ZERO-TRAINING SURFACE STATISTICS FOR PRETRAINING-CORPUS FILTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

The value of pretraining corpora depends not only on their size, but also on how effectively low-quality content is filtered. Heuristic filters are efficient, interpretable, and scalable, but their hand-designed rules are pipeline-specific and do not generalize across languages. We extract the underlying measurements from seven widely used pipelines and summarize the corpus problems they target in an operational taxonomy of seven corpus-quality failure modes. Building on heuristic filters' view of language structure, we form ZenoFilter from how these measurements respond across failure modes, yielding a training-free adaptive detector that measures departures from ordinary text in redundancy and lexical composition and supports an adjustable corpus-level operating point. On real WARC documents, ZenoFilter outperforms the filtering layers of four widely used pipelines: Gopher/C4, FineWeb, Dolma, and RefinedWeb. At matched keep rates on symmetric English, it leaves 2.16–2.63× fewer document-level and 1.74–2.12× fewer token-level residual artifacts while running about six times faster. Against the closest adaptive comparator DCAD, ZenoFilter is never worse across its keep-rate curve. On non-English monolingual text it performs comparably to FineWeb 2 and DCAD, the only comparators that cover it. It processes multilingual and code-switched documents without language gates or language-specific thresholds, and discriminates mixed-language documents comparably to English ones. In downstream validation it achieves lower held-out loss than FineWeb across the six content categories and on the source-aligned Nemotron-CC validation set; the FineWeb advantage persists on the equal-weight mean of the externally defined Pile-22 suite.

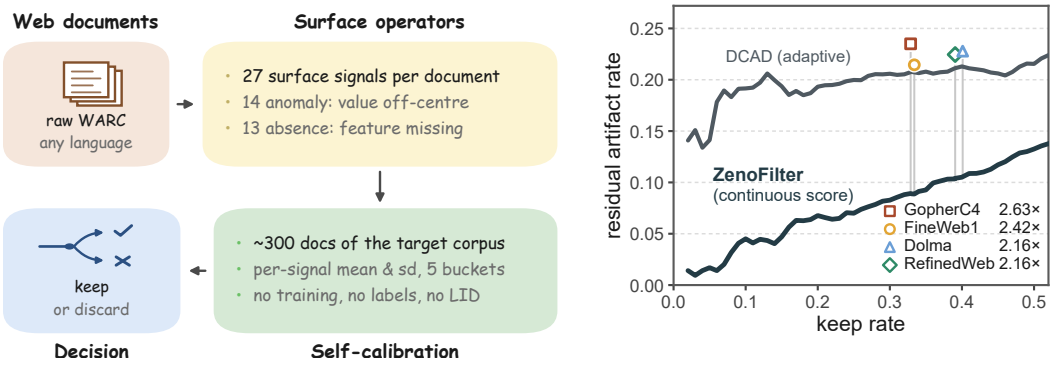

Figure 1: **Left:** twenty-seven surface signals per document, standardized against the moments of about three hundred documents of *the corpus being filtered* within five length buckets, then combined by a maximum and thresholded. No training. **Right:** peers are shown at their native keep rates, matched by retained document count, so each gap reflects both the length mix a filter keeps and its removal of artifacts; Appendix B.2.1 gives the length-standardized comparison. DCAD shares the pool, label and budget axis, and both scores are continuous; at its own operating point (keep 0.82, beyond this panel) the two residual rates are level.

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.

1 INTRODUCTION

Training-data quality is among the most consequential levers in language modelling: under quality-discounted scaling laws, higher-quality data reaches lower final loss at the same compute budget (Subramanyam et al., 2026). Corpus filtering is the principal point at which this quality can be improved before training, at only a fraction of the cost of training itself. Its central content-level decision is noise detection: although pipelines share broad operations such as extraction and cross-document deduplication, they diverge in deciding which remaining documents are usable enough to enter training.

In widely used rule-based pipelines, this decision is operationalized largely through zero-training surface statistics: the heuristic rules of Gopher and C4, the 46 quality signals shipped with RedPajama-V2 (Weber et al., 2024), and 13-gram self-overlap. Later pipelines build on the same statistical substrate. FineWeb layers line-level heuristics on the Gopher rules, while Dolma adopts CCNet's language filtering and deduplication. These statistics are useful because regularities such as repetition, compressibility, and lexical composition map directly onto surface manifestations of noise.

Within this layer itself, the picture is fragmented. The four fixed-rule pipelines reviewed in Section 2 target the same family of surface forms through different thresholds and observation units, and the later pipelines build on the earlier rule families. These rule sets overlap on familiar targets but never fully coincide. Taken together, they do not amount to a shared operational definition of noise, nor do they provide a common basis for comparing what their underlying signals detect.

This lack of a shared framework is compounded by two deployment constraints. First, thresholds tuned by trial and error do not transfer cleanly across language distributions. FineWeb 2 makes this cost visible: extending heuristic filtering to more than a thousand languages requires language-specific thresholds and stopword configurations, as well as disabling filters that fail to transfer beyond English (Penedo et al., 2025). Second, prominent pipelines are gated by English language identification, using either langdetect in C4 (Raffel et al., 2020) or fastText scores in Dolma and FineWeb (Soldaini et al., 2024; Penedo et al., 2024). As a result, multilingual and code-switched documents can be excluded before their content quality is assessed.

Adaptive scoring is a nearby alternative, and it runs into the same language problem. DCAD replaces manually specified thresholds with an Isolation Forest over document statistics, but its features are language-specific: it reads tokenization, lexical resources, KenLM models, and language-identification confidence, so a document's score moves with its language as well as with the quality of its text. We treat DCAD as the closest adaptive comparator and report it separately from the fixed-rule pipelines (Sections 2 and 4.3).
Across the fixed-rule pipelines, we observe that surface statistics are operationalized as standalone accept-or-reject rules with fixed thresholds. We instead treat each statistic as a measurement of how a document departs from the statistical structure of natural text. These departures occur along two measurable dimensions: redundancy and lexical composition.

We develop this view in four steps. (1) We collect the surface statistics shipped by seven widely used pipelines and summarize the corpus problems they target as seven corpus-quality failure modes. (2) We construct a controlled benchmark, evaluating every statistic against every failure mode at five dose levels on two substrates. (3) From this benchmark, we build ZenoFilter, a training-free dual-branch detector (Fig. 1), and compare it with the pipelines' own rules on real crawls. (4) We test whether its detection gains transfer to language-model training in a pre-registered 160M-parameter experiment. Together, these steps provide a common framework for benchmarking, combining, and deploying zero-training filtering signals.

ZenoFilter addresses two complementary failure regimes. Its abnormality branch detects documents with substantive content that departs from natural text, whereas its absence branch detects documents with too little substantive content for such departures to be informative. Each branch aggregates direction-aligned zero-training statistics. A single self-scoring pass selects a few hundred documents from the target corpus and uses their moments to calibrate the frozen detector. This label-free surface-statistical pipeline runs entirely on CPUs, from reference construction through corpus scoring.

We evaluate ZenoFilter on real WARC documents across three snapshots against the filtering layers of four mainstream pipelines (Gopher/C4, FineWeb, Dolma, and RefinedWeb) at matched keep rates. On symmetric English, it leaves 2.16–2.63× fewer document-level and 1.74–2.12× fewer token-level residual artifacts across the practical operating range, while running about six times faster. In a separate multilingual deployment analysis, ZenoFilter processes multilingual and code-switched documents without language-specific gates or thresholds; the full-pool comparison leaves 6.16–7.97× fewer document-level residual artifacts and is reported separately because the peer pipelines apply English-language filtering. In downstream validation, ZenoFilter achieves lower held-out loss than FineWeb across six content categories and on the high-quality, source-aligned Nemotron-CC web validation set (Su et al., 2025); this advantage persists on the equal-weight mean of the externally defined Pile-22 suite. An anonymized copy of the detector and the evaluation code is available at `https://anonymous.4open.science/r/zenofilter-5F3C/README.md`.

2 Related Work

We organize prior work by the role each method plays in corpus curation: surface-statistic filtering, adaptive anomaly-based filtering, prior-based corpus selection, and learned quality selection.

Surface-statistic filtering. C4 and Gopher operationalize document quality through fixed rules over punctuation, length, symbol ratios, stop-word fractions, and repetition (Raffel et al., 2020; Rae et al., 2021). RefinedWeb, Dolma, and FineWeb inherit and extend these rule families (Penedo et al., 2023; Soldaini et al., 2024; Penedo et al., 2024), while RedPajama-V2 exposes dozens of document-level quality signals without prescribing a single filtering recipe (Weber et al., 2024). These systems establish surface statistics as practical, inexpensive instruments for corpus curation. Their signals, however, remain embedded in pipeline-specific accept-or-reject rules and are evaluated mainly through aggregate corpus or downstream outcomes. We instead place signals from seven pipelines in a shared operational taxonomy and compare their detection behavior under controlled noise.

Adaptive anomaly-based filtering. DCAD-2000 is the closest adaptive comparison to our work. It replaces manually specified per-feature thresholds with an Isolation Forest over document statistics (Shen et al., 2025). Its features are language-conditioned, so an anomaly score can reflect language identity and resource coverage alongside surface quality. That preference for natural language is structural, and the under-selection of symbol-bearing content cannot be corrected by adjusting thresholds. ZenoFilter instead applies the same direction-aligned surface measurements to mixed input without assigning each document a language. We therefore treat DCAD as a direct adaptive baseline and report it separately from the fixed-rule comparison.

Prior-based corpus selection. A prior-based selector derives a low-cost score from corpus token statistics and selects on distributional fit (Seo et al., 2026). Its observation space differs from the spaces of the fixed-rule pipelines and the adaptive comparator, and we map how far each of the three overlaps the failure forms scored here in Appendix C.5. We include this selector in the comparison as an upstream diagnostic on its own pool.

Learned quality selection. DCLM, FineWeb-Edu, and QuRating learn scores from curated sources or external judgments, primarily to rank documents by utility, educational value, or resemblance to a desired domain (Li et al., 2024; Penedo et al., 2024; Wettig et al., 2024). This objective is adjacent to the failure forms we score. One deployment arrangement we consider runs ZenoFilter first, removing unusable documents, with a learned selector ranking the remainder; the two do not compete.

3 Method

3.1 Overview: From Pipeline Signals to a Detector

We build a training-free detector that converts pipeline heuristics into calibrated surface measurements, allowing heterogeneous failure forms to be scored in a shared observation space without a language gate. We treat each inherited rule as a source of underlying quantities, and ask whether those quantities can be compared and combined without retaining pipeline-specific thresholds.

Starting from the documented rules of existing filtering pipelines, we first define an operational failure space that specifies what is to be detected. We then recast the underlying rule quantities as threshold-free surface operators, organize them by two measurement dimensions, redundancy and lexical composition, and test their responses under controlled transformations. ZenoFilter combines the resulting operators through two decision paths: an abnormality path with separate mainstream and rare-mode scores, and an absence path with a content-absence score. At deployment, it calibrates these scores against an automatically selected, unlabeled reference from the target corpus; Section 4 compares this self-reference with a label-selected alternative.

3.2 Failure Space and Observation Setting

We define the failure space from the discard criteria documented by C4, Gopher, RefinedWeb, RedPajama-V2, Dolma, DCLM, and FineWeb. We read each rule and convert it into the observable form it is intended to exclude. For example, a symbol-ratio rule and a terminal-punctuation rule are merged when both target the same formatting degradation. A candidate failure mode is retained only if it has a distinct observable manifestation, real instances in the source WARC material, and a specified observation unit. This procedure yields seven corpus-quality failure modes: `near_dup`, `empty`, `fmt_degrade`, `mojibake`, `pii`, `toxic`, and `seo`. The source-to-mode mapping and the treatment of language mixture and out-of-scope criteria are documented in Appendix A.1.

We represent each document through a vector of deterministic surface statistics. Let `P` denote a clean-document distribution and let `T_{d,alpha}` transform a document into failure form `d` at within-document intensity `alpha`. The transformed distribution is `R_{d,alpha} = T_{d,alpha}#P`. If the target corpus contains the form at prevalence `pi`, its document distribution is `Q_{d,alpha,pi} = (1-pi)P + pi R_{d,alpha}`. This notation separates the intensity of corruption inside a document from the fraction of corrupted documents in a corpus. No language-identity variable enters this observation model.

These failure modes span two complementary failure morphologies. **Content abnormality** covers substantive documents that depart from ordinary surface structure, whereas **content absence** covers shells, templates, or failed extractions with too little material for local departures to be informative. Sentence-level rules can miss the latter, while length alone misses text-rich but structurally abnormal documents. This distinction motivates the three-score detector.

3.3 Surface Operators and Two Measurement Dimensions

To compare rules across pipelines, we remove each rule's decision threshold and retain only the scalar it tests. Raw counts are normalized by their corresponding observation units, converting counts into rates and pattern matches into densities. We then expand this rule-derived pool by changing one design choice at a time, including the observation unit, span, normalization, or pattern definition. This controlled construction makes changes in operator response attributable to a specific design choice.

We organize the operators by what they measure, yielding two dimensions. **Redundancy** measures how far a text predicts itself. It includes recurrence at token, n-gram, line, sentence, and document levels, together with code length as a finite-text measure of regularity. This use of compression follows the universal-coding view of Ziv and Lempel (1977): repeated substrings reduce the description length. **Lexical composition** measures the balance between content-bearing material and template, markup, boilerplate, or targeted lexical material.

We characterize each candidate on controlled transformations of clean C4 and DCLM documents across seven failure modes and five within-document intensities. The resulting dose-response profiles establish what each operator measures and reveal overlapping response directions. Because controlled transformations cannot reproduce the weak and mixed failures found in real crawls, we additionally evaluate and expand the candidates on naturally occurring WARC documents annotated under the seven-mode scheme. Dose response establishes coverage and direction, whereas leave-out combination tests on the WARC documents determine whether a candidate reduces residual noise beyond the existing aggregate. The complete construction, operator enumeration, and larger-pool replication are detailed in Appendix D.1.

