal-NLL interpolation are retained separately in Appendices F and K.

E Source accounting and a random-deletion reference

E.1 Accounting identity

Each source family has exactly one original, and each injected descendant has exactly one parent. With F families and I descendants, $N = F + I$. After deleting D rows and losing L entire families, the number of retained members beyond one per surviving family is

$$R = (N - D) - (F - L) = I - D + L. \tag{7}$$

The $D - J$ deleted originals partition into L families that disappear and P families with surviving descendants. Therefore,

$$E = I - R = D - L = J + P, \qquad L \geq \max(0, D - I). \tag{8}$$

These identities do not assume that members of a source family are semantically equivalent. They also do not count naturally occurring redundancy across distinct originals. In the main two-type pool, $D = 72{,}893$ and $I = 59{,}814$, so at least 13,079 families must disappear. At the historical removal budget $D = 65{,}981$, the lower bound is 6,167. The historical full pool includes three pairs of identical originals; all six rows belong to singleton source families and do not contribute to the injected extra-member count.

E.2 Expected recall under independent random completion

In the historical two-type experiment, MinHash deletes $D_0 = 31{,}670$ rows and $J_0 = 29{,}724$ descendants. Uniformly deleting another $D - D_0 = 34{,}311$ rows without replacement from its residual would give

$$\mathbb{E}[\text{Recall}] = \frac{J_0 + (D - D_0)(I - J_0)/(N - D_0)}{I} = 52.4419\%. \tag{9}$$

The additional descendant count is hypergeometric, with standard deviation about 38.5 rows. This is an analytic selection reference, not a trained arm. It is also distinct from the main protocol's Shared-rank random arm, whose shared rank induces dependence on the first stage.

F Implementation and computational cost

F.1 Configuration and reproducibility status

The reproducibility bundle records hashes for the frozen plans, retained masks, training sets, token caches, per-run evaluations, and analysis inputs. The checkpoints evaluated in the main and dense historical protocols are updates 238, 477, 715, 954, 1,192, 1,430, 1,669, 1,907, 2,146, and 2,384.

F.2 Operation and storage estimates

For n examples, sketch dimension d, cluster size m_c, and $q_c = \min(m_c, d)$, dense within-cluster scoring costs $O(\sum_c (m_c d q_c + q_c^3))$ arithmetic operations. Global sorting costs $O(n \log n)$, and sketch storage is $O(nd)$.

A $6ND$ training-style estimate for feature extraction with approximately 1.01×10^9 parameters and $659{,}814 \times 256$ input tokens is about 1.0×10^{18} FLOPs. Continuing the same model for 2.5B tokens is about 1.5×10^{19} FLOPs. The approximately 6.8% ratio reflects more than eight training passes over the retained corpus; it does not transfer to single-pass pretraining. Under a single-pass comparison on the same pool, feature extraction is roughly half a training pass. Probe NLL requires a forward pass, conventionally estimated at about $2ND$. Float16 storage of the 8,190-dimensional sketches takes about 10.8 GB. These are arithmetic estimates. The warm same-text timing experiment is reported separately in Appendix D; neither supplies total pipeline cost.

Table 21: **Implementation settings.** Historical and current selection recipes are distinguished because their outputs are not interchangeable.

Component	Configuration and evidence status
Probe and training start	Pythia-1B, checkpoint 5,000. No other probe checkpoints or target sizes are evaluated here.
Source window	600,000 chunks sampled from the 9,764,864 preshuffled Pile chunks at steps $[7{,}149, 9{,}533)$, after the probe and continuation checkpoint.
L8 sketch	Layer-8 `dense_h_to_4h.weight`; $91 \times 90 = 8{,}190$ projected coordinates; projection seed 0; first 256 tokens; no Adam preconditioning; float16 storage.
Current MinHash	64 unigram signatures; frozen candidate graph; estimated Jaccard threshold 0.8; 34,992 removals; fixed rank shared with residual random.
Current leverage	$K = 512$, $\lambda = 0.01$, clustering seed 0, 50,000 fitting examples, 25 iterations, float64 ridge calculation, fixed ties. Retained masks reproduce from cached features.
Shared-graph audit	Word/character-3-gram TF-IDF top-12 neighbors plus MinHash LSH edges; 64 permutations, hash seed 1729, four rows per band, bucket cap 64; 203,556 test edges. Transductive construction; cached model scores are not re-extracted in the audit.
Historical full-pool leverage	Complete clustering, ridge, sampling, numeric-branch, tie-breaking, and command configuration is not locked. Saved selected sets and their training outcomes are available.
Historical MinHash	Whitespace-tokenized 5-grams; 32 permutations; threshold searched in $[0.5, 0.95]$. Selected threshold and dependency versions are not locked.
Training	Eight GPUs; bf16; per-GPU microbatch 4; accumulation 64; length 512; remaining optimizer and evaluation settings in Section 3.

G Selected-set composition and training trajectories

G.1 Independent numerical checks

The 18 discovery trajectories each contain ten scheduled evaluations, with 1,024,000 labels per evaluation. All six selected sets contain 586,921 rows from the common residual. Endpoint means, paired differences and intervals, and Holm-adjusted tests reproduce from the recorded per-run results; detailed consistency checks are included in the reproducibility bundle.

All three endpoint differences are negative for every control. For three pairs, however, an exact two-sided sign test gives $p = 0.25$ even when all signs agree. The reported significance therefore depends on the prespecified parametric test. Low NLL and fixed-cluster rescoring have identical Holm-adjusted $p = 0.23$ because of step-down monotonicity; their unadjusted values are approximately 0.16 and 0.11. These statistics refer to discovery only; confirmation and the representation extension are reported in Appendix A.

G.2 Selected-set composition

Table 22: **Second-stage deletions and retained-set composition.** Every rule deletes 37,901 residual rows. Overlap is the fraction of deletions shared with Data2Delta. Set tokens include EOS; passes use the matched training budget.

