

SCARV: Structure-Constrained Aggregation for Stable Sample Ranking in Redundant NLP Datasets

Feiyu Wu^{1,2}, Xu Zheng², Zhuocheng Wang², Dai YiMing²,
Hui LI^{1,2*}

Xidian University

¹ State Key Laboratory of Integrated Services Networks

² School of Cyber Engineering

sn0wm1ans@gmail.com, zhengxu200477@gmail.com

Abstract

Sample-level ranking is widely used for data selection, debugging, and training-data attribution, but redundant NLP corpora—exact and near duplicates, paraphrases, and templated examples—can make proxy-induced rankings unstable across random seeds. We study stable sample-level ranking under redundancy and propose SCARV, a modular aggregation layer that operates on top of existing scoring proxies. SCARV normalizes scores within each run, can regularize redundant examples through cluster-level aggregation and allocation, and aggregates across seeds with a robust operator. Across synthetic redundancy, naturally mined QQP redundancy, multiple proxy families, classification and sentence-pair tasks, and end-to-end DistilBERT fine-tuning, stability-aware aggregation improves over bare proxy rankings and yields more reproducible ranking-based decisions, such as selected subsets and suspicious-example sets. Matched mechanism analysis shows that cross-seed aggregation is the dominant generic stabilizer. The incremental effect of structural regularization is small, map-dependent, dataset-dependent, and sometimes negative; a meaningful redundancy map is useful mainly when group-consistent treatment is operationally desired. These results position SCARV as a stability-oriented aggregation framework rather than a universal data selector and emphasize decision reproducibility over uniform gains in downstream utility.

1 Introduction

Sample-level ranking has become a core primitive in data-centric NLP. Researchers and practitioners use example-level scores to inspect corpora, identify suspicious or mislabeled training points, filter data, and construct smaller or higher-quality subsets. Existing pipelines derive such rankings from data valuation, training-data attribution, or training-

dynamics signals, including Shapley-style valuation, learned valuation, influence-based scoring, example forgetting, and dataset cartography (Ghorbani and Zou, 2019; Jia et al., 2019b; Yoon et al., 2020; Koh and Liang, 2017; Pruthi et al., 2020; Park et al., 2023; Toneva et al., 2019; Swayamdipta et al., 2020). Although these methods differ in semantics and motivation, they typically share the same operational pattern: assign a pointwise score to each example, then sort the resulting scores as if the examples were independent units.

For NLP corpora, this independence assumption is often fragile. Real datasets contain exact duplicates, near-duplicates, paraphrases, templatic expressions, repeated reports, and other forms of local redundancy. Deduplication studies have shown that repeated text is common enough to affect memorization, privacy risk, and empirical conclusions in modern language-model pipelines (Lee et al., 2022; Kandpal et al., 2022). Under such redundancy, sample-level ranking becomes unstable in a practically inconvenient way: examples that convey nearly the same information can still receive noticeably different scores because of random initialization, stochastic optimization, bootstrap resampling, or incidental interactions between a proxy and repeated structure. As a result, the local ordering of highly similar examples can vary across random seeds even when the dataset, model family, and base proxy are otherwise fixed. Figure 1 illustrates this failure mode.

This observation suggests that, in redundant NLP datasets, sample ranking is not only a question of what signal to use, but also of how reproducibly that signal can be realized. Recent work has increasingly recognized robustness as a first-class concern in data valuation: Distributional Shapley emphasizes stability under data perturbations, while Data Banzhaf and Weighted Banzhaf explicitly study how stochastic learning induces inconsistent value rankings (Ghorbani et al., 2020; Wang

*Corresponding author: lihui@mail.xidian.edu.cn.