3.4 Response Model and Operator Selection

With the notation of Section 3.2, for an operator `f` with clean mean `mu_f = E_P[f]` and controlled response `Delta_{f,d,alpha} = E_{R_{d,alpha}}[f] - mu_f`, the corpus expectation separates exactly into a prevalence term and a response term:

$$\mathbb{E}_{Q_{d,\alpha,\pi}}[f] = \mu_f + \pi \Delta_{f,d,\alpha} \tag{1}$$

The expected shift therefore separates corpus prevalence from the operator-specific response to a within-document transformation. The five-dose grid traces how `Delta_{f,d,alpha}` evolves with `alpha` for each operator.

The value of a candidate also depends on what the retained operators already measure. For a standardized candidate `z_f` and retained set `S`, let `r_{f|S}` be its residual after affine least-squares projection onto `S` under `P`. With finite second moments, nonsingular retained covariance, and positive residual variance, its contribution to squared Mahalanobis mean separation in the clean covariance geometry is exactly

$$J_{d,\alpha}(f \mid S) = \frac{\left(\mathbb{E}_{R_{d,\alpha}}[r_{f|S}] - \mathbb{E}_P[r_{f|S}]\right)^2}{\mathrm{Var}_P(r_{f|S})} \tag{2}$$

Equation (2) describes the residual mean separation contributed by a candidate under a controlled transformation and provides a common scale for comparing candidates. It is also the increment in squared maximum mean discrepancy (MMD) for a linear kernel whitened by the clean covariance; the exact relation is given in Appendix G.4–G.5.

$$m(f) = |\delta_f| \left(1 - |\rho_{\max}(f, S)|\right) \tag{3}$$

where `delta_f` is the standardized controlled response and `rho_max` is the largest absolute correlation with a retained operator. A larger response and lower overlap raise a candidate's priority. The final inventory combines this screening with normalized quantities from published rules, one-factor extensions, and blind-spot mining on natural WARC documents (Appendix D.1.1).

Controlled transformations reveal how candidates respond to different failure forms, while real WARC documents place those responses in a corpus where weak and mixed failures co-occur naturally. At keep rate `q`, let `Res_q(S)` denote the residual-noise rate obtained with operator set `S`, and define

$$G_q(f \mid S) = \mathrm{Res}_q(S) - \mathrm{Res}_q(S \cup \{f\}) \tag{4}$$

`G_q` is the reduction in residual noise after adding operator `f`. We estimate it over held-out WARC splits and the prescribed keep-rate grid, linking controlled response to filtering gains in natural documents. For document-level residuals under a frozen score on a target mixture with noise prevalence `pi`, let `Gamma_q = TPR_q - FPR_q` at document keep rate `q`. Then

$$\mathrm{Res}_q = \pi - \frac{\pi(1-\pi)}{q}\Gamma_q$$

Thus Equation (4) can be written as the distributional contrast gained by the deployed rule at matched keep. Characteristic-kernel MMD and Wasserstein distance provide complementary views of distributional differences (Appendix G.7–G.8). Implementation and replication details are given in Appendix D.1.

On the development WARC material, each retained operator is oriented toward abnormality. If `a_j` denotes its direction-aligned AUROC, we set the nonnegative weight `omega_j = a_j - 0.5`. This yields 14 abnormality operators, divided into 10 mainstream and 4 rare-mode operators, together with 13 content-absence operators. ZenoFilter is then fixed as a statistical scoring rule: controlled transformations and labeled development WARC data are used once to choose the operator inventory, directions, and weights. Deployment begins after these choices are frozen. At deployment, an unlabeled target-corpus sample supplies the self-reference moments; held-out labels and independent adjudication are used only for evaluation. The complete operator inventory and data-purpose matrix are in Appendices D.1.1 and A.5.

3.5 THREE-SCORE DETECTOR AND LABEL-FREE SELF-REFERENCE

The three scores follow three response directions: mainstream forms, rare-mode forms, and content absence. Mainstream and rare-mode responses differ in strength and prevalence, so separate aggregates preserve their respective rejection paths; content-absence operators form a document-level score for insufficient substantive material. Section 4.4 quantifies what separating the two abnormality layers buys over a single score. The corresponding mean-variance analysis is given in Appendix G.6.

For document `x`, each operator is direction-corrected and standardized with calibration moments. Let `s_j` orient larger values toward the suspicious direction, let `mu_j`$\hat{}${}`cal` and `sigma_j`$\hat{}${}`cal` be the reference moments, and let `epsilon = 1e-3` be the standard-deviation floor:

$$z_j(x) = s_j \frac{f_j(x) - \mu_j^{\text{cal}}}{\max(\sigma_j^{\text{cal}}, \epsilon)}$$

The three scores are

$$A_{\text{main}}(x) = \sum_{j \in M} omega_j z_j(x), A_{\text{rare}}(x) = \sum_{j \in R} omega_j z_j(x), S_{\text{abs}}(x) = \sum_{j \in S} omega_j z_j(x)$$

The final decision is

$$\text{ZenoFilter}(x; k) = \mathbb{1}[\max\{A_{\text{main}}(x), A_{\text{rare}}(x), S_{\text{abs}}(x)\} > k] \tag{5}$$

The weights `omega_j` are the frozen direction-aligned response weights from Section 3.3. Calibration moments are estimated separately within five word-count buckets, and `k` is the corpus-level operating point corresponding to the target keep rate.

Calibration contrast is determined by the reference composition. If a fraction `eta` of the reference carries the same failure form, its mean shifts from `mu_P` to `mu_P + eta Delta`, leaving an affected-to-reference contrast of `(1-eta)Delta`. This attenuation motivates constructing the calibration reference from the target corpus itself. At deployment, an initial random sample of 300 documents supplies provisional moments for scoring an unlabeled target pool. The 60 lowest-scoring documents in each of five length buckets then form a target-matched 300-document reference, from which the calibration moments are estimated once more. Through this single self-scoring pass, the target corpus supplies the reference needed to calibrate the frozen operators, directions, and weights using only unlabeled text. Section 4 evaluates the reference choice separately.

4 EXPERIMENTS

We evaluate ZenoFilter from controlled measurement to downstream use. We first test whether the operators exhibit structured responses across failure forms and doses. We then deploy the fixed detector design on real crawls against four fixed-rule pipelines and DCAD, isolate the contributions of three-score aggregation and label-free self-reference, characterize its observation boundaries, and use a pre-registered study to test whether lower residual artifact rates improve pretraining.

4.1 EXPERIMENTAL SETUP

Evaluation settings. The controlled benchmark crosses seven failure forms with five within-document intensities on clean C4 and DCLM substrates, yielding 70 cells over 46 candidate operators; a held-out FineWeb substrate tests transfer. Deployment uses held-out WARC documents from three Common Crawl snapshots with sentence- and document-level artifact labels.

Comparators. The primary comparison covers the published-threshold layers of Gopher/C4, FineWeb, Dolma, and RefinedWeb on symmetric English. DCAD is evaluated separately as the closest adaptive baseline, using its documented statistics and per-file Isolation Forest. Comparison cohorts are in Appendix A.

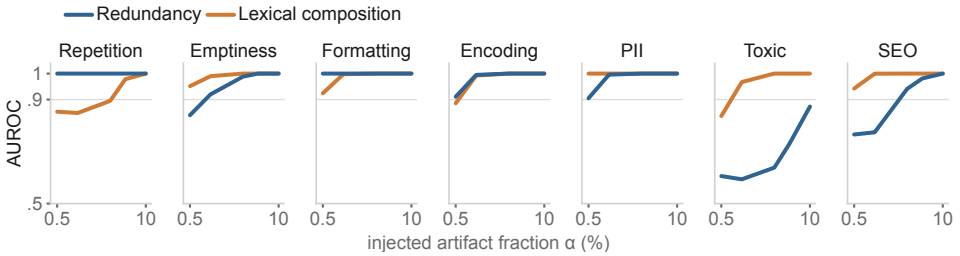

Figure 2: Best direction-adjusted AUROC by measurement dimension across seven corpus-quality failure modes and five injected fractions. The leading dimension varies by failure mode; per-operator results and FineWeb replication are in Appendix A.

Metrics and data separation. We report AUROC for controlled detection and document- and token-level residuals, ratios, and 95% intervals at each comparator's native keep rate; annotator agreement is audited in Appendix C.1. The compared pipelines operate on one shared surface space: the failure forms we score are the ones their documented rules and features target (Appendix C.3–C.5), and the reported ratios are removal capacity attained inside that space. Controlled substrates characterize operator responses; labeled development WARC data selects the inventory, directions, and weights; and the resulting detector is frozen. At deployment, unlabeled target text supplies the one-pass self-reference moments, after which held-out labels and independent adjudication measure performance. Appendix A.5 records label access, purpose, and overlap for every data source.

4.2 Controlled Benchmark

Controlled responses differ by failure form and dose. Figure 2 shows redundancy leading on repetition, lexical composition on emptiness, toxic content, and SEO, and both dimensions responding to formatting, encoding, and PII. Every failure mode reaches AUROC ≥ 0.9 by 3%, with `empty` separating at 0.5–1%; the ordering transfers to FineWeb (Appendix A.6).

Controlled coverage and natural marginal gain serve complementary roles. Six operators reach AUROC ≥ 0.9 in all 70 controlled cells, revealing redundancy in the response map. The final 14+13 inventory adds candidates whose marginal reduction on natural WARC documents distinguishes overlapping controlled profiles through Equation (4).

Natural artifacts occupy a low-separation regime. On natural WARC documents, the best single operator reaches AUROC 0.768 (standardized mean shift about 1.03); expanding the pool from 46 to 110 leaves this ceiling unchanged (Appendix E). This weak, distributed regime motivates the matched-keep deployment evaluation.

4.3 Deployment Comparison

Fixed-rule comparison. On symmetric English, each peer is evaluated at its native keep rate. Peers retain 2.16–2.63 times as many document-level and 1.74–2.12 times as many token-level residual artifacts as ZenoFilter (Table 1). ZenoFilter scores a document in 0.90 ms versus 5.3–5.6 ms for the fixed-rule peers, about six times faster; RefinedWeb is included in this comparison. These symmetric-English results isolate filtering quality under a common language scope; the multilingual full-pool deployment analysis is reported in Appendix B.5, with language-stratified and mixed-language checks in Appendix B.6.

Adaptive comparison. We report adaptive selectors separately from the fixed-rule pipelines. On the symmetric-English evaluation pool of Table 1, at a common keep rate of 0.345 (1,223 documents), the official DCAD implementation leaves a residual artifact rate of 0.2069 versus 0.0932 for ZenoFilter, a 2.22× peer-to-detector ratio. Across the evaluated keep-rate curve, ZenoFilter is no higher than DCAD; the ratio is 2.85× near keep = 0.20 and narrows to 1.03× at DCAD's own operating point, whose default threshold retains 82 per cent of the pool. In the same run on the same pool, ZenoFilter scores a document in 0.86 ms against 8.15 ms for the official DCAD implementation, a 9.5× ratio (Appendix B.4). A prior-based selector occupies a different observation space (Section 2), and we report it as an upstream diagnostic on its own pool, where it leaves 0.3389, or 2.90× the ZenoFilter rate measured on that pool. Implementation details and the feature-to-mode analysis are in Appendix B and Appendix C.

Table 1: Matched-keep deployment on symmetric English ($n = 3{,}546$). Residuals are document-level; token ratios use retained tokens. Full-pool results are in Appendix B.

	GopherC4	FineWeb 1	RefinedWeb	Dolma
Keep rate (the peer's own)	0.329	0.334	0.391	0.401
Peer rule set, residual ↓	0.235	0.215	0.225	0.228
ZenoFilter, residual ↓	**0.089**	**0.089**	**0.104**	**0.105**
Absolute reduction	0.146	0.126	0.121	0.122
95% CI	[.130, .162]	[.110, .141]	[.104, .136]	[.105, .139]
Reduction ↑	**2.63×**	**2.42×**	**2.16×**	**2.16×**
Token-level ratio ↑	**1.96×**	**1.74×**	**2.05×**	**2.12×**

Table 2: Reference ablation at keep 0.134 on the 50k pool; only the calibration reference changes. Relative changes use the deployed self-reference as denominator. Independent-audit intervals and the label-assisted-only audit are in Appendix D.

Reference used for calibration	Residual ↓	vs. self-reference
Self-reference (deployment default; no labels)	**0.046**	—
Label-assisted reference	0.061	+33%
Same-source, unadjudicated reference	0.082	+78%
Independently confirmed-clean subset	0.039	−15%
External curated corpus	0.120	+161%
Reference contaminated at 10% / 30% / 50%	0.090 / 0.092 / 0.202	+96 / +100 / +339%

Independent validation. An independent failure-mode review, separate from the operational annotator's decision boundary, confirms the ranking: across four peer operating points, peers retain 1.32–1.48 times as many confirmed artifacts as ZenoFilter, with all 95% intervals above one (Appendix C.7).

Coverage and stability. ZenoFilter has lower residual on six of seven failure modes and reaches parity on encoding damage, where both residuals are below 0.2%. The ordering persists on claimed-mode and shared-core subsets, under negative controls, across snapshots, and across five seeds (Appendices B and C).

4.4 Component Ablations

No single signal explains the gain. Collapsing the two anomaly layers raises residual from 5.9% to 8.2%; the split saves 27.7% ± 5.6% and replicates at 31.1% on an independent 84k pool. Reweighting recovers only 16%, and the two branches cover complementary failure patterns (Appendix D.3).