Rule	Desc.	Original	NLL	Overlap	Tokens (M)	Passes
Shared-rank random	1,786	36,115	2.929	5.3%	296.004	8.45
Low NLL	24,518	13,383	2.766	60.5%	300.439	8.32
Lexical + low NLL	26,193	11,708	2.847	58.2%	300.338	8.32
Frozen scores	33,623	4,278	2.813	68.3%	299.995	8.33
Rescore, fixed clusters	29,317	8,584	2.809	66.1%	300.065	8.33
Data2Delta (main)	28,948	8,953	2.814	–	300.230	8.33

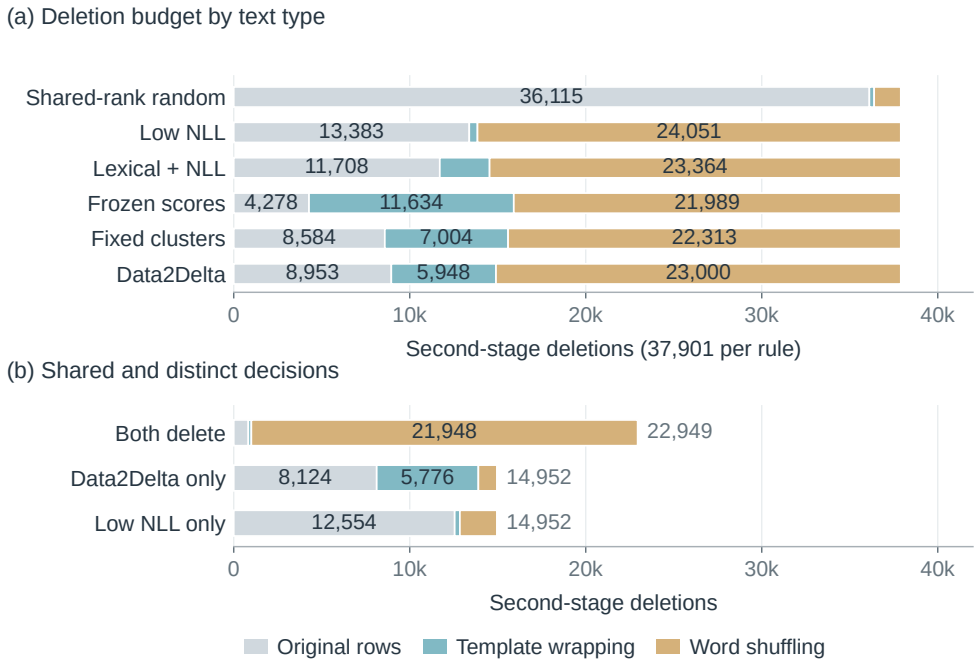


Figure 6: **Allocation of the second-stage deletion budget.** (a) Exact counts by provenance type; each rule deletes 37,901 rows. (b) The intersection and exclusive portions of the Data2Delta and low-NLL deletion sets. Segment labels identify larger counts; totals are shown for the three portions in (b). The 60.5% overlap divides the intersection by either deletion-set size; deletion-set Jaccard is 43.4%. Counts describe one fixed mask per rule. Full counts are in Appendix G.

Data2Delta and low NLL share 60.5% of their second-stage deletions. Data2Delta alone deletes 6,828 descendants: 5,776 template-wrapped and 1,052 word-shuffled. Low NLL alone deletes 2,398 descendants: 295 template-wrapped and 2,103 word-shuffled. Their remaining differences concern original rows. Neither an overlap percentage nor a marginal token total controls the training effect of the changed examples.

G.3 Provenance coverage on the training pool

Using the frozen parent mapping, each original and its injected copies define one provenance family. Coverage records whether any member remains, including a corrupted copy, with equal weight for every family irrespective of training value. The common residual represents 598,099 families in 624,822 rows. We count families present before the second stage and absent afterward, and independently verify that the reduction in extra members is $E = D - L$ for each mask. At the common second-stage budget $D = 37{,}901$, the unavoidable loss lower bound is $37{,}901 - 26{,}723 = 11{,}178$. Counts in Table 23 are incremental to the same MinHash residual, not totals from the unfiltered pool. All ten masks are evaluated without new selection or training; source hashes and the family mapping accompany the analysis script.

G.4 Observed parent availability

Table 25 and Figure 7 compare descendant deletion rates by whether the original parent remains after MinHash. This split is observational, unlike the controlled intervention in Appendix D: MinHash and rank determine group membership, and the absent/present groups have different NLL distributions.

Table 23: **Provenance coverage on the common training residual.** Every rule removes 37,901 rows from 624,822 candidates representing 598,099 source families. L counts families lost completely; $E = 37{,}901 - L$ counts removed extra members. Counts are incremental to the shared residual and describe one mask per rule. A surviving member may be a perturbed copy.

Selector	Families lost L	Families retained (%)	Extra members removed E
Shared-rank random	34,667	94.20	3,234
Low NLL	16,100	97.31	21,801
Lexical + low NLL	12,023	97.99	25,878
Frozen scores	15,329	97.44	22,572
Fixed clusters	15,958	97.33	21,943
Data2Delta (8,190-d)	15,807	97.36	22,094
Data2Delta (1,024-d)	17,153	97.13	20,748
BGE-M3 (1,024-d)	35,202	94.11	2,699
Independent uniform	34,670	94.20	3,231
Cluster-budget random	16,517	97.24	21,384

Table 24: **Exact second-stage deletion counts.** Each rule removes 37,901 rows. "Original" describes provenance, not a quality label. The middle block adds the training-extension controls; the lower block partitions the Data2Delta and low-NLL deletion sets.

Rule or portion	Original	Template	Word shuffle	Total
Shared-rank random	36,115	272	1,514	37,901
Low NLL	13,383	467	24,051	37,901
Lexical + low NLL	11,708	2,829	23,364	37,901
Frozen full-pool scores	4,278	11,634	21,989	37,901
Rescore, fixed clusters	8,584	7,004	22,313	37,901
Data2Delta	8,953	5,948	23,000	37,901
Gradient (1,024-d)	9,378	7,704	20,819	37,901
BGE-M3 (1,024-d)	35,054	1,324	1,523	37,901
Independent uniform	35,297	951	1,653	37,901
Cluster-budget random	9,812	5,935	22,154	37,901
Both delete	829	172	21,948	22,949
Data2Delta only	8,124	5,776	1,052	14,952
Low NLL only	12,554	295	2,103	14,952

G.5 TEN-POINT DISCOVERY TRAINING TRAJECTORIES

Endpoint negative-pair counts are 3/3 for every control. Across all ten evaluations, the counts are 30/30 for residual random and lexical-plus-NLL selection, 28/30 for frozen scores, and 24/30 for both low NLL and fixed-cluster rescoring. These are descriptive trajectory counts, not independent sample sizes.