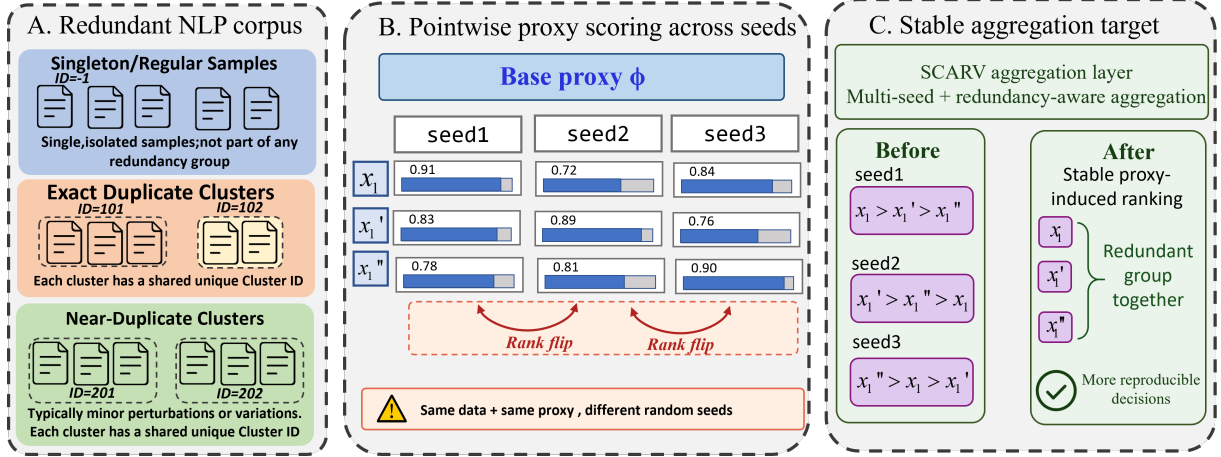


Figure 1: Ranking instability under redundancy. Redundant NLP corpora contain exact duplicates and near-duplicate clusters. Pointwise proxy scoring can assign locally similar examples different relative orders across random seeds, producing rank flips. SCARV targets a stability-oriented aggregation layer over existing proxies rather than redefining sample value.

and Jia, 2023; Li and Yu, 2023). Yet an important practical question remains underexplored for NLP: given an existing sample-level scoring proxy and a dataset with exact or approximate redundancy, how should noisy pointwise scores be aggregated so that the resulting ranking is more stable across seeds?

We study this problem as stable sample-level ranking under redundancy. Our goal is deliberately narrower than proposing a universally optimal data selector. We do not seek to replace existing proxies with a new monolithic notion of sample value, nor do we assume that greater ranking stability automatically implies stronger downstream pruning performance. Instead, we ask a modular question: given a base proxy and an approximate redundancy map, how can we transform per-example scores into a ranking that is more stable across random seeds while preserving the main signal of the original proxy? This question is complementary to deduplication. Deduplication asks whether repeated content should be removed or consolidated before training; we ask how rankings should be stabilized once redundancy is already present or approximately identified.

To address this question, we propose SCARV (Structure-Constrained Aggregation for Stable Sample Ranking), a modular aggregation framework that operates on top of an existing proxy. SCARV combines two ingredients: robust multi-seed aggregation, which suppresses seed-specific fluctuations by aggregating proxy scores across repeated runs, and a structure-aware aggregation/allocation step, which uses redundancy clus-

ters to regularize how local scores are pooled and redistributed within groups of similar examples. Conceptually, SCARV does not redefine sample importance; it regularizes how a chosen proxy behaves under redundancy and stochastic training.

A central principle in our study is to separate three objectives that are often conflated in data-centric NLP: valuation alignment, ranking stability, and downstream selection utility. A ranking can be more stable without becoming a better approximation to leave-one-out value. Likewise, a better-aligned valuation signal need not be the best direct selector under a fixed pruning budget (Wang et al., 2024b). We therefore evaluate stability, selection utility, noisy-label retrieval, and leave-one-out alignment as distinct axes. This framing leads to a deliberately nuanced conclusion: SCARV improves reproducibility of proxy-induced rankings and ranking-based decisions, but should not be read as a universal guarantee of downstream utility gains.