Reference construction. With the detector fixed, independent review favors length-stratified self-reference at all four full-pool operating points, reducing residual artifact rates by 0.045–0.052; symmetric-English estimates agree. Table 2 isolates the reference choice at a fixed keep rate. Per-point intervals and audit details are in Appendix D.

4.5 Downstream Validation

Design. Four arms are trained (Table 3); the first three share the same extracted, deduplicated, English base and training budget, and the fourth is a descriptive control that keeps deduplication out of the picture. The pre-registered contrast is ZenoFilter minus FineWeb on held-out validation loss; token matching and the complete-block protocol are in Appendix F.

Results. Table 3 reports the pre-registered FineWeb comparison: at a matched budget, the shared-seed paired contrast places ZenoFilter 0.0858 nats/token below FineWeb. We then test this comparison in two complementary settings. Across the six content categories (books, news, mathematics, encyclopedic text, scientific text, and code), ZenoFilter has lower held-out loss than FineWeb; on the high-quality, source-aligned Nemotron-CC web validation set, the difference is 0.0908 nats/token in ZenoFilter's favor (trainer/bf16 evaluation on 2,048 complete blocks with three shared seeds for the former; independent CPU/fp32 B574 evaluation for the latter). Matched random retention is stronger on some of these primary endpoints, including by 0.0228 nats/token on Nemotron-CC. This reversal motivates the external check below. On the equal-weight mean of the externally defined Pile-22

Table 3: Final held-out validation loss at 160M parameters and 3.2B tokens. Per-arm means use all available seeds. On the three shared seeds, ZenoFilter minus FineWeb is -0.0858 nats/token and ZenoFilter minus matched random retention is $+0.0195$; Appendix F reports the protocol and intervals.

Arm	Seeds	Loss ↓	Role
ZenoFilter matched	3	**3.5936**	confirmatory arm
FineWeb 1	5	3.6771	designated peer
Random matched	5	**3.5530**	supporting control
Deduplication-ignored random	3	3.6379	descriptive control

suite, ZenoFilter remains 0.0549 bpb below FineWeb, while matched random retention is 0.0791 bpb below ZenoFilter. The comparison uses Pythia 160M and 3.2B tokens; interval estimates are reported in Appendix F.

Domain composition. The matched-random result leads us to inspect the fixed Pile-22 components individually. ZenoFilter remains below FineWeb on 18 of 22 components, while the random-retention advantage is concentrated in symbol-bearing content. Across 12 held-out content endpoints, the random-minus-ZenoFilter ordering tracks estimated content coverage (Spearman $\rho = 0.90$ with published markers and 0.93 with strict markers). This preference for natural-language text is a family property: every filter we evaluate on natural-language surface statistics keeps less symbol-bearing material than prose (Appendix F.5). A 1B-token continuation pilot that replenishes code and mathematical text moves the symbolic endpoints in the predicted direction; audit checks and the pilot seed-level limits are in Appendix F.5. On that basis we propose a further deployment step (Section 5).

5 Discussion and Conclusion

What the reference determines. Reference-only ablations separate source match from purity. Among source-matched references, independent confirmation lowers residual from 0.061 to 0.039, whereas injecting artifacts raises it; external curated text, nominally pure, still leaves 0.106–0.120 across two independent draws. Source shift can therefore dominate this gain (Table 2). ZenoFilter constructs its reference from the target corpus itself, using a few hundred unlabeled documents and no deployment labels. Because that reference fixes both the deviation scale and the filtering behavior, matching the source and composition of the target is what brings calibration closest to deployment.

Limits and research directions. Our results characterize surface-form filtering in isolation; its marginal value inside curation stacks with deduplication, language identification, learned quality ranking, and domain reweighting remains to be measured under matched compute. The seven failure modes are grounded in published rules and open-vocabulary audits. The prior-based selector ranks documents by corpus statistics, a different premise and a different selection from ours (Appendices C.5 and F.5). Whether that selection trains a better model remains for training to settle, at a matched token budget. The domain-coverage analysis points to one concrete next step: impose explicit coverage targets for symbol-bearing content on top of ZenoFilter's screen, and compare the resulting corpus against ZenoFilter and matched random retention at the same token budget (Appendix F.5).

Broader interpretation and conclusion. Our comparisons are deliberately restricted to a shared observable surface space: the documented failure forms targeted by the four heuristic rule sets and the part of DCAD's feature space that maps onto those forms. These approaches all encode the surface form of ordinary natural-language web text as an acceptance region, using either fixed thresholds or adaptive anomaly scores. ZenoFilter turns that common premise into measurements that can be compared, combined, and calibrated to a target corpus. It leaves 2.16–2.63$\times$ less residual than the four rule sets at matched keep rates and runs about six times faster; in token-matched 160M training, it also improves over FineWeb on held-out validation. The domain decomposition further shows that downstream composition shifts with how strongly a domain departs from this natural-language reference, most visibly for symbol-bearing content. This links the downstream pattern to the filtering mechanism itself. ZenoFilter therefore provides a training-free, CPU-scale surface-quality layer and a common measurement framework for existing filters.

REFERENCES

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, et al. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Maximilian Idahl, Benedikt Droste, Bjoern Pluester, and Jan Philipp Harries. propella-1: Multi-property document annotation for LLM data curation at scale. *arXiv preprint arXiv:2602.12414*, 2026.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, 2022.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, et al. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

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 *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One pipeline to scale them all — adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Yeongbin Seo, Gayoung Kim, Jaehyung Kim, and Jinyoung Yeo. Prior-based noisy text data filtering: Fast and strong alternative for perplexity. In *International Conference on Learning Representations (ICLR)*, 2026.

Yingli Shen, Wen Lai, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. DCAD-2000: A multilingual dataset across 2000+ languages with data cleaning as anomaly detection. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025. arXiv:2502.11546.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, 2024.

Hyunjin Song, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga, and Hajo G. Boomgaarden. In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, 37(4):550–572, 2020.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Anirudh Subramanyam, Yuxin Chen, and Robert L. Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining. In *International Conference on Learning Representations (ICLR)*, 2026.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. Improving pretraining data using perplexity correlations. In *International Conference on Learning Representations (ICLR)*, 2025.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. RedPajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. arXiv:2411.12372.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. QuRating: Selecting high-quality data for training language models. In *International Conference on Machine Learning*, 2024.

Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

AI USE STATEMENT

Large language models were used in three distinct roles. **(i) Writing assistance.** Drafts of this manuscript were produced with assistance from AI writing tools; the authors reviewed, revised, and take responsibility for all text, including every experimental claim and number. **(ii) Research tooling.** An LLM-based annotator applies the rule set of Section 3.2 to label sentences and documents, and independent LLM judges were used for the audits of Appendix C; both are objects of study rather than aids to the authors, and their validation is reported in Appendix C. **(iii) Code assistance.** Parts of the experimental code were written with AI coding assistants and were reviewed and tested by the authors. No AI system is an author, and no AI system contributed research ideas that are claimed as contributions.

REPRODUCIBILITY STATEMENT

The seven-failure-mode taxonomy, controlled benchmark, evaluation pools, and data-purpose matrix are in Appendix A; the complete deployed operator inventory is in Appendix D.1.1; the peer reimplementation audit, fairness variants and throughput protocol are in Appendix B; the label-validation and coverage analyses are in Appendix C; signal mining and ablations are in Appendix D; negative results and boundaries are in Appendix E; the downstream protocol is in Appendix F; and the statistical derivations behind Equations (1)–(3) are in Appendix G. The operator inventory, directions and weights are frozen on disjoint development data before deployment. Calibration consumes only per-signal moments of about three hundred unlabeled target-corpus documents, and the deployment threshold implements the predefined target keep rate.

A BENCHMARK, POOLS, AND THE SEVEN FAILURE MODES

A.1 DERIVING THE FAILURE MODES

We read the quality rules that seven pretraining-corpus filtering pipelines document, using Albalak et al. (2024) as a reference for the broader data-selection taxonomy. Each rule is read at the level of the failure form it intends to exclude rather than at the level of its implementation: Gopher's symbol-ratio rule and C4's terminal-punctuation rule are implemented differently but target the same observable form, and are merged. A candidate failure mode is retained only if it has an independent manifestation, real instances in the WARC material the pipelines themselves use, and a specified unit of observation. Three components fall outside the resulting failure modes and are excluded on stated grounds: language identification is a selection criterion rather than a failure form; domain blocklists judge provenance rather than text; and educational-value scoring is quality ranking rather than form identification. Two earlier failure modes were merged into the final seven: mixed-language material is dominated in WARC by low-information multilingual spam and folds into SEO.

A.2 EVALUATION POOLS

The early and extended pools span three Common Crawl snapshots, the extended pool adding six further WARC files; cross-snapshot stability is checked on 2025-30 against 2025-51. Pool sizes and labeled artifact rates are 31k at 0.347, 84k at 0.347, 533k at 0.435, 10k at 0.437, and the 50k analysis pool at 0.439. Failure-mode composition on the analysis pool, in documents, is near-duplication 5,918, formatting degradation 5,793, emptiness 4,083, personal information 2,775, SEO 1,805, encoding damage 846 and toxic vocabulary 713.

Table A.1. Pool sizes and labeled artifact rates across five pools spanning three Common Crawl snapshots, with cross-snapshot stability checked on 2025-30 against 2025-51, and failure-mode composition on the analysis pool.

Pool	Documents	Artifact rate
Early	31k	0.347
Extended	84k	0.347
Large-scale labeling	533k	0.435
Evaluation	10k	0.437
Analysis	**50k**	**0.439**

Failure mode (analysis pool)	Documents
Near-duplication	5,918
Formatting degradation	5,793
Emptiness	4,083
Personal information	2,775
SEO	1,805
Encoding damage	846
Toxic vocabulary	713

A.3 CONTROLLED RESPONSE MAP

Figure A.1 displays the complete response map underlying the dimension-level summary in Figure 2. It reports direction-adjusted AUROC for every retained operator across seven failure modes and five doses. The block structure shows redundancy within response families and a change in the leading family across failure modes.

A.4 MINIMUM COVER

A forward greedy cover with tolerance 0.02, counting a cell as covered when an operator reaches within that tolerance of the best signal for the cell, needs six operators for the seventy main cells. The cover is not unique: direction-aligned clusters, one of compression-family signals and one of type-token ratios, make members interchangeable, and two greedy runs select different representatives of the same clusters.

A.5 DATA LIFECYCLE AND LABEL ACCESS

Table A.2 separates the five roles played by data in detector construction, deployment calibration, and evaluation. Controlled transformations and labeled development WARC documents determine the fixed measurement rule. Deployment calibration then uses only an unlabeled, source-matched target sample. Operational labels and independent adjudication are introduced after scoring and are used only to estimate performance. In the reported experiments, development, self-reference, and

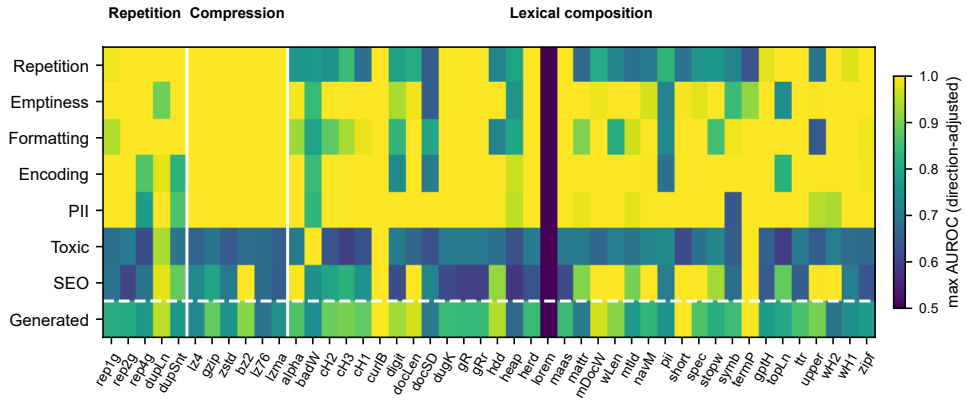

Figure A.1. Complete direction-adjusted AUROC response map across seven corpus-quality failure modes, five doses, and the retained surface operators. Blocks expose within-family redundancy and changes in the leading measurement family across failure modes.

held-out evaluation documents are hash-disjoint; independent adjudication intentionally resamples strata from the held-out evaluation sets.

Table A.2. Data-purpose matrix. Labeled development data fixes the statistical scoring rule; deployment calibration uses only unlabeled target text, and both evaluation channels enter after scoring.

Stage	Material	Label access	Purpose	Relation to evaluation
Controlled response	Clean C4/DCLM plus injected transforms	Failure mode and dose known	Characterize response, direction, and overlap	Separate substrates
Development WARC	Natural crawl documents	Seven-mode development labels	Select inventory, directions, and weights; then freeze	Hash-disjoint
Target self-reference	Random target-corpus sample	None	One-pass length-stratified calibration moments	Source-matched; hash-disjoint in experiments
Held-out deployment	Natural WARC documents	Operational labels after scoring	Matched-keep residuals and operating curves	Disjoint from development and reference
Independent audit	Retained-set strata	Independent per-mode adjudication after scoring	Validate ranking and criterion-reuse effects	Resampled from held-out evaluation

A.6 A.6 Held-out substrate transfer

The controlled ordering replicates on the held-out FineWeb substrate. We retain the cell-level AUROCs and intervals from `F1_matrix8_fineweb_results.json` and `F1_15_fineweb_validation_results.json`; these results test transfer of response ordering and are reported separately from the C4 and DCLM primary matrix.