Table 25: **Second-stage descendant deletion rates (%).** "Absent" means that MinHash already removed the original parent. Parent status is observational. Parent-absent/present group sizes are 13,452/2,994 for templates and 3,736/23,026 for word shuffling.

	Template wrapping		Word shuffling	
Rule	Absent	Present	Absent	Present
Low NLL	3.3	0.9	78.2	91.8
Lexical + low NLL	1.3	88.6	23.9	97.6
Frozen full-pool scores	68.0	83.1	69.8	84.2
Rescore, fixed clusters	40.1	53.8	71.4	85.3
Data2Delta (main)	34.8	42.1	77.5	87.3

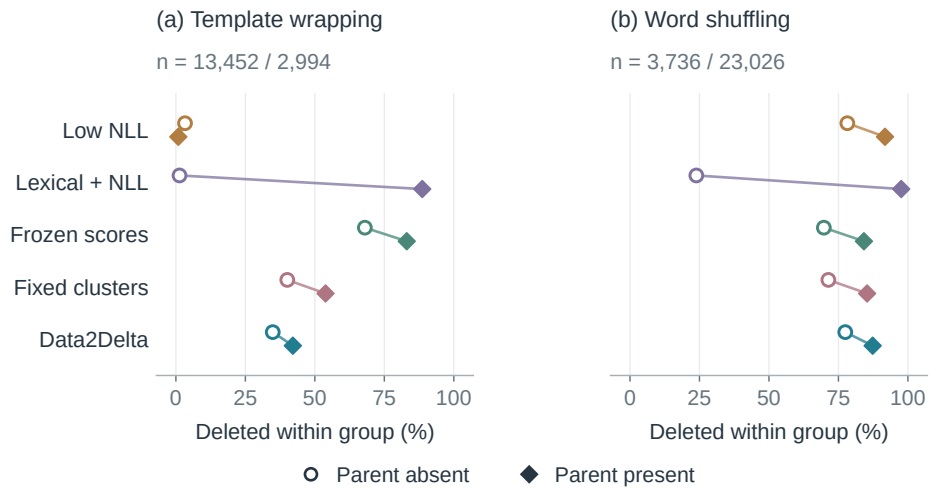


Figure 7: **Deletion rates by observed parent availability.** Open circles denote a parent absent from the common residual; filled diamonds denote a parent still present. Panels separate template wrapping and word shuffling, with group sizes listed in absent/present order. Each percentage uses its own residual subgroup as the denominator. Parent status is not randomized: absent/present groups have mean NLL 3.04/1.95 for templates and 5.21/5.83 for word shuffling. The comparison describes selection behavior, not a causal parent effect. Exact rates are in Table 25.

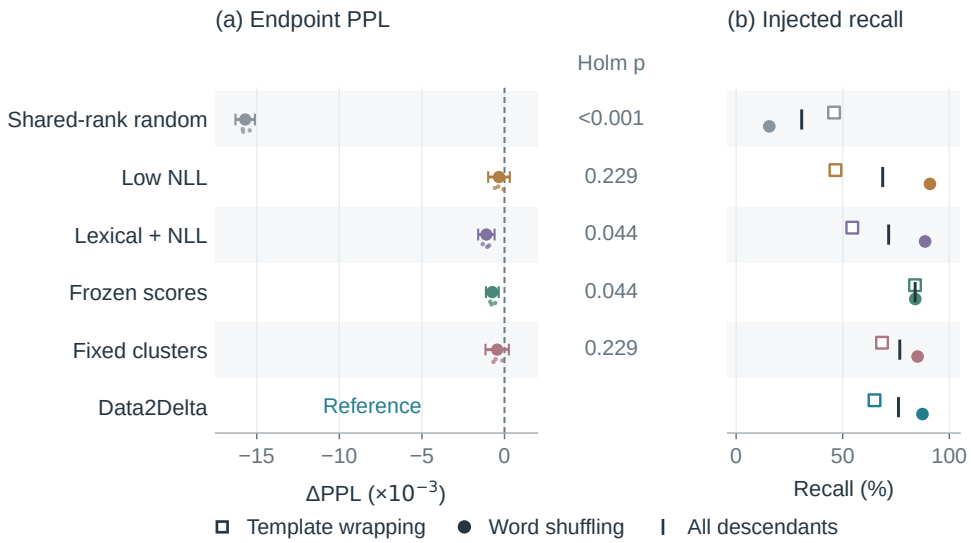


Figure 8: **Endpoint differences and removal recall for the same six sets.** (a) Data2Delta-minus-control mean PPL differences, pointwise paired 95% t intervals, three per-seed differences, and Holm-adjusted NLL tests across five comparisons. Negative differences favor Data2Delta; the full range includes the random comparison and its interval. (b) Recall across both deletion stages, with template-wrapped, word-shuffled, and all descendants distinguished by marker shape. Vertical bars mark all-descendant recall on the same axis. Recall denominators are the original injected groups, rather than the residual subgroups in Figure 7.



Figure 9: **Training on the common residual.** (a) Three-seed mean PPL for residual random, low NLL, and Data2Delta. (b–e) Data2Delta minus each model-informed control, with a marked mean curve and three thin individual paired trajectories. Negative values favor Data2Delta. Rankings change at intermediate budgets; fixed-cluster rescoring, for example, has lower mean PPL at 1.5B. The ten checkpoints on each trajectory are repeated measurements.

H FULL-POOL SELECTION UNDER SEPARATE PROTOCOLS

These experiments precede the main residual protocol. Their complete selection configuration is not locked, and their training sets are not outputs of the frozen configuration used in the main evaluation. We retain them as evidence about fixed historical sets, without pooling their results with the main experiment.

H.1 FULL-POOL DETECTION AND RESIDUAL SELECTION

Historical MinHash removes 31,670 rows (4.80%) and 29,724 descendants: recall is 99.35% for template wrapping and 0.01% for word shuffling. At a matched removal budget of 65,981 rows, sequential MinHash–leverage selection has 93.80% recall, independent full-pool leverage has 93.68%, and MinHash with independent random completion has expected recall 52.44%. The sequential selector removes only 72 more descendants than full-pool leverage. This small count difference is not evidence of a training benefit from residual rescoring. Full-pool and sequential leverage have $P = 93$ and 128, and lose 9,854 and 9,747 families, respectively; the budget lower bound is 6,167.