The empirical results support this interpretation. Across SST-2, AG News, MRPC, MNLI, and QQP, stability-aware aggregation substantially improves over bare proxy rankings under redundancy. However, matched mechanism analysis shows that robust multi-seed aggregation is the dominant generic stabilizer. Once preprocessing and cross-seed operator are fixed, the incremental effect of the structure-aware stage is small, map-dependent, dataset-dependent, and sometimes negative. The submitted full-pipeline QQP comparison is positive, whereas the operator-controlled matched structural effect is small and protocol-

dependent. A small seed budget alone does not guarantee a structural benefit, because map construction also has a cost. The practical benefit is clearest in reproducible ranking-based decisions and group-consistent audit workflows, not in universal improvements to downstream accuracy or AUROC.

Contributions. First, we formulate stable sample-level ranking under redundancy as a distinct problem in data-centric NLP and argue that valuation alignment, ranking stability, and downstream selection utility should be evaluated separately. Second, we propose SCARV, a modular aggregation framework that combines robust multi-seed aggregation with optional redundancy-aware structural regularization on top of an existing proxy. Third, we provide an empirical study showing that stability-aware aggregation substantially improves bare proxy-induced rankings and decision reproducibility across synthetic redundancy, naturally mined QQP redundancy, multiple proxy families, and classification/sentence-pair tasks. Fourth, our matched factorial, timing, map-quality, tie, diversity, and audit diagnostics clarify the mechanism and scope: robust multi-seed aggregation is the dominant generic stabilizer, while structural regularization is best treated as a conditional tool for trusted redundancy maps and group-consistent auditing rather than as a universal source of ranking gains.

2 Related Work

Sample-level valuation and scoring. Sample-level valuation methods assign importance scores to training examples based on their contribution to model utility. Shapley-style and learned valuation methods provide principled or learned notions of data value, while Beta Shapley broadens the family of semivalue-based value notions (Ghorbani and Zou, 2019; Jia et al., 2019b; Yoon et al., 2020; Kwon and Zou, 2022). Our work is related in motivation but different in scope: SCARV does not introduce a new valuation functional. Instead, it operates downstream of a chosen proxy and asks how the induced ranking can be made more stable when redundancy and stochastic training interact.

Training-data attribution. A closely related line of work studies how individual training examples affect predictions or learned parameters.

Beyond influence functions, TracIn, and TRAK, representer-point methods offer another scalable route to example-level attribution (Koh and Liang, 2017; Pruthi et al., 2020; Park et al., 2023; Yeh et al., 2018). These methods are natural sources of sample-level scores, but they do not by themselves address redundancy-induced ranking instability. SCARV is complementary to this literature: it can sit on top of loss-based or influence-based scores and regularize the final ranking without replacing the underlying attribution mechanism.

Ranking, selection, and data-centric diagnostics. Sample-level ranking is also central to neighboring areas such as training-dynamics analysis, data filtering, and subset construction. Alongside example forgetting and dataset cartography, curriculum learning and core-set-style selection study how example order or subset composition affects learning efficiency and downstream performance (Toneva et al., 2019; Swayamdipta et al., 2020; Bengio et al., 2009; Sener and Savarese, 2018). Our empirical findings are aligned with the broader view that these objectives should not be conflated: SCARV is not a new curriculum or coreset method, and it is not presented as a universal selector. Its primary role is to make proxy-induced rankings more reproducible.

Redundancy, deduplication, and data curation. Our setting is especially connected to work on repeated content and deduplication in NLP corpora. Prior studies show that exact and near-exact repetition materially affect memorization, privacy risk, and evaluation reliability in language-model training (Lee et al., 2022; Kandpal et al., 2022). This literature establishes that redundancy is not a synthetic corner case. SCARV addresses a complementary problem: once duplicate or near-duplicate structure is present or approximately identified, how should sample-level scores be aggregated so that the resulting ranking is more stable across random seeds?