B Fairness Variants, Robustness, and Cost

B.1 Rigor devices

Directions, weights and the threshold are fixed constants (B.1.1). Negative controls include scoring only clean reference documents, which produces no separation by construction, and within-document token shuffling, which collapses every response toward the clean reference and confirms that responses depend on how material is arranged rather than on aggregate composition. Transfer is calibration transfer because the detector fits nothing: two-way cross-snapshot re-runs preserve the advantage and frozen-seed re-runs reproduce it in five of five seeds. Improvement intervals use one thousand bootstrap resamples of the 50k pool.

B.1.1 Direction sensitivity

The shell branch's thirteen signal directions are fixed constants, each set to the sign of the signal's association with the artifact label measured on a pool disjoint from the evaluation set (direction-invariant ROC-AUC, Table B.1). The AUC magnitudes are reported to make the margin explicit; the least secure directions sit closest to the 0.5 boundary (`placeholder_kw_density` **0.514**,

stopword_frac **0.461**). To test sensitivity to these choices, we recomputed the main-table deployment comparison with all six least-secure directions flipped (Table B.1). The advantage persists at $\geq$**1.78**$\times$ for every peer arm, versus 2.16–2.63$\times$ under the signed directions; on the full-pool caliber the same comparison drops from 6.16–7.97$\times$ to at worst 5.14$\times$. The direction choice therefore changes the magnitude of the reported ratios but cannot manufacture the advantage, and the direction values are recorded with the code for reproduction.

Table B.1. Thirteen fixed signal directions and direction-invariant ROC-AUC on the training pool ($n = 111{,}805$, disjoint from the evaluation set). The two least-secure directions sit closest to the 0.5 boundary.

Signal	Dir.	AUC	Signal	Dir.	AUC
doc_len	-1	0.494	newline_density	$+1$	0.748
lines_per_char	$+1$	0.741	stopword_frac	-1	0.461
ttr	-1	0.393	nav_kw_density	$+1$	0.545
template_kw_density	$+1$	0.600	boilerplate_kw_density	$+1$	0.542
line_char_mean	-1	0.259	placeholder_kw_density	$+1$	0.514
sents_per_char	-1	0.387	menu_line_frac	$+1$	0.562
			sentence_completeness	$+1$	0.611

Table B.2. Symmetric-English residual ratios at matched keep when the six least-secure signal directions are flipped. The advantage persists at $\geq 1.78 imes$ for every peer arm.

Arm	Signed directions	Six least-secure flipped
FineWeb 1	2.42	**1.80**
Gopher-C4	2.63	**1.97**
Dolma	2.16	**1.78**
RefinedWeb	2.16	**1.81**

B.2 Fairness variants

Four protocol variants test whether the advantage depends on the comparison protocol. Length-bucket calibration removes the length bias a global ranking would carry. Under matched false-positive rate, with quantile scans on both sides so that neither side benefits from a looser threshold, the advantage persists; we note that two learned baselines collapse to zero scores at document level, with zero fractions between 0.55 and 0.91, so a fixed false-positive-rate scan is undefined for them and we mark them degenerate rather than report a detection rate of zero as if it were measured. Under symmetric granularity, both sides take their best configuration at the same granularity. Under two extraction settings, the advantage is preserved in both. A threshold-independent comparison is reported separately below because it asks a different question.

B.2.1 Length-stratified residuals

The deployment panel matches filters by retained document count, a criterion that does not control document length. DCAD's first official feature is the word count, and at a matched count the filters retain different length mixes, so we recompute every keep rate on the fixed axis under five word-count quintiles and ten within-pool deciles, standardizing each kept set to the ranked pool's own length mix; strata holding fewer than five retained documents are dropped, and the covered length weight is reported. Three properties follow. The standardized value depends on the number of strata and need not move monotonically (ZenoFilter at keep 0.050 gives 0.0394 with five strata and 0.0170 with ten; DCAD at the same budget gives 0.1508 and 0.1809). Standardizing moves our own curve as well, and by a large amount at tight budgets (ZenoFilter at keep 0.100 goes from 0.0451 to 0.0718). And the tight end is only partly identified for DCAD, whose keep 0.050 row rests on 80 per cent of the length weight. At the working point the effect is small and the ordering is unchanged: at keep 0.345 ZenoFilter moves from 0.0932 to 0.0968 while DCAD moves from 0.2069 to 0.2068 under five strata, a ratio of 2.14; under ten deciles the pair is 0.1003 against 0.2020. We therefore report the panel raw and treat the length question as a sweep, because a single standardized curve would hide the rung it was drawn on. At the peer's own retained count the same standardization moves FineWeb from 0.2145 to 0.2073 and ZenoFilter from 0.0887 to 0.0921.

Table B.3. Length-stratified residual artifact rates on the fixed deployment axis. Each keep rate is recomputed unstratified and standardized to the ranked pool's word-count mix with five quintiles and ten within-pool deciles; strata holding fewer than five retained documents are dropped. The DCAD rows at keep 0.100 and 0.050 cover 80 per cent of the length weight, all other rows 100 per cent. The standardized value depends on the number of strata and need not move monotonically, and it moves the detector's own curve as well; the working point is the least affected, where the ordering is unchanged.

		ZenoFilter			DCAD		
Keep	k	raw	quint.	dec.	raw	quint.	dec.
0.895	3,174	0.3812	0.3812	0.3812	0.3812	0.3812	0.3812
0.700	2,482	0.2434	0.2434	0.2431	0.2740	0.2748	0.2743
0.500	1,773	0.1325	0.1337	0.1344	0.2155	0.2140	0.2111
0.345	1,223	**0.0932**	0.0968	0.1003	**0.2069**	0.2068	0.2020
0.200	709	0.0677	0.0884	0.0906	0.1932	0.2126	0.1993
0.100	355	0.0451	0.0718	0.0729	0.1915	0.1992	0.2052
0.050	177	0.0169	0.0394	0.0170	0.1412	0.1508	0.1809

B.3 Threshold-independent discrimination

On the same pool of 332,843 documents, direction-invariant ROC-AUC is 0.906 for the deployed detector and 0.716 for the strongest single baseline we evaluate at that setting. AUC is direction-invariant here, that is, the maximum of the value and its complement.

Figure B.1 shows the corresponding ROC curves. ZenoFilter dominates the comparator throughout the practically relevant false-positive range, so the matched-keep result is not tied to one isolated threshold.

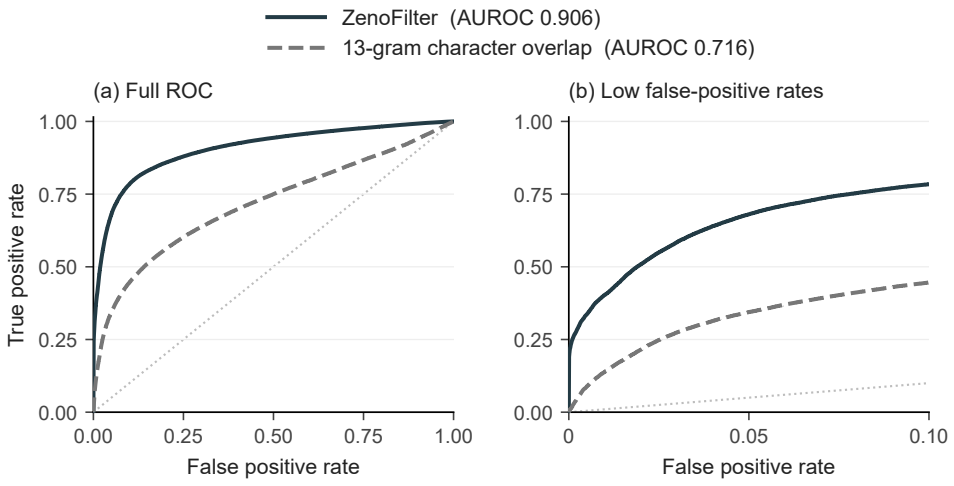

Figure B.1. Threshold-independent ROC comparison on 332,843 held-out documents. ZenoFilter reaches direction-invariant AUROC 0.906 versus 0.716 for 13-gram character self-overlap.

B.4 Cost

Measured on one machine over the same five-thousand-document sample, the detector scores a document in **0.90 ms**, of which **0.894 ms** is signal computation in C and 0.002 ms is standardization and aggregation; it requires neither a tokenizer nor a language-identification stage because its signals are character- and pattern-level. A single peer pipeline's full rule set costs 5.29 to 5.61 ms per document, of which **4.53 ms** is the tokenization its word-level rules require. The ratio is therefore between **5.9 and 6.3**. The comparison is conservative for the detector: the peer rule sets short-circuit, rejecting a document at the first failing rule, so that at Dolma's keep rate roughly five sixths of documents never reach the expensive repetition scan, whereas the detector computes every signal on every document. Calibration costs about one core-second for three hundred reference documents.

Cost of the adaptive comparator. On the symmetric-English evaluation pool, in the same run that produced the residual comparison of Section 4.3, the official DCAD-2000 implementation costs a median **8.15 ms/doc** against **0.86 ms/doc** for ZenoFilter, a **9.5**× ratio. DCAD's cost is dominated by repeated SentencePiece tokenization, GlotLID, and the flagged-word resource; ZenoFilter runs its character-level signals without language identification or model inference. An earlier run on a 9,748-document multilingual sample measured 11.77 ms/doc against 0.61 ms/doc (19.3×); that sample and its tokenisation path differ from this one, and the two ratios are not interchangeable.

Table B.4. Measured per-document cost on one machine over the same five-thousand-document sample. The comparison is conservative: peer rule sets short-circuit while the detector computes every signal on every document.

Component	Scope	Time per document
ZenoFilter signal computation (C)	all 46 signals	0.894 ms
ZenoFilter standardization + aggregation		0.002 ms
ZenoFilter total	no tokenizer, no language ID	**0.90 ms**
Peer rule set (any of the four)	full published rule list	5.29–5.61 ms
of which tokenization (word-level rules)		4.53 ms
Speedup	conservative: peers short-circuit	**5.9–6.3**×
Calibration	300 reference documents	≈1 core-second

B.5 Full-pool deployment

The main text reports the symmetric-English comparison, where document- and token-level residuals isolate filtering quality under a common language scope. We report the multilingual full-pool comparison over 10,000 documents separately as a deployment analysis: ZenoFilter processes multilingual and code-switched documents without language-specific gates or thresholds, whereas the peer pipelines apply English-language filtering. The full-pool results are therefore reported at document level only in Table B.1; token-level ratios are reported on symmetric English, the peers' working caliber (Table 1). FineWeb 2 appears only in this deployment caliber. At every peer's native keep rate the detector retains **6.00–7.97**× **fewer residual artifacts** per document (FineWeb 2: 6.00×).

Table B.5. Full-pool matched-keep deployment at document level ($n = 10{,}000$); FineWeb 2 appears only in this caliber. The multilingual pool mixes languages that the peer pipelines are not designed to filter, so this caliber is reported at document level only; token-level ratios are reported on symmetric English, the peers' working caliber (Table 1).

Arm	Keep rate	Peer resid.	ZenoFilter resid.	Reduction [95% CI]	Ratio
FineWeb 1	0.125	0.214	**0.029**	0.185 [.176, .194]	**7.42**×
Gopher-C4	0.123	0.234	**0.029**	0.204 [.195, .213]	**7.97**×
Dolma	0.149	0.227	**0.037**	0.190 [.180, .199]	**6.16**×
RefinedWeb	0.139	0.225	**0.031**	0.194 [.184, .202]	**7.23**×
FineWeb 2	0.109	0.160	**0.027**	0.134 [.124, .143]	**6.00**×

B.6 Language-stratified and mixed-language checks

The canonical 10,000-document pool contains 3,546 high-confidence English documents, 5,570 non-English monolingual documents, and 303 mixed-language candidates. All documents receive finite scores (coverage 1.000). The four English peer filters retain only 0.0–1.7% of the non-English group, so the full-pool ratios in B.5 combine filtering with language coverage. FineWeb 2 is the only peer with a non-English operating point; under the annotator labels, ZenoFilter leaves 0.034 residual versus 0.160 for FineWeb 2 (4.65×). An independent multilingual audit of 2,132 documents gives per-language ratios spanning 0.67 to 2.17, every one of them smaller than that audit's noise, so no language shows a resolved difference in either direction. These strata establish scoring coverage and expose language-dependent calibration, especially for English vocabulary signals, and they explain the language composition of the full-pool deployment result. The same cohort supports a direct comparison on disputed documents. Where the detector and a peer disagree, documents kept only by the detector carry an issue rate of 0.192 against 0.451 for those kept only by the peer (difference +0.258, 95% CI [+0.174, +0.343]), and every peer's interval excludes zero; on the chunk-level mixed cohort

the same contrast is 0.258 against 0.446 (+0.188, CI [+0.057, +0.319]). On that cohort, detector AUROC against the final seven-axis labels is 0.822 [0.778, 0.865], against 0.748 [0.695, 0.795] for an English monolingual control and 0.761 [0.736, 0.786] for a non-English monolingual control, all above chance; the direct differences are +0.069 [+0.006, +0.134] and +0.055 [+0.008, +0.102]. The three English hard-rule peers reach binary discriminative power of only 0.578–0.587 on this cohort, against 0.637–0.653 for the detector binarized to each peer's own keep rate, with same-document paired differences of +0.050, +0.069, and +0.071, all three intervals excluding zero. The frozen V0 calibration was retained after the source-disjoint Pilot-2 comparison, while all candidate multilingual repairs were stopped by the pre-registered gate.