Only 195 template-wrapped descendants remain after this historical MinHash stage, all with retained parents. Residual leverage removes 27 of them (13.8%), versus 99.47% recall for template wrapping in the full-pool run. Residual word-shuffle recall remains 88.16%.

Table 26: **Historical injected recall by perturbation (%).** Native deletion counts appear in parentheses and differ across methods. WS: word shuffle; TW: template wrapping; MI: markup insertion; CN: character noise; EX: excerpt.

Pool and method (deletions)	WS	TW	MI	CN	EX	All
Two-type: MinHash (31,670)	0.01	99.35	–	–	–	49.69
Two-type: BGE cosine (72,013)	26.33	48.42	–	–	–	37.37
Two-type: Data2Delta cosine (72,893)	80.12	90.18	–	–	–	85.15
Two-type: Data2Delta leverage (65,981)	87.89	99.47	–	–	–	93.68
Two-type: MH + Data2Delta (65,981)	88.16	99.44	–	–	–	93.80
Two-type: random (65,981)	9.48	10.07	–	–	–	9.78
Five-type: MinHash (20,146)	0.02	99.42	0.33	0.47	51.89	30.42
Five-type: BGE cosine (72,622)	26.47	48.69	74.82	50.34	19.19	43.91
Five-type: Data2Delta leverage (65,983)	89.63	94.72	84.91	54.51	13.80	67.51
Five-type: random (65,983)	9.39	9.18	10.06	9.88	10.61	9.82

BGE-cosine and Data2Delta-cosine use the same selection rule: cluster the representation vectors; visit examples within each cluster in increasing cosine similarity to the centroid; delete an example if its cosine similarity to an already retained representative exceeds a threshold. Binary search selects a threshold near the removal target. BGE-cosine is a SemDeDup-style baseline using BGE-M3 representations (Abbas et al., 2023; Chen et al., 2024).

H.2 CONTINUATION ON TWO FIXED HISTORICAL SETS

The two historical sets each contain 586,921 examples. The Data2Delta set starts from full-pool leverage removal of 65,981 examples, then removes another 6,912 uniformly sampled rows to match the largest native removal count among the compared methods, 72,893 (Data2Delta-cosine). The random and Data2Delta sets retain 53,335 and 3,737 descendants, respectively (9.09% and 0.64% of retained rows). They represent 539,379 and 583,317 source families, and contain 295,714,813 and 300,269,890 tokens including EOS. The 2.5B-token budget corresponds to 8.45 and 8.33 passes.

Training otherwise follows Section 3. Discovery uses seeds 100–102; confirmation uses 200–202. The original three-budget results are given in Table 27. Their 1B point is update 953 (999,292,928 effective prediction tokens), whereas the dense curve uses update 954 (1,000,341,504). The maximum NLL difference at this shifted checkpoint is 4.7×10^{-5}. Recorded NLL at aligned updates 238 and 2,384 matches exactly. This is a check of replayed metrics at those points, not a bytewise comparison of all model states.

Table 27: **Historical training on fixed full-pool sets.** Three seeds per stage. ΔPPL is historical Data2Delta minus random; intervals are paired pointwise 95% t intervals. NLL tests are Holm-adjusted over three budgets within each stage.

Stage	Tokens	Random	Data2Delta	ΔPPL [95% CI]	Holm p
Discovery	0.25B	13.4556	13.4451	$-0.0105[-0.0114, -0.0096]$	0.00082
Discovery	1B	13.3199	13.3039	$-0.0161[-0.0165, -0.0156]$	0.00011
Discovery	2.5B	13.2421	13.2224	$-0.0197[-0.0222, -0.0172]$	0.00088
Confirmation	0.25B	13.4556	13.4452	$-0.0104[-0.0120, -0.0089]$	0.0012
Confirmation	1B	13.3200	13.3037	$-0.0162[-0.0169, -0.0155]$	0.0003
Confirmation	2.5B	13.2420	13.2227	$-0.0194[-0.0210, -0.0178]$	0.00075

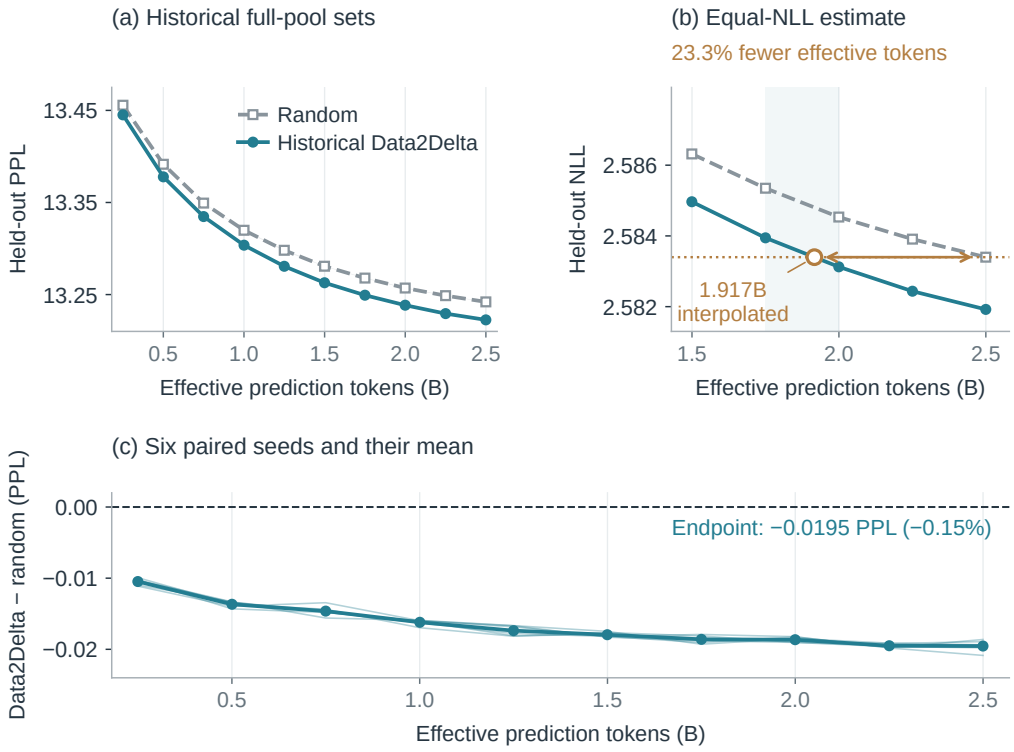


Figure 10: **Historical fixed full-pool sets.** The Data2Delta curve denotes a historical selection, distinct from the main residual algorithm. (a) Ten-point mean PPL for six seeds per arm; sample SD is narrower than the lines. (b) A local mean-NLL comparison: reaching the random endpoint at 2.4998B tokens requires an interpolated 1.917B Data2Delta tokens, bracketed by measurements at 1.750B and 2.000B. (c) Paired Data2Delta-minus-random differences for six seeds and their mean.