Positioning. Taken together, prior work has developed data value notions, attribution mechanisms, and data-centric ranking paradigms for diagnostics or selection. Our paper lies at the intersection of these threads but targets a narrower question: stable sample-level ranking under redundancy in NLP datasets. SCARV is therefore best understood as a stability-oriented aggregation layer rather than a new monolithic proxy, a new value functional, or a

universal data selector.

3 Method

3.1 Problem Setup

Let $D = \{(x_i, y_i)\}_{i=1}^n$ denote a training set. For each random seed $r \in \{1, \dots, R\}$, a base sample-scoring proxy ϕ produces one score per training example,

$$s_i^{(r)} = \phi((x_i, y_i); D, \theta^{(r)}), \quad (1)$$

where $\theta^{(r)}$ is the model obtained under seed r . The proxy ϕ may be loss-based, influence-oriented, confidence-based, or any other scoring rule that returns per-example values.

In addition, we assume access to an approximate redundancy map $\mathcal{C} = \{C_1, \dots, C_K\}$ over the training set. Each example belongs to exactly one cluster, where singleton clusters represent samples for which no redundant partner is identified. The clusters may come from exact duplicates, synthetic near-duplicates, or naturally mined similarity structure, and they need not be perfect.

Our goal is not to redefine the proxy itself. Instead, given the noisy pointwise scores $\{s_i^{(r)}\}$ and a redundancy map $\mathcal{C}$, we seek a final ranking π that is more stable across random seeds while preserving the primary signal of the underlying proxy.

3.2 SCARV: Stability-Oriented Aggregation

SCARV is a modular aggregation layer on top of an existing proxy. Figure 2 summarizes the pipeline. Starting from per-seed proxy scores, SCARV first constructs a redundancy-aware score within each seed and then aggregates these scores across seeds with a robust operator. The framework is designed to stabilize a proxy-induced ranking rather than to define a new valuation functional. It regularizes how an existing scoring signal behaves under redundancy and stochastic training.

Within-seed score preparation and structure-aware aggregation. Because raw proxy magnitudes may vary across seeds, we first normalize scores within each run:

$$\bar{s}_i^{(r)} = \frac{s_i^{(r)} - \min_j s_j^{(r)}}{\max_j s_j^{(r)} - \min_j s_j^{(r)} + \epsilon}. \quad (2)$$

The minimum and maximum are computed globally over the complete scored training subset for that seed, not over a mini-batch, cluster, or local

neighborhood. All matched comparisons use the same normalized scores before either enabling or disabling structural allocation. In the default implementation, we then add a light diversity term,

$$u_i^{(r)} = (1 - w)\bar{s}_i^{(r)} + wd_i, \quad w = 0.2, \quad (3)$$

where $d_i \in [0, 1]$ is the normalized mean cosine distance to the $k = 10$ nearest neighbors of example i . We treat $(w, k) = (0.2, 10)$ as the frozen submitted default, not as a tuned optimum. The diversity term introduces a stability–fidelity trade-off: larger w can stabilize rankings while moving them farther from the raw proxy ordering.

Let $c(i) \in \{1, \dots, K\}$ denote the cluster index of example i . For each cluster C_k , SCARV computes a cluster-level summary

$$a_k^{(r)} = A(\{u_j^{(r)} : j \in C_k\}), \quad (4)$$

and allocates this signal back to cluster members through

$$\hat{s}_i^{(r)} = B(u_i^{(r)}, a_{c(i)}^{(r)}, |C_{c(i)}|). \quad (5)$$

The default implementation uses arithmetic mean aggregation,

$$A(\{u_j^{(r)} : j \in C_k\}) = \frac{1}{|C_k|} \sum_{j \in C_k} u_j^{(r)}, \quad (6)$$

and hard replacement for non-singleton clusters,

$$B(u_i^{(r)}, a_{c(i)}^{(r)}, |C_{c(i)}|) = \begin{cases} u_i^{(r)}, & |C_{c(i)}| = 1, \\ a_{c(i)}^{(r)}, & |C_{c(i)}| > 1. \end{cases} \quad (7)$$

Thus singletons are unchanged, while all members of a non-singleton cluster receive the same within-seed score. This hard replacement can create ties and partially converts an item ranking into a group-level partial ranking. We therefore evaluate tie-aware Spearman ranks, Kendall τ_b , explicit top- k tie policies, and soft shrinkage diagnostics in addition to the submitted hard-replacement default.