C Label Validation and Coverage

C.1 Agreement

Automated annotation must be validated on the task where it is used (Pangakis et al., 2023; Song et al., 2020). On a hundred-sentence gold standard sampled independently of detector development, sixty carrying an artifact and forty clean, adjudicated by the authors, the rule annotator reaches Cohen's kappa 0.720 with precision 0.960 and recall 0.800. An independent LLM re-audit of the same sentences with full document context reaches **kappa 0.811** with **precision 0.946** and **recall 0.898**, the disagreements concentrating on repetition and formatting fragments, the two categories that require document context to judge. Document-level audits that hand a judge a whole document and no rule prompt agree far less, between 0.23 and 0.30 across three annotator calibers. Two effects bound that number. A document can carry several failure modes at once, so two audits that select different modes from the same document disagree although neither is wrong; and the same judge's own artifact rate moves by twelve points between rounds on identical material, since attainable agreement depends on how the judgment is elicited (Wettig et al., 2024). We therefore anchor precision and recall to the gold standard and use those audits only for the aggregate level.

Table C.1. Label quality audited from both sides. The first two rows test whether the annotator applies its own failure modes correctly; the next three ask an independent judge for free descriptions with no category vocabulary; the last row is bounded by the judge's own drift and is used only at the aggregate level.

Check	Agreement	What it establishes
Human gold standard, 100 sentences	$\kappa = 0.720$	$P = 0.960$, $R = 0.800$
Independent LLM re-audit, full context	$\kappa = 0.811$	$P = 0.946$, $R = 0.898$
Open-vocabulary, 92 flagged docs	90.2% confirmed	every description maps to a failure mode
Open-vocabulary, 300 clean-side docs	objects in 68.8%	81.5% of that maps back to a failure mode
Open-vocabulary, problem-free check	12.3%	no failure mode above 25%
Document-level, no rule prompt, 600 docs	$\kappa = 0.23$–0.30	judge's own rate drifts 12 points

C.2 Open-vocabulary audit

The checks above show that the annotator applies its own failure modes correctly; they cannot detect a failure form outside them, since a rule that does not exist cannot fire. We therefore gave documents to an independent judge with no category vocabulary, asking only for a free description of what was wrong, and mapped the descriptions onto failure modes afterwards by a procedure with no access to the annotator's rules. On ninety-two documents the annotator had flagged, the judge confirmed 90.2 per cent and every description mapped onto an existing failure mode or a sub-type. Applied to three hundred documents the annotator calls clean, stratified by length, the judge objects in 68.8 per cent of them, but 81.5 per cent of what it objects to maps back onto an existing failure mode, consistent with the measured recall. Run on the annotator's positives, the same protocol finds 12.3 per cent problem-free with no failure mode above 25 per cent. We disclose a sampling defect in the first of these audits: a script error kept twenty clean documents out of the sample, so that audit measures a confirmation rate on known positives rather than a detection rate.

C.3 CLOSURE

On the 84k pool all 181,782 labeled sentences and all 29,302 labeled documents fall inside the seven failure modes with no residual other category, and reverse-mapping every artifact-screening rule the surveyed pipelines document places all of them inside the taxonomy. DCAD and the prior-based selector sit outside this closure count, since both produce corpus-relative selection scores and neither documents a per-defect discard rule. Their feature mappings show which of the seven failure modes their observable features can respond to (C.5); the exhaustiveness claim rests on the rule sources and the open-vocabulary audit above.

C.4 FIXED-RULE COVERAGE

Counting only rules that discard a document for the failure mode in question, FineWeb 1 covers three of the seven failure modes, Gopher/C4 four, Dolma four and RefinedWeb three; counting indirect handling as well the counts are **five, six, six and four** (Table C.2). All four cover near-duplication, emptiness and formatting degradation explicitly. FineWeb 1 and Dolma specify PII patterns but apply them to redact matches rather than to discard the document, so documents carrying personal information enter their retained corpora; we therefore count that as indirect. No pipeline treats SEO material with a content-level discard rule. Cross-document deduplication is treated as a shared preprocessing prefix rather than as detection power; it is applied by every pipeline we compare and its effect on model quality is established separately (Lee et al., 2022).

Table C.2. Failure-mode coverage of the surveyed pipelines' published filtering layers (✓ = content-level discard rule, ∼ = indirect handling such as redaction or a classifier used for ranking rather than discarding). ZenoFilter covers all seven failure modes; the closest peer covers four directly and six including indirect handling.

Failure mode	Gopher/C4	FineWeb	Dolma	RefinedWeb	FineWeb 2	DCLM	ZenoFilter
Near-duplication	✓	✓	✓	✓	✓	✓	✓
Emptiness	✓	✓	✓	✓	✓	✓	✓
Formatting degradation	✓	✓	✓	✓	✓	✓	✓
Toxic vocabulary	✓	—	✓	—	∼	—	✓
SEO (+low-quality multiling.)	∼	∼	∼	∼	∼	—	✓
Personal information	—	∼	∼	—	∼	—	✓
Encoding damage	∼	—	—	—	—	—	✓
Failure modes claimed (✓)	4/7	3/7	4/7	3/7	3/7	3/7	**7/7**
With indirect handling	6/7	5/7	6/7	4/7	6/7	3/7	**7/7**

C.5 ADAPTIVE-SELECTOR FEATURE MAPPING AND FAILURE-MODE DIAGNOSTIC (SEPARATE FROM FIXED-RULE COVERAGE)

DCAD's official implementation combines eight document features with a per-file Isolation Forest. We map a feature to a failure mode only when the correspondence is observable in the documented feature or target: **explicit** means that the method describes the phenomenon as a target or problem; **indirect** means that a feature can respond to it without defining it as a separate defect class; **none** means that no defensible correspondence is available. This is the adaptive-selector counterpart to the fixed-rule coverage matrix in C.4. It is a diagnostic of the selector's observable feature space, not additional evidence that the seven failure modes are exhaustive.

DCAD is therefore reported as an adaptive comparator, separately from the fixed-rule pipelines. Both methods fit scores relative to the target corpus, but their feature spaces and decision rules differ. On the symmetric-English evaluation pool at keep = 0.345 (1,223 documents), DCAD leaves 0.2069 residual artifacts and ZenoFilter 0.0932 (a 2.22× ratio); on the same pool and in the same run, DCAD costs 8.15 ms/doc against 0.86 ms/doc for ZenoFilter (9.5×). DCAD's feature-level notation aversion is reported with the adaptive comparison.

The prior-based selector joins this adaptive group as well. Its starting point differs from every other comparator in this paper: it ranks documents by how well they fit corpus token-frequency statistics, not by their surface-statistical features. We read its matched-retention result at the selection layer, alongside the other adaptive comparator.

For the claimed-mode analysis, we retain DCAD's qualitative correspondences: `toxic` is an explicit target, while `near_dup`, `fmt_degrade`, and `mojibake` are indirect correspondences. We do not report a separate independent per-mode adjudication ratio for DCAD; the 1.32–1.48 peer/ZenoFilter intervals in C.7 therefore remain restricted to the four fixed-rule operating points and are not extended to DCAD.

Table C.3. Feature-to-failure-mode mapping for the official DCAD selector. ✓ denotes an explicit target, ∼ an indirect observable correspondence, and — no defensible mapping. This diagnostic table is excluded from the fixed-rule coverage counts because DCAD produces a corpus-relative anomaly score rather than a documented per-defect discard rule.

Failure mode	DCAD	Feature-level basis
Near-duplication	∼	`word_repetition` responds to repetition but is not a dedicated near-duplicate rule
Emptiness	—	No content-absence feature
Formatting degradation	∼	Repetition and special-character features can respond to formatting anomalies without naming the target
Encoding damage	∼	`char_repetition` has a weak correspondence to the documented encoding-artifact concern
Personal information	—	No PII feature among the eight inputs
Toxic vocabulary	✓	`flagged_words` corresponds directly to the vocabulary-filtering target
SEO	—	No independent search-bait or template-SEO criterion

C.6 External construct alignment

Rule-source coverage and the open-vocabulary audit establish the taxonomy within the surface-failure space. We add a convergent-validity check against Propella-1, an independently defined 18-property annotation scheme (Idahl et al., 2026). Eleven ordinal properties can be compared with our surface measurements after excluding content length, whose relationship is circular. On 49,361 Nemotron web documents, length-controlled partial correlations form a structured map: educational value aligns most with compression (rho = -0.226), reasoning indicators with document structure (-0.283), and information density with lexical composition (-0.221). Safety and PII effects are weak or unstable under full confounder control. The same compression mechanism does not transfer unchanged to FinePDFs, where educational and reasoning correlations are carried by uppercase structure and compression loses significance after entropy control. Figure C.1 therefore provides external construct alignment while also locating semantic and corpus-specific boundaries; the completeness argument continues to rest on source coverage and empirical closure.

C.7 C.7 Claimed-mode subset and independent review

Restricting the comparison to failure modes explicitly claimed by each peer leaves a 2.1–4.2-fold residual-artifact advantage for ZenoFilter; the shared-core subset has the same direction. An independent per-mode review gives peer/ZenoFilter confirmed-artifact ratios of 1.32–1.48 across four peer operating points, with all four 95% intervals above one. These comparisons use different cohorts and levels from the primary table and are not pooled with its ratios.

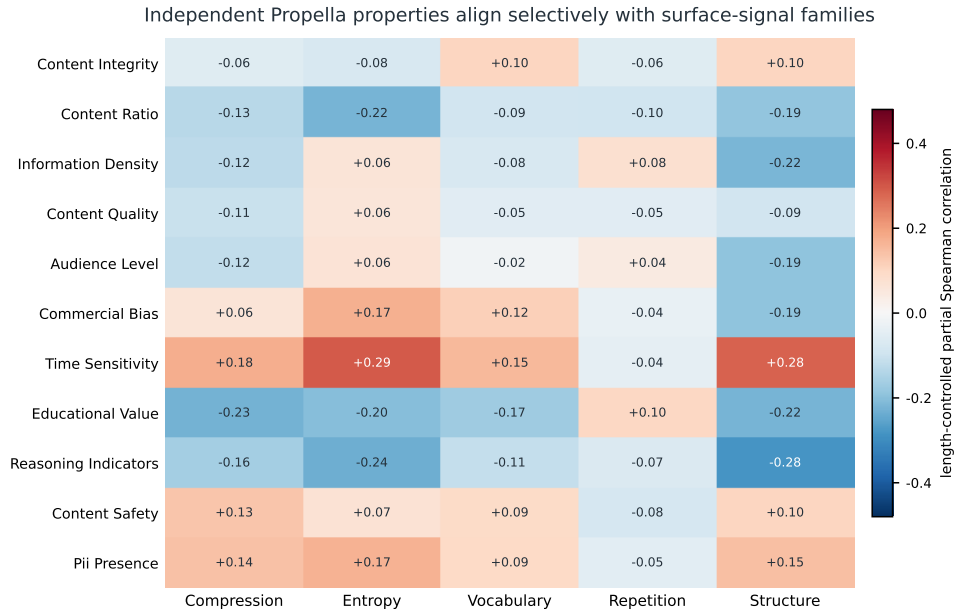

Figure C.1. External construct alignment on 49,361 Nemotron web documents. Each cell is the largest absolute length-controlled partial Spearman correlation within a surface-signal family. The structured, selective pattern supports convergent validity without treating semantic properties as additional failure modes.

D Signal Mining and Ablations

D.1 Mining

Candidate signals are enumerated by replacing one design dimension at a time and holding the others fixed, first on a 31k pool to establish signal families and then on a 52k pool to re-mine the sparse modes, where samples are between 2.6 and 2.9 times more numerous; the second pass reproduces the ordering of the first. Sentence-level repetition rate reaches per-type AUROC 0.936 against a previous pool best of 0.880. Bigram and trigram repetition ratios reach 0.753 and 0.749 and cover the 87.3 per cent of weakly repetitive documents that a line-duplication score below 0.01 misses, which is the case for mining: the quantity the pipelines themselves record barely fires in the weak-repetition regime that dominates real crawl material.

D.1.1 Complete deployed operator inventory

Tables D.1 and D.2 list the exact 27 operators used at deployment. Direction is the implemented sign before weighting, with + meaning that larger standardized values increase the branch score and − meaning that smaller values do. Weights are the frozen direction-aligned AUROC margins from the disjoint development pool, rounded to three decimals here; exact constants are released with the implementation. Provenance distinguishes normalized published-rule quantities, one-factor operator extensions, and operators added by natural-WARC blind-spot mining. None of these quantities is fitted on the deployment corpus.

D.2 The two-layer split

Collapsing the two layers into a single fourteen-signal aggregate raises total residual from 5.9 to 8.2 per cent, a relative reduction from the split of **27.7 per cent** with a two-standard-deviation interval of 5.5 points over one thousand resamples, reproduced at **31.1 per cent** on an independent 84k pool. The per-mode pattern is what a maximum merge predicts: toxic vocabulary improves by **89 per cent**, SEO by **45**, encoding damage by **40**, emptiness by **9** and formatting by **8**, while near-duplication gives back **17 per cent**. Sparse modes otherwise diluted inside one sum gain most and the densest mode gives a little back. Reweighting within a single layer is not a substitute and recovers only **16 per cent**.

Table D.1. The fourteen abnormality operators used at deployment. Direction is the implemented score sign; weights are frozen AUROC margins rounded to three decimals.