All six same-seed differences are negative at all ten budgets. Their mean increases in magnitude from -0.0105 to -0.0195 PPL. The sets have substantially different perturbation and source composition, so this improvement cannot be attributed specifically to gradients or leverage.

H.3 Earlier padded training and downstream tasks

Earlier training supplied padded examples. Seed-0 endpoint PPL was 13.2654 for random, 13.2618 for BGE, and 13.2352 for leverage. Averaging Avg8 over training seeds 0, 1, and 2 gives 44.2366% for leverage and 44.1918% for random: a paired difference of $+0.0447$ percentage points, with a 95% paired t interval of $[-0.3972, +0.4867]$ computed from the three seed-level Avg8 differences ($n = 3$, $df = 2$). Within each seed, Avg8 averages eight task metrics on a common propor-

tion scale: LAMBADA, ARC-Easy, ARC-Challenge, PIQA, SciQ, OpenBookQA, HellaSwag, and WinoGrande. ARC-Challenge, OpenBookQA, and HellaSwag use normalized accuracy; the others use accuracy. Evaluation is zero-shot with lm-eval 0.4.12.

This earlier protocol gives leverage 1.54% more effective prediction tokens and can repeat one document within a microbatch. We therefore do not use it as evidence for matched-token training utility or claim downstream equivalence.

I SHARED-GRAPH DIAGNOSTICS

The shared-graph experiment fixes $N = 10{,}560$, $F = 9{,}607$, $I = 953$, and $D = 1{,}056$. It crosses four edge scores with five endpoint-retention rules, using the same 203,556 candidate edges, row identities, and deletion budget. For each edge score, the complete edge order is fixed across rules. The endpoint rules prioritize a fixed random rank, longer capped token length, lower probe NLL, higher cached L8 leverage, or longer full-text character length. No uncertainty over samples or graph construction is estimated.

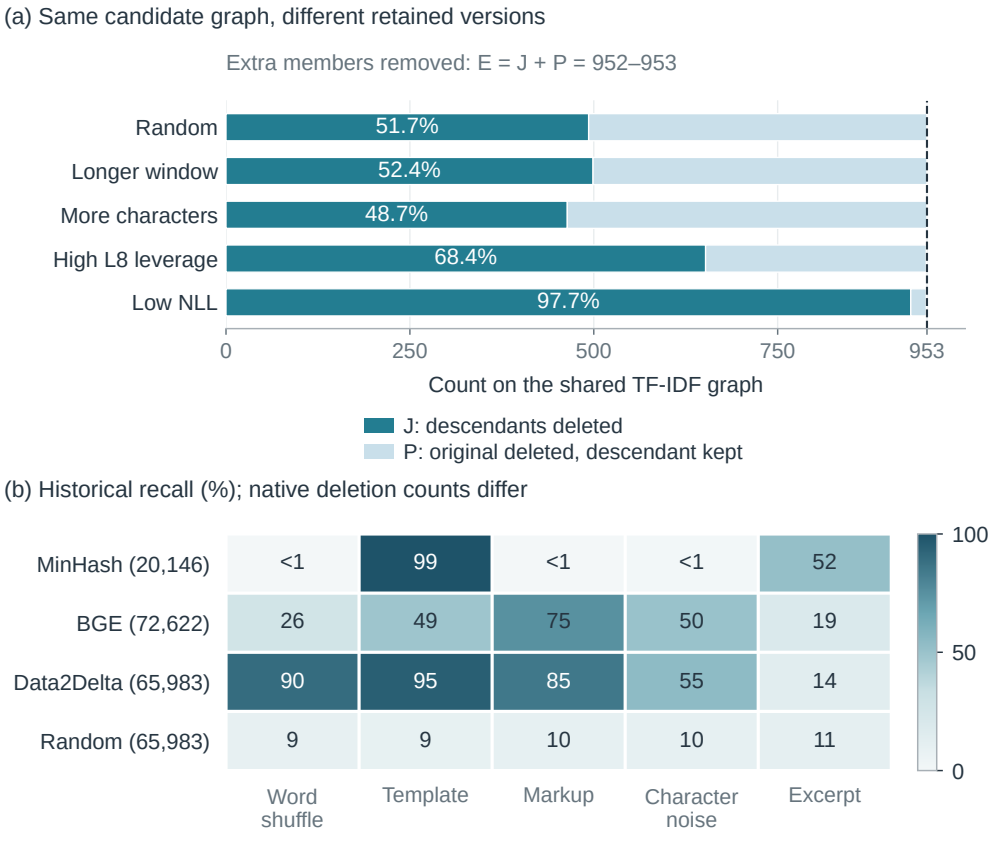


Figure 11: **Interpreting injected recall.** (a) Source-family accounting on a fixed TF-IDF graph: changing the endpoint rule preserves nearly the same extra-member removal $E = J + P$ while changing recall. (b) Historical five-type recall, with native deletion counts in parentheses; these budgets differ across methods. Neither panel measures training utility.