Robust cross-seed aggregation. After the within-seed stage, each example has one score per seed, $\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(R)}$. SCARV forms the final score

$$\tilde{s}_i = \text{Agg}(\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(R)}), \quad (8)$$

and ranks samples by

$$\pi_{\text{SCARV}} = \text{rank}(\tilde{s}_1, \dots, \tilde{s}_n). \quad (9)$$

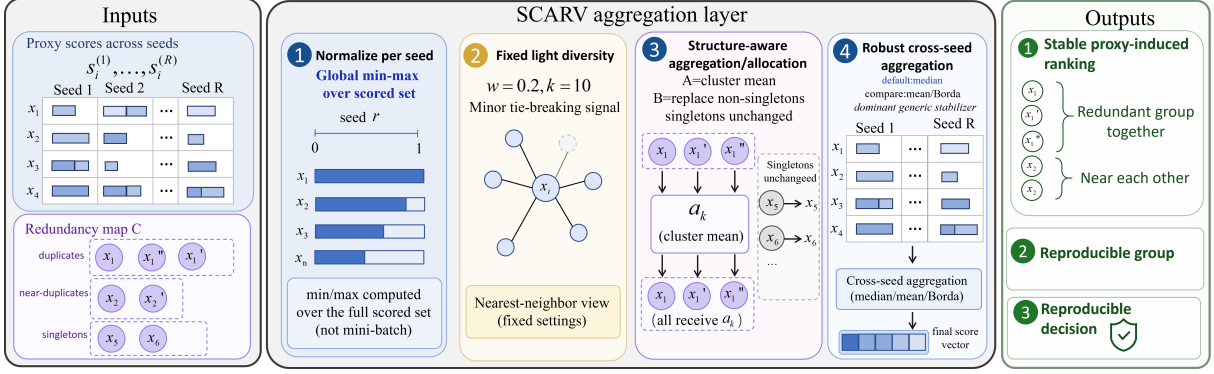


Figure 2: SCARV pipeline. Given proxy scores from repeated runs and a redundancy map, SCARV normalizes scores within each seed, optionally regularizes redundant examples through cluster-level aggregation and allocation, and aggregates across seeds with a robust operator. Multi-seed aggregation provides the largest generic stability gain in our study; structure-aware allocation is a conditional regularizer whose effect depends on the map, dataset, and audit objective.

In the default instantiation, $\text{Agg}(\cdot)$ is the median:

$$\tilde{s}_i = \text{median}(\hat{s}_i^{(1)}, \dots, \hat{s}_i^{(R)}). \quad (10)$$

This robust aggregation suppresses seed-specific outliers while preserving the central tendency of the proxy signal.

Although the default instantiation uses the median for robustness, we explicitly compare against seed-mean and seed-Borda aggregation in the experiments. Borda ranks examples within each seed in descending score order, so higher proxy-prepared scores receive better ranks; ties use average ranks before averaging ranks across seeds. This comparison is important because strong seed-only aggregation can be an upper baseline for pure stability when enough repeated proxy runs are available.

Special cases used in our experiments. The framework subsumes several informative variants:

- **Bare proxy:** no structure-aware stage and no cross-seed aggregation.
- **Seed-only aggregation:** bypass the structure-aware stage and apply only cross-seed aggregation. We report seed-median as the default robust baseline and also evaluate seed-mean and seed-Borda as stronger pure-stability variants.
- **Cluster-only aggregation:** apply the within-seed structural stage but do not aggregate across seeds.
- **Full SCARV:** use both the structure-aware and cross-seed stages under the submitted default preprocessing.