Operator	Definition and unit	Dir.	Branch	Weight	Provenance
repeat_2gram_rate	Repeated word-bigram occurrences / all word bigrams	+	Main	.172	Rule normalization
repeat_4gram_rate	Repeated word-4-gram occurrences / all word 4-grams	+	Main	.218	Rule normalization
lz4_ratio	LZ4-compressed bytes / UTF-8 bytes	+	Main	.123	One-factor extension
dup_line_frac	Occurrences in repeated normalized lines / nonempty lines	+	Main	.210	Rule normalization
gzip_ratio	Gzip-compressed bytes / UTF-8 bytes	+	Main	.068	One-factor extension
char_entropy_unigram	Shannon entropy over Unicode code points (bits)	+	Main	.179	One-factor extension
bigram_repeat_ratio	Nondistinct alphabetic bigrams / all alphabetic bigrams	+	Main	.156	WARC mining
trigram_repeat_ratio	Nondistinct alphabetic trigrams / all alphabetic trigrams	+	Main	.188	WARC mining
dup_sent_frac2	Occurrences in repeated sentences / all sentences	+	Main	.167	WARC mining
title_line_frac	Short uppercase-initial lines / nonempty lines	+	Main	.087	WARC mining
pii_regex	Email and phone matches per million characters	+	Rare	.078	Rule normalization
badwords_rate	Toxic-lexicon matches / word tokens	+	Rare	.021	Rule normalization
seo_vocab_frac	Distinct SEO entries observed / 20-entry lexicon	+	Rare	.003	WARC mining
mojibake_frac	Nonoverlapping mojibake-sequence matches / characters	+	Rare	.022	Rule normalization

Table D.2. The thirteen content-absence operators. All belong to the absence branch; direction and weight follow the same convention as Table D.1.

Operator	Definition and unit	Dir.	Weight	Provenance
doc_len	Unicode character count	−	.063	Blind-spot mining
lines_per_char	Nonempty lines / characters	+	.200	Blind-spot mining
ttr	Distinct alphanumeric/apostrophe tokens / tokens	−	.051	Blind-spot mining
template_kw_density	Distinct template-phrase hits / characters	+	.052	Blind-spot mining
line_char_mean	Mean raw characters per nonempty line	−	.200	Blind-spot mining
sents_per_char	Sentence segments / characters	−	.094	Blind-spot mining
newline_density	Newline characters / characters	+	.205	Blind-spot mining
stopword_frac	Stopword matches / alphanumeric-apostrophe tokens	−	.055	Blind-spot mining
nav_kw_density	Navigation entries observed / navigation lexicon	+	.028	WARC extension
boilerplate_kw_density	Boilerplate entries observed / boilerplate lexicon	+	.037	WARC extension
placeholder_kw_density	Placeholder entries observed / placeholder lexicon	+	.009	WARC extension
menu_line_frac	Menu-like lines / nonempty lines	+	.041	WARC extension
sentence_completeness	One minus terminal/segment completeness	+	.101	WARC extension

Table D.3. Cost of collapsing the two-layer detector into a single fourteen-signal aggregate. Rare modes diluted inside one sum gain the most; the densest mode gives a little back.

Failure mode	Residual change when the two layers collapse to one sum
Toxic vocabulary	+89%
SEO	+45%
Encoding damage	+40%
Emptiness	+9%
Formatting degradation	+8%
Near-duplication	−17%
Total residual (50k pool)	**5.9% → 8.2%; the split saves 27.7% ± 5.5**
Total residual (84k pool)	**the split saves 31.1%**
Reweighting within one layer	recovers only 16%

D.3 Aggregation variants

At keep rate 0.5, per-mode maximum leaves residual 0.201 against 0.254 for a plain sum; response weighting improves on uniform weighting by a further 5 to 9 per cent; truncated sums and an earlier two-branch variant are strictly worse.

D.4 A trained ceiling

Abandoning the zero-training constraint, a trained likelihood-ratio classifier reaches AUROC 0.96 on the same material. We report it to mark the headroom outside the constraint; it is not our method and does not enter any comparison.

D.5 Reference construction

The deployed reference is obtained by one length-stratified self-scoring pass over a random target-corpus sample. With the detector held fixed, independent per-mode adjudication compares this self-reference against a label-assisted reference at the same four peer operating points. On the multilingual full pool, self-reference reduces residual by 0.045-0.052 at every point and all intervals exclude zero. On symmetric English, all four point estimates also favor self-reference by 0.009-0.017; the Gopher-C4 interval excludes zero and the other three end within 0.003 of it. Error decomposition localizes the reversal to label-assisted-only retained documents, whose annotator miss rate is 0.23-0.36. Five independent reference draws reproduce the ranking with residual standard deviation at most 0.005, while separate cross-snapshot re-estimation supports the same stability conclusion. Figure D.1 reports the per-point effects and intervals.

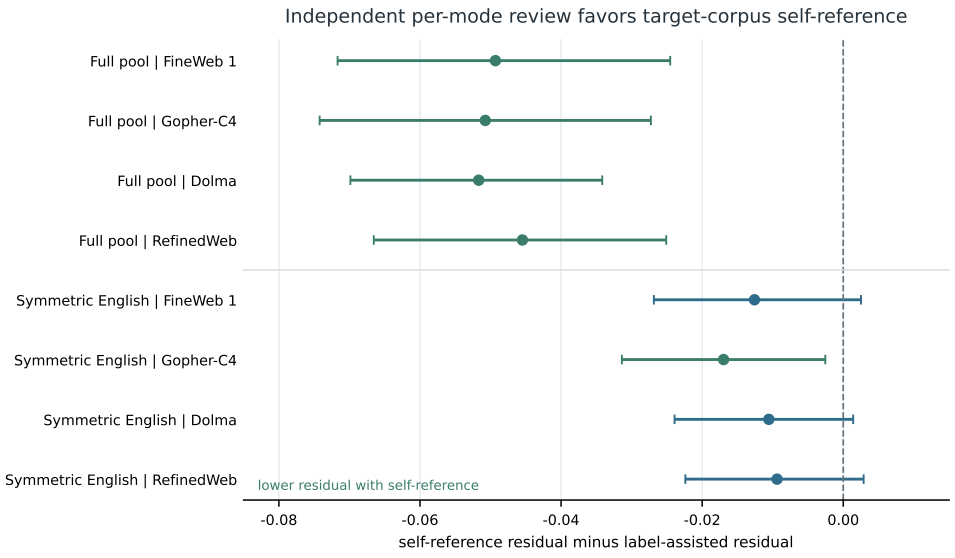

Figure D.1. Self-reference minus label-assisted residual under independent per-mode adjudication. Negative values favor self-reference. All four full-pool intervals exclude zero; symmetric-English point estimates agree, with one interval excluding zero.

E Negative Results and Boundaries

E.1 Controlled-pool diagnostics

The historical comparison combines a Gaussian-model KL/Pinsker calculation, an out-of-sample classifier, and the best observed matrix signal. The first is a model-dependent reference and the classifier measures achieved performance, so neither certifies an information ceiling for the actual feature distribution. Near-perfect observed performance on several controlled failure modes and the limited gains from pool expansion support empirical saturation on that grid. They do not establish transfer to natural failures, where the best single signal reaches 0.768.

Figure E.1 retains the historical comparison; its bound and ceiling labels denote model-based and fitted diagnostics, respectively. The controlled and natural evaluations remain distinct.

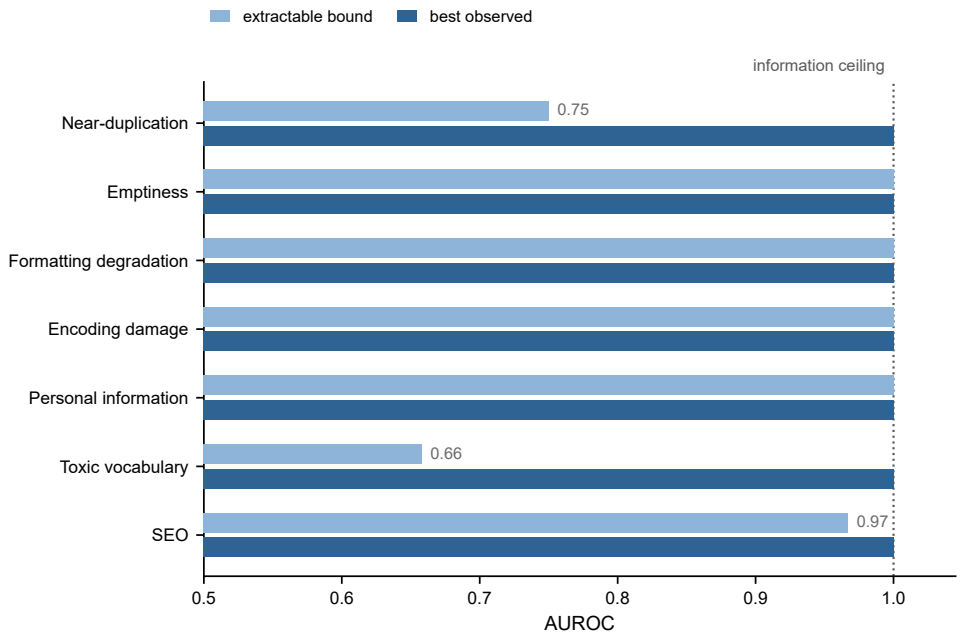

Figure E.1. Historical controlled-pool diagnostics by failure mode. The information-bound and extractable-ceiling labels refer to a Gaussian-model calculation and fitted classifier performance; neither is a certified ceiling for the actual joint feature distribution.

E.2 Perplexity

A 410M language model's perplexity does not detect these artifacts. On an 84k sample its AUROC is 0.412, below chance, 0.368 after length correction, and it correlates negatively with the zero-training signals at -0.586. Surface artifacts and low perplexity are different properties, which is the empirical basis for treating surface statistics and learned quality classifiers as complementary.

E.3 A historical distributional diagnostic

The archived nine-mode comparison gives Spearman correlation 0.25 ($p = 0.5165$) between the legacy maximum predicted single-operator effect and the mean of 20 marginal Gaussian plug-in TV values, averaged over two substrates at intensity 0.03. We retain the historical result as a diagnostic only, and we do not use it to validate a distributional prediction or to calibrate detection limits. Appendix G.5-G.8 separates mean-response geometry, distributional discrepancy, and achieved filtering performance.

E.4 Effective-sample bridge

Any quality-discounted scaling extrapolation must use the residual artifact rate conditional on retention: initial prevalence times one minus the true-positive rate, divided by the keep rate (G.8). False positives reduce the retained clean mass; they are not an additional artifact term. The archived effective-data calculations are model-based extrapolations and are not revalidated by the distributional identities here. They must specify document versus token weighting, retained data amount, and the assumed quality and scaling laws before numerical reuse. They remain separate from downstream training evidence and do not determine its reported values.

E.5 Length sensitivity

The length boundary reported in the boundaries subsection is established on real material rather than on controlled grids: removing short documents from the retained sets leaves the deployment advantage unchanged.

F Downstream Protocol

F.1 Protocol

Figure F.1 summarizes the frozen comparison. ZenoFilter, FineWeb, and matched random retention start from the same extracted, deduplicated, English-identified base; the deduplication-ignored control starts before deduplication. FineWeb fixes the realized token budget, and the sole confirmatory contrast compares token-matched ZenoFilter with FineWeb after identical 160M training schedules and blocked seeds.

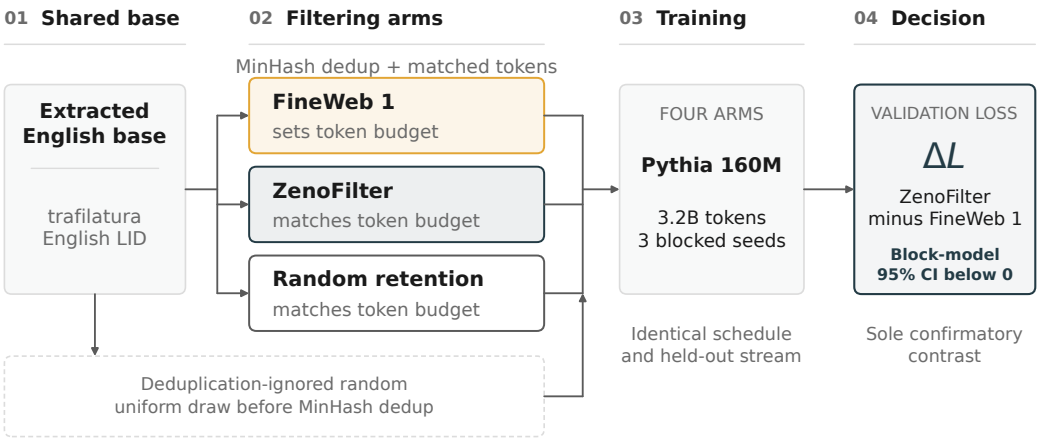

Figure F.1. Frozen downstream protocol. The three matched arms share preprocessing and training compute; the deduplication-ignored control starts before deduplication. FineWeb fixes the token budget, and ZenoFilter minus FineWeb validation loss is the sole confirmatory contrast.

Four arms enter the downstream analysis: the detector (zeno_len_tm at the deployed token-matched point, archived as zeno_raw_len_tm), FineWeb 1 as the designated strong peer, random_len_tm as the length-matched selection-null control, and random_en_nodedup as a deduplication-ignored control. The original roster also contained unfiltered text, Gopher-C4, Dolma, RefinedWeb, and a length-stratified ablation, but these arms were not trained. A residualised arm was trained and deprecated as a non-treatment on 2026-09-15; it is reported separately and excluded from the four-arm analysis. FineWeb fixes the reference corpus at its native keep rate. The three matched arms consume **exactly the same realized token count**: ZenoFilter keeps the lowest-scoring documents until its retained corpus reaches FineWeb's token count (which, because ZenoFilter's retained documents are shorter on average, keeps more documents than FineWeb), and the random control samples documents at random until the same token count. Token-matching (equivalently compute, at fixed batch and model) makes every matched arm train the same number of tokens and the same number of passes, so the confirmatory contrast isolates data quality rather than data volume or repetition. The 17,000-WARC expansion is expected to leave about 0.91B one-pass Pythia tokens in the binding arm at the point estimate and about 0.81B at the conservative bound. All four analysis arms carry the same three seeds, while FineWeb and random_len_tm carry two additional seeds. Each arm trains Pythia 160M on the same smallest whole-effective-batch token count at or above 3.2B (Hoffmann et al., 2022), realized as 3,200,253,952 tokens; each dataloader epoch is trimmed to a whole 512-sequence effective batch so no arm receives a short gradient-accumulation update at its epoch boundary. Exact arm-specific corpus passes are disclosed and must not exceed four, with a one-pass floor of exactly 800,063,488 tokens. This cap is conservative relative to evidence that up to four repeated epochs produced negligible loss changes at fixed compute (Muennighoff et al., 2023), but is not treated as proof that repetition is inert in our setting. Documents are tokenized without truncation and packed without padding, with one EOS boundary after every document. We initialize from the official

Pythia-160M configuration and follow the published 160M/3.2B setup of Thrush et al. (2025): an effective batch of 512 sequences of length 2,048, learning rate 5e-3, 10% warmup, cosine decay, Adam betas 0.9 and 0.95 with epsilon 1e-8, weight decay 0.1, gradient clipping at 1.0, and bf16. Optimizer step 750 corresponds to 786,432,000 tokens, below the exact 800,063,488-token launch floor and therefore within every arm's first pass, and is reported as a descriptive common first-pass sensitivity point; it cannot replace or rescue the 3.2B confirmatory result.