Of the 10,560 rows, 10,545 reach the 256-token cap, motivating the separate full-character-length control. The leverage priority uses cached, L2-normalized JL256 features, absolute $\lambda = 0.01$, fixed BGE clusters, and static scores inside greedy edge deletion. This differs from the 8,190-dimensional globally ranked selector in Algorithm 1. For TF-IDF edges, high leverage and low NLL both retain

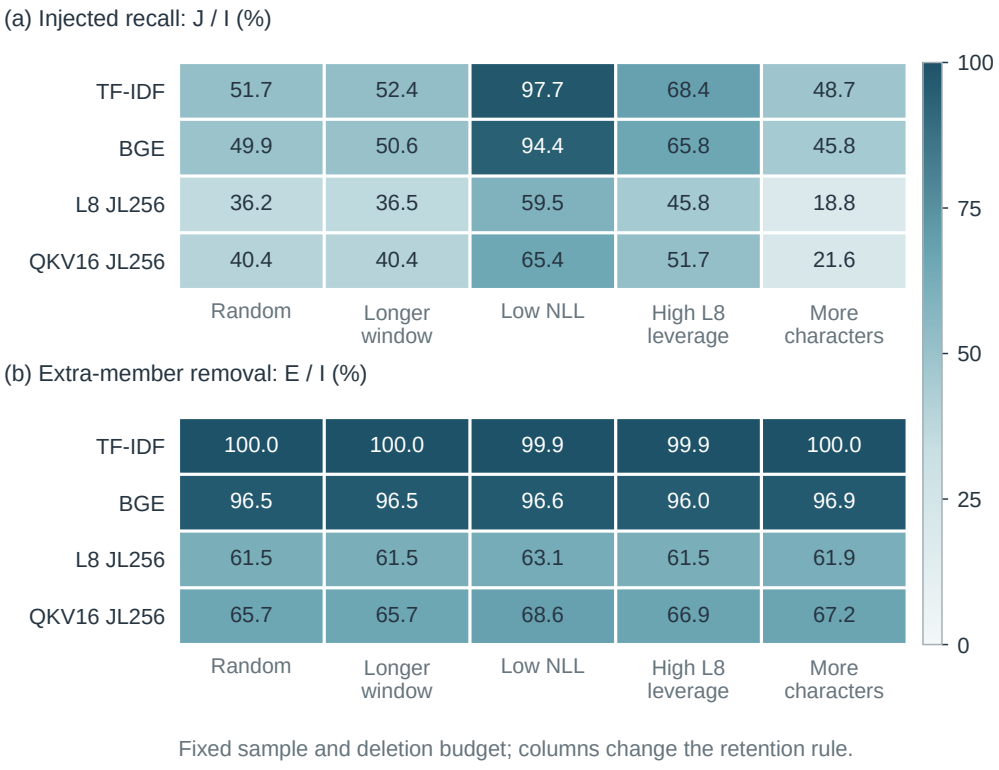


Figure 12: **All four edge scores and five retention rules.** Injected recall (top) and E/I (bottom) share a 0–100 scale. Low NLL has the highest recall in every row. Gradient edge scores have lower E/I than TF-IDF and BGE in this graph. Counts are exact for the fixed sample.

9,503 families, of which 9,474 coincide. Their retained row masks differ on 666 rows. The two runs share 1,009 deletion-triggering edges and each has 47 additional such edges.

Table 28: **Exact shared-graph counts.** $N = 10{,}560$, $F = 9{,}607$, $I = 953$, $D = 1{,}056$ for every row. Recall and E/I are percentages.

Edge score	Retain priority	J	Recall	P	E	E/I	L	R
TF-IDF	Random	493	51.7314	460	953	100.0000	103	0
TF-IDF	Longer window	499	52.3610	454	953	100.0000	103	0
TF-IDF	Low NLL	931	97.6915	21	952	99.8951	104	1
TF-IDF	High L8 leverage	652	68.4155	300	952	99.8951	104	1
TF-IDF	More characters	464	48.6884	489	953	100.0000	103	0
BGE native	Random	476	49.9475	444	920	96.5373	136	33
BGE native	Longer window	482	50.5771	438	920	96.5373	136	33
BGE native	Low NLL	900	94.4386	21	921	96.6422	135	32
BGE native	High L8 leverage	627	65.7922	288	915	96.0126	141	38
BGE native	More characters	436	45.7503	487	923	96.8520	133	30
L8 JL256	Random	345	36.2015	241	586	61.4900	470	367
L8 JL256	Longer window	348	36.5163	238	586	61.4900	470	367
L8 JL256	Low NLL	567	59.4963	34	601	63.0640	455	352
L8 JL256	High L8 leverage	436	45.7503	150	586	61.4900	470	367
L8 JL256	More characters	179	18.7828	411	590	61.9098	466	363
QKV16 JL256	Random	385	40.3987	241	626	65.6873	430	327
QKV16 JL256	Longer window	385	40.3987	241	626	65.6873	430	327
QKV16 JL256	Low NLL	623	65.3725	31	654	68.6254	402	299
QKV16 JL256	High L8 leverage	493	51.7314	145	638	66.9465	418	315
QKV16 JL256	More characters	206	21.6159	434	640	67.1563	416	313

J REMOVAL-BUDGET ALIGNMENT IN FULL-POOL SELECTION

Table 29: **Historical counts after matching removal budgets.** Random completion is included. Source accounting is distinct from semantic redundancy.

Pool; deletions	Method	Recall	E/I	L
Two-type; 72,893	BGE cosine	37.4695%	76.1126%	27,367
Two-type; 72,893	Data2Delta leverage	93.7523%	93.9747%	16,683
Two-type; 72,893	Random	10.8319%	20.5170%	60,621
Five-type; 72,622	BGE cosine	43.9065%	76.7490%	26,700
Five-type; 72,622	Data2Delta leverage	67.8527%	69.3820%	31,108

Table 30: **Counts before and after random completion.** The native budgets differ.

Pool	Method	Native D	Added	Final D	E: native $\to$ final	L: native $\to$ final
Two	BGE cosine	72,013	880	72,893	45,481 $\to$ 45,526	26,532 $\to$ 27,367
Two	Data2Delta cosine	72,893	0	72,893	51,209 $\to$ 51,209	21,684 $\to$ 21,684
Two	Data2Delta leverage	65,981	6,912	72,893	56,127 $\to$ 56,210	9,854 $\to$ 16,683
Two	Random	65,981	6,912	72,893	11,085 $\to$ 12,272	54,896 $\to$ 60,621
Five	BGE cosine	72,622	0	72,622	45,922 $\to$ 45,922	26,700 $\to$ 26,700
Five	Data2Delta leverage	65,983	6,639	72,622	41,110 $\to$ 41,514	24,873 $\to$ 31,108

Native removal counts differ and do not support a matched-budget ranking. At the aligned two-type budget, BGE-cosine removes the original but retains a descendant in 23,114 families, compared with 133 for Data2Delta-leverage. Thus their E/I difference (17.86 percentage points) is much smaller than their recall difference (56.28 points). In the historical implementation, random completion shares one generator across arms in arm order. The audit checks the saved sets and their multiset inclusion relationships.