This nested formulation separates three possible sources of improvement: the base proxy itself, robust multi-seed aggregation, and redundancy-aware structural regularization. For causal mechanism claims about structure, we use matched pairs that hold preprocessing and cross-seed operator fixed: Struct-Mean versus Seed-Mean, Struct-Median versus Seed-Median, and Struct-Borda versus Seed-Borda.

Scope. SCARV uses a modular proxy interface: any scoring rule ϕ can be plugged in as long as it produces comparable per-example scores across seeds, but empirical support in this paper is limited to the proxy families and task types we evaluate. In our experiments, we instantiate this interface with loss-based, influence-style, hybrid, margin, and confidence-based proxies. We evaluate valuation alignment, ranking stability, and selection utility separately. SCARV is designed primarily to improve ranking stability while tracking the semantics of an informative base proxy; it does not assume that stabilization alone will optimize downstream selection utility.

4 Experiments

We evaluate SCARV along five questions: (1) how stabilization interacts with the semantics of the base proxy, (2) whether aggregation improves ranking stability under redundancy, (3) what part of the framework is responsible for the gain under matched comparisons, (4) when the structure-aware stage is worth its map-construction cost, and (5) whether improved stability supports reproducible and group-consistent audit decisions. Following our framing, we treat valuation alignment,

ranking stability, and downstream utility as related but distinct targets.

4.1 Experimental Setup

We use SST-2, MRPC, MNLI, and QQP from GLUE (Wang et al., 2018), together with AG News (Zhang et al., 2015). Unless otherwise stated, the controlled experiments use TF-IDF features with logistic regression, which allows broad sweeps over redundancy, proxy, clustering, and seed-budget conditions. To address whether the phenomenon is an artifact of shallow models, we additionally run an end-to-end fine-tuning suite with distilbert-base-uncased on SST-2, MRPC, QQP, and MNLI. We study exact duplicates, synthetic near-duplicates, and naturally mined redundancy. The audited default synthetic regime uses redundancy rate $r = 0.3$ and perturbation strength $s = 0.6$.

Unless noted otherwise, reported main-text numbers use $R = 5$ internal aggregation seeds. For aggregation methods, one outer run produces a final ranking from a group of R internal scoring seeds, and stability is then computed by comparing final rankings across outer runs rather than by treating seed-pair correlations within the same group as independent observations. We separate three quantities throughout: stability-aware aggregation over bare rankings, the submitted Full-SCARV pipeline, and the matched structural increment obtained by toggling structure while holding preprocessing and cross-seed operator fixed.

Alignment. We use leave-one-out (LOO) alignment only as a diagnostic rather than as a primary target. Appendix results show a mixed picture: TracIn-style scores are generally more positively aligned with LOO than loss or margin proxies, while stability-aware aggregation can improve, preserve, or reduce LOO alignment depending on the dataset and proxy. We therefore do not treat SCARV as a method for improving valuation alignment. Its role is to stabilize a chosen proxy-induced ranking.

4.2 Stable Ranking under Synthetic Redundancy and Across Tasks

Table 1 reports the main stability results. Two patterns are consistent. First, seed aggregation and full SCARV substantially outperform the bare proxy across tasks, confirming that cross-seed aggregation stabilizes redundant-data rankings rather than

exposing a dataset-specific artifact. Second, the relative ordering between seed-median and full SCARV is regime dependent. In the audited SST-2 synthetic setting, the default seed-median baseline is stronger than full SCARV for the loss proxy, whereas on MRPC and MNLI full SCARV is competitive with or slightly above seed-median.

These results substantially weaken the interpretation that ranking instability is confined to a single toy setting. The same qualitative trend transfers beyond the original SST-2/AG News setup to paraphrase matching (MRPC) and natural language inference (MNLI). At the same time, Table 1 illustrates the central mechanism of the paper: full SCARV is reliably above the bare proxy, but its comparison to seed-median is regime-dependent. We therefore treat the method as a stability-oriented aggregation layer rather than as a universally dominant replacement for the seed-median baseline.