The sole confirmatory outcome is common held-out validation loss, computed on 2,048 complete blocks selected at deterministic equal intervals across the entire packed validation stream rather than from its prefix; every run records the same block-index hash. We use loss rather than task accuracy because accuracy at this model scale is sparse, discrete, and seed-sensitive and does not establish a stable corpus ordering; the completed power audit is reported in F.5. The sole confirmatory contrast is ZenoFilter minus FineWeb. We claim a downstream advantage only if its two-sided 95% confidence interval lies wholly below zero. Because there is one confirmatory contrast, no multiplicity adjustment is required. ZenoFilter minus random retention is a hierarchical supporting contrast that tests selection against uniform sampling at the same token budget; it cannot rescue a failed primary contrast. The other trained controls and the LAMBADA, ARC-Easy, PIQA, and SciQ evaluations are retained as auxiliary checks and do not enter the confirmatory decision; we freeze lm-evaluation-harness 0.4.3, zero-shot prompting, the paper's 5,000-example limit, fixed evaluator seeds, and no response cache, and report cross-arm average rank without using it as a second decision channel.

The same seeds (42, 1234, and 20260812) form blocks across the four analysis arms. We fit an additive complete randomized-block model with arm and seed effects. Its residual degrees of freedom are (4 minus 1) times (3 minus 1), or 6, and the standard error of a balanced pair contrast is the square root of twice the residual mean square divided by three. We report the effect, two-sided 95% t interval, three raw paired differences, per-arm dispersion, and model residuals. Because this interval pools residual variance across arms, we additionally report a direct paired-t interval with two degrees of freedom and all leave-one-seed-out estimates as assumption and influence diagnostics; they cannot rescue or silently replace the frozen block-model decision. Every run is configured from the outset with the full 3.2B-token schedule but pauses at 0.1B tokens for mechanical checks only: finite loss, hashes and token counts, validation separation, throughput, and checkpoint resumability. Before launch we also block duplicate arm rows, exact/prefix and anchored-13-gram validation overlap, corrupted packed-file hashes, inadequate disk/GPU capacity, and mismatched software/hardware fingerprints. Performance rank is not an early-stopping or rerun rule. Passing runs resume the same checkpoint under the unchanged full schedule, so the check consumes no additional training tokens; failed attempts retain the same arm and seed. The effective-sample bridge remains separate from these runs.

Table F.1. The four downstream analysis arms. ZenoFilter, FineWeb, and matched random retention share the extracted, deduplicated, English-identified base; the descriptive control omits deduplication. All four use Pythia 160M, 3.2B tokens, and the same three analysis seeds.

Arm	Token budget	Role
FineWeb 1	defines the matched budget (native keep)	designated confirmatory peer
Random matched	matched	hierarchical supporting contrast
ZenoFilter token-matched	**matched**	**sole confirmatory contrast**
Deduplication-ignored random	matched compute	descriptive control

F.2 PRIMARY COMPARISON

On the held-out endpoint the detector's corpus reaches 3.593629 nats, the length-matched random control 3.553022, the strong peer 3.677139, and the deduplication-ignored control 3.637855 (Table F.2; trainer evaluation on the 2,048 complete validation blocks, bf16). Every arm is Pythia-160M at 3,200,253,952 tokens over 3052 steps. The detector arm and the deduplication-ignored control carry three seeds each; the remaining arms carry five. Paired contrasts use the seeds shared by both arms (Table F.3). The detector sits 0.085827 nats below the strong peer, negative on all three shared seeds, and 0.019453 nats above the length-matched random control, positive on two of the three. A direct paired-t interval on the peer contrast has two degrees of freedom and spans (-0.173486, +0.001831); a four-arm by three-seed randomized-block interval has six and spans (-0.163638, -0.008017). The two differ in how residual variance is pooled.

Table F.2. Per-arm final held-out validation loss, trainer evaluation on the 2,048 complete validation blocks, bf16. Step 3052, 3,200,253,952 tokens. Seed labels: s0812 = 20260812, s0813 = 20260813, s0814 = 20260814. Arm names are abbreviated here: zeno_raw = the detector (zeno_raw_len_tm), random_len = random_len_tm, nodedup = random_en_nodedup, resid = resid_len_tm, the deprecated control. The detector and nodedup arms carry three seeds; the others carry five.

Arm	n	s42	s1234	s0812	s0813	s0814	Mean
zeno_raw	3	3.5641	3.5578	3.6589	—	—	**3.5936**
fineweb	5	3.6837	3.6466	3.7080	3.6894	3.6579	**3.6771**
random_len	5	3.5342	3.5970	3.5913	3.5322	3.5104	**3.5530**
nodedup	3	3.6684	3.6177	3.6275	—	—	**3.6379**
resid	5	3.6865	3.5869	3.6649	3.6490	3.6652	**3.6505**

Table F.3. Paired contrasts on the seeds shared by both arms; negative values favour the left arm. Pipeline A is the trainer evaluation (2,048 blocks, bf16); pipeline B is the CPU fp32 held-out evaluation (128 of the 574 WikiText blocks, 128 of the 3,748 val_common blocks). The two pipelines are not interchangeable and their absolute levels differ, so contrasts are comparable within a pipeline and not across.

Contrast	Pipeline	Endpoint	n	Mean
zeno_raw − fineweb	A	val_common	3	-0.0858
zeno_raw − random_len	A	val_common	3	$+0.0195$
fineweb − random_len	A	val_common	5	$+0.1241$
nodedup − random_len	A	val_common	3	$+0.0637$
nodedup − zeno_raw	A	val_common	3	$+0.0442$
resid − random_len	A	val_common	5	$+0.0975$
zeno_raw − random_len	B	WikiText	3	-0.0527
zeno_raw − fineweb	B	WikiText	3	-0.0384
zeno_raw − random_len	B	val_common	3	$+0.0100$

F.3 Held-out generalisation

The downstream validation has two complementary scopes. The high-quality Nemotron-CC web set is source-aligned with the training corpora because both are drawn from Common Crawl; this label does not assert document overlap, identical extraction, or identical processing. The six cross-content categories cover books, news, mathematics, encyclopedic text, scientific text, and code; books and code each have two held-out datasets, while the category-level statement treats them as one content category. Across these endpoints, ZenoFilter is below FineWeb in held-out loss, including a 0.0908-nat/token difference on Nemotron-CC. On WikiText the four corpora order detector (4.1913), strong peer (4.2248), length-matched random (4.2429), deduplication-ignored control (4.3170); the detector is 0.0527 nats below the random control (t = -5.05) and 0.0384 below the peer. LAMBADA perplexity orders them the same way, at 152.26, 167.88, 185.16 and 231.85. On the in-distribution web endpoint of the same pipeline the detector and the random control are not separated (+0.00998, t = +0.31). On the equal-weight mean across the Pile's 22 fixed components, the detector is 0.0791 bits per byte above the random control and 0.0549 below the peer (Table F.4). Its loss is below the peer on 18 of the 22 components.

F.4 Residual artifacts and length composition

The detector's corpus carries a residual-artifact rate of 0.1357 against 0.374 on the eligible pool, 0.2224 for the peer, 0.4000 for the length-matched random control and 0.3974 for the deduplication-ignored control (document-weighted, the same endpoint throughout). Deduplication removes documents that are dirtier than those it keeps, 0.4798 against 0.3736. The detector's own length composition accounts for none of its rate: holding its length distribution fixed and replacing its within-stratum selection with a random draw raises the rate from 0.1357 to 0.3998. The length-matched random control returns 0.4000 observed against 0.3999 under the same substitution, which is the check that the calculation is sound.

Table F.4. Held-out bits per UTF-8 byte on the Pile's 22 components (NeelNanda/pile-10k, each domain capped at 150 documents, fp32). bpb is the Pile's own tokenization-invariant metric. Values are the mean over the three matched seeds (1234, 20260812, 42); paired differences are taken per seed then averaged. Several domains carry very few documents (BookCorpus2, Gutenberg and Ubuntu IRC two each; PhilPapers five; EuroParl six; Books3 nine; YoutubeSubtitles eleven), and the equal-weight macro counts each such cell as heavily as a 150-document one. ZF − Rand is positive in 18 of 22 domains; ZF − FW is negative in 18 of 22.

Domain	Docs	ZF	Random	FineWeb	ZF−Rand	ZF−FW
ArXiv	91	1.7981	1.5864	2.2721	+0.2117	−0.4740
BookCorpus2	2	1.4330	1.4640	1.4354	−0.0310	−0.0024
Books3	9	1.7104	1.6900	1.7361	+0.0204	−0.0257
DM Mathematics	99	2.8980	2.5082	2.9895	+0.3898	−0.0915
Enron Emails	47	2.3031	2.1548	2.3211	+0.1483	−0.0180
EuroParl	6	2.2321	2.1475	2.2437	+0.0846	−0.0116
FreeLaw	150	1.4662	1.4325	1.4499	+0.0337	+0.0163
Github	150	2.5264	2.2024	2.7409	+0.3240	−0.2146
Gutenberg	2	2.1596	2.1563	2.1760	+0.0033	−0.0164
HackerNews	81	2.2331	2.1838	2.2110	+0.0493	+0.0220
NIH ExPorter	104	0.9876	0.9963	0.9979	−0.0088	−0.0104
OpenSubtitles	27	1.3643	1.3673	1.3785	−0.0030	−0.0142
OpenWebText2	150	1.4360	1.4374	1.4381	−0.0013	−0.0020
PhilPapers	5	1.5476	1.5095	1.5637	+0.0381	−0.0161
Pile-CC	150	1.4025	1.3973	1.3954	+0.0051	+0.0071
PubMed Abstracts	150	1.0592	1.0588	1.0850	+0.0004	−0.0258
PubMed Central	150	1.3952	1.3411	1.4916	+0.0541	−0.0964
StackExchange	150	2.1539	2.0110	2.2996	+0.1429	−0.1457
USPTO Backgrounds	150	1.0137	0.9905	1.0095	+0.0232	+0.0042
Ubuntu IRC	2	2.4789	2.3417	2.5431	+0.1372	−0.0642
Wikipedia (en)	150	1.4959	1.4851	1.5038	+0.0107	−0.0079
YoutubeSubtitles	11	2.9296	2.8230	2.9491	+0.1066	−0.0195
Macro (22 domains)		**1.8193**	**1.7402**	**1.8741**	**+0.0791**	**−0.0549**

F.5 Task-level evaluation and domain-composition analysis

The completed task battery contains 52 tasks across the frozen evaluator outputs. In a balanced four-corpus-by-three-seed power audit, only five tasks exceed the prespecified detectable-variation threshold (F ¿ 4.76), while the median task-level F statistic is 1.08. Several tasks are near or below their random baselines, and individual-seed differences can reverse the corpus ordering. At this model scale, task accuracy is too sparse, discrete, and seed-sensitive to establish a stable corpus ordering; we therefore use held-out loss for the downstream comparison.

The component analysis explains the different ordering against matched random retention. Across the fixed Pile-22 components, the detector remains below FineWeb on 18 of 22 components, but matched random retention has the lower equal-weight mean by 0.0791 bpb. On the 12 content endpoints used for the composition probe, the ordering of the random-minus-detector contrast follows the detector's estimated content coverage (Spearman rho = 0.90 with the published markers and 0.93 with the strict markers). The leave-one-out and marker-sensitivity checks preserve this ordering (Table F.5).

The seven symbol-bearing Pile components used in the subset check are ArXiv, DM Mathematics, Enron Emails, FreeLaw, Github, StackExchange, and Ubuntu IRC.

The same analysis separates several controls. After splitting held-out documents by seen versus unseen status and controlling for length, domain, and crawl block, the random-minus-detector contrast remains larger on unseen documents (-0.319 versus -0.289), while the seen coefficient is small and positive (+0.0245, t = 1.38). Across the same pool, the strict code-marker audit places the detector's code preference at 0.19, compared with 0.07 for FineWeb, 0.11 for Gopher/C4 and Dolma, and 0.44 for the prior-based selector; the official DCAD implementation is reported separately because its feature and operating-point protocol differs. Table F.4 gives the complete Pile-22 component values, and Table F.5 records the composition, seen-status, and marker-sensitivity checks.

Continuation pilot. A targeted continuation pilot replenishes code and mathematical text in the detector corpus. At 1B additional tokens, the symbolic group (two code endpoints, mathematics, and

Table F.5. Domain-composition sensitivity checks. The coverage probe uses the 12 held-out content endpoints and two marker definitions. The seen-status row reports the controlled random-minus-detector contrast; negative values favour random retention. The marker row is a sensitivity check, not a claim that either marker set is a semantic ground truth. The Pile subset row uses the seven symbol-bearing components listed in the accompanying analysis record.