Table 31: **Historical dense curves.** Six training seeds per arm; mean ± sample SD. Budgets are rounded effective prediction-token counts.

Tokens (B)	Random PPL	Historical Data2Delta PPL	ΔPPL
0.25	13.455622 ± 0.000409	13.445148 ± 0.000460	-0.010474
0.50	13.391498 ± 0.000321	13.377831 ± 0.000230	-0.013667
0.75	13.349330 ± 0.000263	13.334701 ± 0.000598	-0.014629
1.00	13.319832 ± 0.000439	13.303632 ± 0.000285	-0.016199
1.25	13.298095 ± 0.000719	13.280720 ± 0.000268	-0.017375
1.50	13.280783 ± 0.000265	13.262836 ± 0.000210	-0.017947
1.75	13.267909 ± 0.000392	13.249313 ± 0.000221	-0.018596
2.00	13.257102 ± 0.000445	13.238454 ± 0.000368	-0.018647
2.25	13.248827 ± 0.000308	13.229330 ± 0.000413	-0.019497
2.50	13.242068 ± 0.000247	13.222530 ± 0.000744	-0.019538

K EQUAL-NLL INTERPOLATION FOR FULL-POOL SELECTION

Let u be random's six-seed mean NLL at $T_R = 2{,}499{,}805{,}184$ effective prediction tokens. For adjacent Data2Delta observations at T_l and T_r, with mean NLL values $\ell_l \geq u \geq \ell_r$, define

$$\widehat{T}_D = T_l + \frac{\ell_l - u}{\ell_l - \ell_r}(T_r - T_l), \qquad \widehat{s} = 1 - \frac{\widehat{T}_D}{T_R}. \tag{10}$$

With $u = 2.583398716$, $T_l = 1{,}750{,}073{,}344$, and $T_r = 1{,}999{,}634{,}432$, linear NLL–token interpolation gives $\widehat{T}_D = 1{,}916{,}565{,}331$ tokens and $\widehat{s} = 23.33\%$. Linear PPL–token and linear NLL–log-token interpolation give 23.33% and 23.48%, respectively. This small spread is interpolation-method sensitivity, not a confidence interval. The interpolated crossing is not a separately trained checkpoint.

Using observed Data2Delta budgets as targets and interpolating random's curve gives savings of 13.9%–21.7% within the common measured range. Targets beyond random's endpoint are not extrapolated. The late curve is shallow, so a small vertical PPL difference corresponds to a larger horizontal token distance. The endpoint absolute PPL gap is 0.01954, approximately 0.15%. These estimates concern the historical fixed sets, exclude selection overhead, and do not establish net compute savings, downstream equivalence, or savings for the main residual protocol.

L SINGLE-EXAMPLE SIGNALS AND REFERENCE-SET DIAGNOSTICS

Single-example scalars. On the 10,560-row shared-graph test sample with 1,056 deletions, globally deleting by descending NLL gives 47.53% injected recall; deleting by descending QKV16 gradient norm gives 55.72%. A supervised diagnostic using nine scalar features, including four exact and projected L8/QKV16 norms, has AUROC 0.8643. It uses injection labels on its development set and tests whether individual-example signals predict those labels. It is not a label-free selection method or an evaluation of naturally occurring redundancy.

Reference-set intervention. This small experiment fixes candidate texts and replaces originals in an equal-sized reference set with unrelated texts. An external-ridge redundancy score is evaluated relative to the reference. There are 128 sources, 131 candidates, and 129 examples in each reference condition. Positive changes indicate a larger redundancy score when the same-source reference is present. Intervals use a source-family bootstrap (Table 32).

Only L8's word-shuffle interval includes zero. Since scales differ, these values cannot be interpreted as ratios of representation quality.

Small-sample global leverage. On the same test rows, with 32 fixed BGE clusters, JL256, L2 normalization, and 10% deletion, six ridge settings give recall ranges of 13.85%–14.59% for L8, 15.11%–15.84% for QKV16, and 16.68%–20.99% for BGE. Approximately 330 rows per cluster and 256 dimensions place this experiment in the $n_c > d$ regime, unlike the main selector

Table 32: **Change under the reference-set intervention.** Mean change and 95% bootstrap interval. Score scales differ across representations.

Representation	All perturbed candidates	Word-shuffled candidates
L8	13.7686 [11.1609, 16.1740]	0.2259 [−0.0383, 0.4791]
QKV16	13.3792 [10.8119, 15.6617]	0.7354 [0.3971, 1.0938]
BGE	13.5094 [12.6086, 14.3568]	9.5592 [8.8337, 10.3118]

(Appendix M). On the same graph, unigram-MinHash edges and a random endpoint priority give $J = 493$, $E = 952$, and $L = 104$. Its edge score differs from the TF-IDF random row in Table 28; the equal J is coincidental.

M GEOMETRY OF RIDGE LEVERAGE

These standard matrix identities explain the score for fixed clusters, representations, and $\lambda > 0$. They do not add a training-utility or batch-selection guarantee. Write $M = Z_c^\top Z_c$ and $A = M + \lambda I$.

M.1 REGULARIZED WHITENING

Equation (5) is $\ell_i = \|A^{-1/2} z_i\|_2^2$, the squared length after regularized whitening. For example, four copies of e_1 and one copy of e_2 have scores 0.2, 0.2, 0.2, 0.2, and 0.5 when $\lambda = 1$. Changing the sign of any row leaves its outer product and every score unchanged. Thus leverage cannot distinguish aligned from anti-aligned versions of the same direction, unlike a signed gradient inner product.

M.2 SINGLE DELETION AND THE LOG DETERMINANT

Deleting only row i gives $A_{-i} = A - z_i z_i^\top \succ 0$. The matrix determinant lemma yields

$$\det A_{-i} = \det A\,(1 - \ell_i), \qquad \log\det A - \log\det A_{-i} = -\log(1 - \ell_i). \tag{11}$$

Consequently $0 \le \ell_i < 1$, and, with all other rows fixed in the same cluster, the lowest leverage minimizes the one-row log-determinant decrease. With fixed representations and assignments, define $\Phi = \sum_c \log\det(M_c + \lambda I)$. Deleting row i changes only its cluster's term, so $\Phi - \Phi_{-i} = -\log(1 - \ell_i)$. Since this function increases with ℓ_i, global leverage ranking also orders these single-example decreases across clusters. Algorithm 1 deletes multiple rows using initial scores. Such decreases are not additive, so this identity does not prove batch log-determinant optimality, spectral preservation, or family coverage. Greedy DPP optimization also involves a different objective and update rule (Chen et al., 2018).