4.3 Full-Pipeline Mechanism Analysis

We next test whether the gains are artifacts of the original fixed-slice design and what part of the method is actually responsible for them. Under five stratified subset draws and five redundancy-layout seeds on SST-2 and AG News, mean stability increases from 0.7556 for the bare proxy to 0.8970 for full SCARV and 0.9012 for seed-median. At a selection budget of 0.3, subset overlap rises from 0.5415 to 0.6699 for SCARV and 0.6803 for seed-median, while the selection gap drops from 0.0309 to 0.0232 and 0.0239, respectively. Thus, the stability story is robust to subset-draw and layout variation, but this comparison alone does not isolate the structural module.

Table 2 reports the full-pipeline mechanism decomposition. It compares the complete default Full-SCARV pipeline with bare, seed-only, cluster-only, and dedup-first alternatives under the full-pipeline protocol. This overall-pipeline table is the appropriate main-text companion to the main cross-task results, but it should not be read as an isolated estimate of the structure-aware allocation step.

The decomposition shows that the default full pipeline is meaningfully above bare, cluster-only, and dedup-first baselines, while strong seed-only aggregation remains a difficult comparator. A separate operator-controlled matched diagnostic asks a narrower causal question by toggling only hard cluster allocation while holding preprocessing, score bank, seeds, rank direction, and cross-seed operator fixed. That diagnostic supports the

Setting	Proxy	Bare	Seed-med.	Full SCARV
SST-2 (synthetic)	Loss	0.6040	0.8422	0.8018
SST-2 (synthetic)	TracIn-style	0.9564	0.9769	0.9774
MRPC (TF-IDF)	Loss	0.8092	0.9223	0.9368
MRPC (TF-IDF)	TracIn-style	0.8998	0.9276	0.9351
MNLI (TF-IDF)	Loss	0.3337	0.6330	0.6395
MNLI (TF-IDF)	TracIn-style	0.8793	0.9316	0.9388

Table 1: Global ranking stability across tasks. Values are mean pairwise Spearman correlations across random seeds. In the default synthetic SST-2 regime, seed-median is strongest on the loss proxy; on MRPC and MNLI, full SCARV is competitive with or slightly above seed-median.

Method	Mean Spearman
Approx. dedup-then-rank	0.7350
Cluster-only	0.7861
Bare	0.7880
Full SCARV	0.8899
Seed-median	0.9159
Seed-Borda	0.9445
Seed-mean	0.9470

Table 2: Full-pipeline mechanism decomposition. Values are mean Spearman stability. This table compares complete pipelines; operator-controlled structural-increment diagnostics are discussed separately.

same qualitative conclusion of no general structural dominance, but it has a different score-bank and sample-universe scope from the full-pipeline table. We therefore use it to qualify the causal interpretation rather than to replace the main mechanism result; Appendix Table 7 reports the additional controlled protocol. The Full-SCARV versus Seed-median significance table is retained only as an overall-pipeline comparison in the appendix.

4.4 Natural Redundancy

A central concern for redundancy-based methods is whether they work only under synthetic perturbations. Table 3 addresses this issue on QQP. Compared with the earlier weak SST-2 natural setting, naturally mined QQP clusters provide a more informative test bed: sentence-embedding clustering covers 6.5% of the training set, with 74 non-singleton clusters, mean cluster size 2.65, and maximum size 8. Although this coverage is still modest, it is non-trivial and sufficient for testing whether the structure-aware stage helps under naturally mined redundancy.

In this regime, full SCARV can exceed the default seed-median pipeline across the three cluster-discovery procedures and both proxy families. This

result shows that the paper is not restricted to synthetic duplicate injection, but it should not be interpreted as structural dominance over seed-only aggregation. The submitted full-pipeline QQP comparison is positive, whereas the operator-controlled matched structural effect is small and protocol-dependent. Moreover, non-singleton coverage is at most 6.5%, so global Spearman is diluted by many singleton examples. We therefore report covered-only diagnostics and proxy-fidelity checks in the extended analysis: global stability that does not collapse under a natural map is not, by itself, evidence that the map is semantically correct.