Check	Estimate	Reading
Coverage probe (12 endpoints)	$\rho = +0.90$ published; $+0.93$ strict	Random-minus-ZF ordering follows estimated coverage.
Leave-one-out	published $+0.87$–$+0.93$; strict $+0.91$–$+0.96$	The sign is unchanged after removing any endpoint; removing code gives $+0.867/+0.850$, and removing books gives $+0.939/+0.952$.
Seen status	Random-minus-ZF: -0.289 seen, -0.319 unseen; seen coefficient $+0.0245$ ($t = 1.38$)	The random advantage is not larger on documents seen by the random arm.
Code-marker sensitivity	ZF/FineWeb retention ratio: $0.92\times$ published, $2.97\times$ strict	The marker definition changes the absolute code share but not the direction of the endpoint ordering.
Pile symbol-bearing subset	7 domains: FineWeb$-$ZF $= +0.1417$ bpb; random$-$ZF $= -0.1982$ bpb	Across all $\binom{22}{7} = 170{,}544$ subsets, 99.8% have FineWeb$-$ZF > 0 and 100% have random$-$ZF < 0.

arXiv) improves by -0.169 nats/token on average, while the generic-text group changes by -0.016 nats/token. The intervention therefore moves the targeted endpoints in the predicted direction while leaving general-text losses nearly unchanged, supporting the domain-composition interpretation. The full stream construction, endpoint-level values, dose trajectory, and seed-level limitations are reported in the project record; the main text uses only the directional conclusion.

F.6 Boundaries

The in-distribution web endpoint orders the arms the other way, with the length-matched random control first.

F.7 Training-trajectory check

Re-running the held-out web endpoint on existing intermediate checkpoints (32 fixed blocks, paired across checkpoints) shows the detector's advantage over FineWeb growing through training and then plateauing: FW$-$ZF is $+0.0187$ at step 2250 and $+0.0358$ at steps 3000 and 3052, with the random-retention and deduplication-ignored contrasts moving the same way. Earlier checkpoints exist for two arms only, and they show the artifact effect absent at the very start of training: at step 96 the higher-artifact arm is the better one, by $+0.1127$ nats/token, while the sign has flipped and stabilised near -0.09 by step 2250. Because 3.2B tokens on a 160M model is the standard budget of twenty tokens per parameter, the early point sits below every evaluation regime we report; the direction is that the benefit of a clean corpus accrues with training rather than at its start. Absolute losses on this fixed block subset are not comparable to the published evaluation calibers, but the cross-checkpoint comparisons are paired.

G Statistical Derivations

Figure G.1 gives the dependency structure behind the derivations below. Prevalence determines the expected response, reference moments place it on a comparable scale, the screening criterion combines response strength with directional novelty, and the branch maximum preserves rare-mode decision paths. Distributional assumptions enter only where separation is related to AUROC.

G.1 Standardization

Each operator is direction-corrected and standardized against moments estimated on the reference set within five length buckets. Bucketing is necessary because local statistics are sensitive to document length, so that a global standardization would read length differences as anomalies. The sign aligns larger standardized values with the abnormal direction and is fixed. The floor on the standard deviation prevents a near-constant signal, whose reference standard deviation approaches zero, from

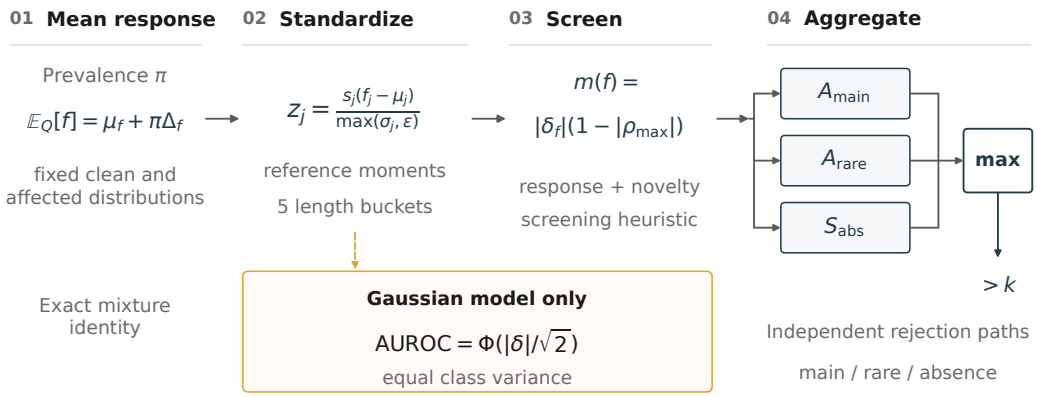

Figure G.1. Dependency structure of the statistical argument. Distributional assumptions enter only when standardized separation is related to AUROC; calibration, screening, and aggregation remain deterministic.

amplifying a bounded numerator without limit; without it a single degenerate signal can dominate the aggregate.

G.2 Linearity in prevalence

For fixed clean and affected conditional distributions, the mixture expectation is the clean mean plus prevalence times the conditional mean difference. This is an identity for any integrable operator. The standardized expectation is likewise linear when the reference moments are fixed. Changing within-document intensity changes the affected distribution; moving from controlled transformations to real WARC failures can change both conditional distributions. The identity therefore separates prevalence from response but does not establish linearity in intensity or transfer of the response to real failures.

G.3 Separation and AUROC

For Gaussian class-conditional scores with common variance, the direction-aligned AUROC equals the standard normal distribution function evaluated at the absolute standardized mean difference divided by the square root of two. Ranking by this difference is valid within that model. Unequal variances, heavy tails, and multimodality can invalidate this mapping; an AUROC near one alone is not evidence of model failure. The controlled benchmark reports measured AUROC, and the deployed weighted scores are not assumed to attain the Gaussian optimum.

G.4 The screening criterion

Equation (3) is a response-and-redundancy heuristic, not a first-order approximation or a lower bound on Equation (2). For one retained coordinate, unit clean variances, correlation rho, and conditional mean shifts delta-f and delta-s, the exact increment is

$$J(f \mid s) = \frac{(\delta_f - \rho\delta_s)^2}{1 - \rho^2}, \qquad |\rho| < 1.$$

The heuristic omits the retained mean shift and replaces residual standardization with a correlation penalty. For example, delta-f = 1, delta-s = 2, and rho = 1/2 give zero exact increment but heuristic value 1/2. Conversely, delta-f = 0 and delta-s = 1 at the same correlation give increment 1/3 but heuristic value zero. A candidate with no marginal response can therefore improve a linear contrast by canceling correlated clean variation. The heuristic can miss such candidates, which limits screening coverage. Also, the square root of one minus rho squared has no linear term near zero, whereas one minus absolute rho does; they are not first-order equivalent. Held-out combination gains assess the shortlisted candidates without claiming an exhaustive optimum.

G.5 Multivariate form and linear-kernel MMD

Write R for the affected distribution at a fixed failure mode and intensity, and let F denote a fixed feature vector with clean mean mu-P and positive-definite clean covariance Sigma-P. For finite second moments, define the kernel and discrepancy

$$k_P(u, v) = (u - \mu_P)^\top \Sigma_P^{-1}(v - \mu_P), \qquad D_F^2 = \mathrm{MMD}_{k_P}^2(P_F, R_F) = \Delta_F^\top \Sigma_P^{-1} \Delta_F.$$

Here P-F and R-F are the feature pushforwards and Delta-F is their mean difference. The identity follows by taking the squared norm of the difference of mean embeddings under clean whitening. On adding a coordinate, block inversion of the clean covariance gives exactly Equation (2): the augmented whitened representation consists of the retained representation and its unit-variance regression residual, so their squared mean differences add. No Gaussian or equal-class-covariance assumption is required for this algebra. Singular covariance and zero residual variance are excluded; ridge regularization defines a modified geometry, not the same quantity.

Under the additional assumption that both feature distributions are Gaussian with common covariance, the likelihood-ratio score is the Fisher direction and

$$\mathrm{AUROC}_{\mathrm{opt}} = \Phi(D_F/\sqrt{2}), \qquad \mathrm{TV}(P_F, R_F) = 2\Phi(D_F/2) - 1.$$

These performance identities do not extend to arbitrary distributions or to the frozen branch scores. We use diagonal calibration and frozen weights; estimating a full inverse covariance, especially with correlated coordinates and about 60 reference documents per length bucket, would require a separately validated regularization scheme. The trained comparator in Appendix D.4 is empirical evidence about a different decision rule, not a proved information ceiling.

G.6 Why a maximum rather than a sum

If a mainstream component has zero conditional mean shift and zero covariance with a rare-mode component, adding it increases variance without increasing the rare-mode contrast. This motivates separate branches. Low prevalence alone does not imply these conditions or prove that a maximum is optimal. The maximum retains separate rejection paths but also changes the clean-score tail, so its operating point must be compared at matched keep. Appendix D.2 supplies this empirical comparison.

G.7 Distributional scope and nonlinear diagnostics

MMD extends the test-function class used to compare distributions. A characteristic kernel, such as a Gaussian kernel with fixed positive bandwidth, distinguishes any unequal feature distributions at the population level, including changes invisible to linear-kernel MMD. For example, equal mixtures of point masses at -1 and 1 versus at -2 and 2 have zero mean separation and TV one. A nonlinear diagnostic can detect this difference, although the one-sided linear-score AUROC is one half. A kernel on the same features cannot recover distinctions erased by the feature map.

For a fixed kernel with diagonal bounded by kappa squared, the RKHS unit ball contains only functions bounded by kappa, giving the first inequality below. For Wasserstein-1 on a metric space of finite diameter L, maximal coupling gives the second:

$$\mathrm{MMD}_k(P_F, R_F) \le 2\kappa\, \mathrm{TV}(P_F, R_F), \qquad W_1(P_F, R_F) \le L\, \mathrm{TV}(P_F, R_F).$$

These inequalities lower-bound TV from a discrepancy; they do not turn a small MMD or Wasserstein distance into a small TV upper bound. Point masses at zero and at a positive epsilon have TV one while both Gaussian-kernel MMD and Wasserstein-1 tend to zero as epsilon tends to zero. The bounded-kernel inequality does not apply to the unbounded linear kernel of G.5. On a fixed feature map and with a fixed test-function class, TV, MMD, and Wasserstein-1 all scale exactly with corpus prevalence under the two-component mixture; their squared versions scale quadratically. No analogous identity is asserted for within-document intensity.

G.8 DISTRIBUTIONAL CONTRAST AND RETAINED NOISE

Fix the feature map, reference moments, and weights. At document keep rate q, let h denote the induced rejection rule taking values in [0,1], allowing randomized treatment of threshold ties. Let u be its clean rejection rate and v its noise rejection rate. With noise prevalence pi and 0 ¡ q, direct conditioning gives

$$1 - q = (1 - \pi)u + \pi v, \qquad \Gamma_q = v - u, \qquad \mathrm{Res}_q = \frac{\pi(1 - v)}{q} = \pi - \frac{\pi(1 - \pi)}{q}\Gamma_q.$$

At the same prevalence and keep rate, two operator sets therefore satisfy

$$G_q(f \mid S) = \frac{\pi(1 - \pi)}{q}\left[\Gamma_q(S \cup \{f\}) - \Gamma_q(S)\right], \qquad |\Gamma_q| \le \mathrm{TV}(P_F, R_F).$$

This is an exact population identity, and also a counting identity on a finite labeled pool using its observed prevalence and retention. It connects Equation (4) to the distributional contrast achieved by the deployed rule. TV optimizes over all measurable rejection rules, whereas this rule uses a particular score and operating point; MMD optimizes over an RKHS function class and does not directly determine either rule's rejection contrast. Consequently, greater full-distribution discrepancy need not yield a positive combination gain. Estimates on held-out WARC data remain subject to sampling uncertainty and label error.

For token-level residuals, the same conditioning argument requires token-level prevalence, token retention, and rejection rates conditional on token labels, with each token inheriting its document's rejection decision. Matching document keep does not match token keep, so the document-level gain identity cannot be substituted directly into the token-level comparison.

Finally, transfer from controlled to natural failures requires controlling a separate distribution shift. For the same frozen rule and observation map, its rejection contrast obeys

$$|\Gamma_{\mathrm{real}}(h) - \Gamma_{\mathrm{syn}}(h)| \le \mathrm{TV}(P_{F,\mathrm{real}}, P_{F,\mathrm{syn}}) + \mathrm{TV}(R_{F,\mathrm{real}}, R_{F,\mathrm{syn}}).$$

This follows by adding the changes in the two rejection probabilities. Recalibrating the feature map or threshold changes the rule and falls outside this fixed-rule comparison. We do not estimate these shift bounds from marginal Gaussian fits or small MMD values; the real-data combination tests provide the operational evidence.

H EXTENDED RESEARCH DIRECTIONS

H.1 COMPLETE CURATION STACKS

A matched-compute factorial evaluation should place surface filtering before and after deduplication, language identification, learned quality ranking, and domain reweighting. The target quantity is the marginal contribution of each stage and its interactions, rather than the performance of an isolated filter.

H.2 INDUCING FAILURE FAMILIES

The current failure modes begin from published rules. A complementary discovery route would embed raw crawl documents, identify stable anomaly clusters without a mode vocabulary, and test whether the clusters recur across snapshots, languages, and domains before mapping them back to observable mechanisms.

H.3 HYBRID RESIDUAL DETECTION

The Propella alignment map shows that some independently specified properties intersect the surface space while others remain weak or corpus-specific. A hybrid pipeline can exploit this division by

applying ZenoFilter at corpus scale and routing only its residual to learned, embedding-based, or provenance-aware detectors.

H.4 Reference representativeness

Self-reference estimates local moments efficiently, but a deployment criterion should also detect when a sample ceases to represent the target mixture. Future work can combine moment stability, coverage over length and script strata, and change-point tests to decide when a reference must be rebuilt.