M.3 EFFECTIVE DIMENSION AND TWO SAMPLE-SIZE REGIMES

Summing scores within a cluster gives

$$\sum_{i \in c} \ell_i = \operatorname{tr}\big(M(M + \lambda I)^{-1}\big) = \sum_j \frac{\mu_j}{\mu_j + \lambda} = d_\lambda(c) \le \min(n_c, d). \tag{12}$$

Let $B = M_{-i} + \lambda I$, with row i removed from the moment matrix. Sherman–Morrison gives

$$\ell_i = \frac{s_i}{1 + s_i}, \qquad s_i = z_i^\top B^{-1} z_i \ge \frac{\|\Pi_{-i}^{\perp} z_i\|_2^2}{\lambda}, \tag{13}$$

where $\Pi_{-i}^{\perp}$ projects onto the orthogonal complement of the span of the other rows. If $n_c < d$ and the squared orthogonal component is much larger than λ, the score approaches one. General position alone does not imply this condition: two nearly parallel unit vectors can be linearly independent yet have orthogonal components far smaller than λ, yielding scores near $1/(2 + \lambda)$. When $n_c \gg d$, the other rows can span the full space, making this lower bound zero; directional second-moment energy then governs the score.

The main protocol has about 1,220 rows per cluster and $d = 8{,}190$, whereas the small-sample diagnostic has about 330 rows and $d = 256$. In the main residual, score quantiles at 10%, 50%, and 90% are 0.97882, 0.98449, and 0.98744, and the deletion threshold is 0.97449. Score–NLL Spearman correlation is -0.2068. Mean NLL is 2.9283 before selection, 4.7005 among deleted rows, and 2.8139 among retained rows. This association supports examining single-example loss as a control; the near-unit scores do not by themselves establish near-linear dependence among the deleted examples.

M.4 A TWO-MEMBER FAMILY'S SYMMETRY

Suppose one cluster contains unit vectors p (parent) and c (child), with other rows O: $A = pp^\top + cc^\top + O^\top O + \lambda I$. For $p \neq c$, let $w = (p - c)/\|p - c\|$ and $H = I - 2ww^\top$. Then $Hp = c$ and $Hc = p$. If $HO^\top OH = O^\top O$, then $HAH = A$, and

$$\ell_c = c^\top A^{-1} c = p^\top H A^{-1} H p = p^\top A^{-1} p = \ell_p. \tag{14}$$

For $p = c$, equality is immediate. Under these fixed same-cluster conditions, a parent–child score difference requires asymmetry in their relationship to the remaining rows. Different cluster assignments or a change in the reference set invalidate the stated exchange symmetry; the identity alone does not explain the observed PPL ordering.

N EXTENDED RELATED WORK

Data2Delta sits at the intersection of corpus pruning and model-aware selection. The relevant distinctions are the representation, the candidate-relative scoring rule, and how deletion counts are allocated; our experiments separate representation substitution from within-cluster membership choice.

Deduplication and representative selection. Lee et al. (2022) study near-duplicate examples and repeated substrings in language-model data. SemDeDup removes near neighbors within pretrained-representation clusters and compares member-selection rules, with similar accuracy for its tested CLIP rules (Abbas et al., 2023). D4 combines deduplication and diversification (Tirumala et al., 2023). Self-supervised prototype geometry also supports data pruning (Sorscher et al., 2022), and density-based pruning sets cluster-specific pruning rates from concept complexity (Kamal et al., 2024). Difficulty-only pruning can lose coverage at high pruning rates, which motivates coverage-centric (Zheng et al., 2023) and diversity–difficulty (Maharana et al., 2024) selection. FineWeb evaluates deduplication choices through training (Penedo et al., 2024), and NEWS-COPY supplies manually labeled natural duplicate news for detection evaluation (Silcock et al., 2023). Frequency- and length-aware subdocument selection separates detecting repeated spans from deciding how many copies to retain (Wang et al., 2026).

Gradient representations and data value. CRAIG and GRAD-MATCH select subsets whose weighted gradients approximate those of the full training or validation set (Mirzasoleiman et al., 2020; Killamsetty et al., 2021). TracIn estimates influence through gradient inner products along training (Pruthi et al., 2020); LESS uses projected gradients for targeted instruction selection, and TRAK uses them for scalable attribution (Xia et al., 2024; Park et al., 2023). LoGra exploits factored gradient structure for efficient projection (Choe et al., 2024). Data2Delta uses projected gradients for candidate-relative geometric scoring and global deletion, without a target set. Gradient- and error-norm scores (Paul et al., 2021) retain information that Data2Delta discards through row normalization. MATES and DsDm study model-aware pretraining selection (Yu et al., 2024; Engstrom et al., 2024), and Quad balances influence-based quality with cluster-level diversity (Zhang et al., 2025). Gradient alignment depends on model capability (Kazdan et al., 2026), and separating representation from selection algorithm matters in targeted instruction selection (Nayak et al., 2026); we hold the probe checkpoint fixed.

Loss filtering, mixtures, and set geometry. Perplexity-based pruning provides a direct single-example baseline (Marion et al., 2023; Ankner et al., 2025). Other pipelines select by domain weights (Xie et al., 2023a), similarity to a target distribution (Xie et al., 2023b), or rated quality (Wettig et al., 2024), and DCLM benchmarks such curation choices at scale (Li et al., 2024); Data2Delta

instead prunes within a pool without a target distribution or quality labels. Its single-deletion log-determinant identity relates to DPP objectives, for which greedy updates are an alternative to static decisions (Kulesza & Taskar, 2012; Chen et al., 2018). DiSF greedily selects pretraining files that make feature-covariance eigenvalues more uniform (Fan et al., 2025). Unlike density-based pruning and Quad, Data2Delta sets no explicit per-cluster budget. Repeated-data studies motivate examining a finite pool reused across passes (Hernandez et al., 2022; Muennighoff et al., 2023); our approximately eight-pass continuation evaluates selection in this regime.