4.5 Seed-Budget Frontier and Transformer Robustness

A natural objection is that SCARV may be unnecessary once stronger seed-only baselines are allowed, or that the observed instability may be an artifact of shallow TF-IDF models. We address both concerns with two additional audited experiments.

First, we run a seed-budget frontier with $R \in \{1, 2, 3, 5, 7, 10\}$, comparing full SCARV against a post-hoc best seed-only upper baseline under the same aggregation budget. This table is a seed-count diagnostic, not a complete compute frontier: it varies the number of proxy seeds but does not by itself charge the end-to-end cost of feature extraction, map construction, map loading, diversity kNN, structural allocation, and metric evaluation. A separate whole-machine-idle timing study in the appendix distinguishes fresh-map, amortized-map, and existing-map regimes and shows that low seed count should not be equated with low total compute.

Second, we run end-to-end distilbert-base-uncased fine-tuning on SST-2, MRPC, QQP, and MNLI. The transformer results confirm that the instability and stabilization

Cluster mining	Coverage (%)	Loss Seed-med. / Full	TracIn Seed-med. / Full
TF-IDF cosine	2.6	0.936 / 0.948	0.907 / 0.922
Sentence-embedding	6.5	0.936 / 0.948	0.907 / 0.923
Lexical char-gram	6.1	0.936 / 0.948	0.907 / 0.924

Table 3: Natural redundancy on QQP. Coverage is the fraction of training examples in non-singleton clusters. Values compare the submitted seed-median and Full-SCARV pipelines; matched structural deltas are smaller and are discussed separately.

R	Full > best upper	Mean Δ_{best}	Median Δ_{best}
1	16	+0.0180	+0.0233
2	15	+0.0095	+0.0126
3	0	-0.0374	-0.0335
5	0	-0.0374	-0.0307
7	0	-0.0318	-0.0251
10	0	-0.0187	-0.0138

Table 4: Seed-budget frontier. Δ_{best} is Full SCARV Spearman minus the post-hoc strongest seed-only method among seed-mean, seed-median, and seed-Borda under the same seed count. End-to-end wall-clock conclusions require separate map-cost accounting.

effects are not artifacts of TF-IDF/logistic regression. Averaged over confidence, loss, and margin proxies, full SCARV improves Spearman stability from 0.893 for bare rankings to 0.943, and it also improves local Jaccard overlap and selected-subset reproducibility. Full SCARV is essentially tied with seed-median on average. However, seed-mean and seed-Borda remain stronger pure-stability baselines, reaching about 0.975 average Spearman in the best seed-only upper-baseline column.

4.6 Practical Value: Reproducible Audit Queues

The strongest practical benefit of stabilization is decision reproducibility, not a universal increase in mean utility. Figure 7 visualizes this result. Under subset selection, seed-median and full SCARV substantially increase overlap between selected subsets across runs and reduce the selection gap, even when mean utility changes only modestly. In the audited SST-2 setting with the loss proxy at budget 0.3, subset overlap rises from 0.4229 for bare to 0.6168 for seed-median and 0.5306 for full SCARV, while the selection gap falls from 0.0293 to 0.0200 and 0.0205, respectively. These results support the interpretation that stability primarily improves the repeatability of ranking-based decisions.

The audit-queue interpretation makes the role of structure more concrete. For noisy-label retrieval,

Scenario	Recommended practice
Pure global stability and several proxy seeds are affordable	Prefer Seed-Mean or Seed-Borda.
Redundancy-map quality is unknown	Use seed-only aggregation.
Trusted map exists and audit decisions require group consistency	Consider structure-aware regularization and report group metrics.
Cheap proxy plus fresh semantic map	Usually add seeds before paying map-construction cost.
Hard replacement creates many ties	Analyze soft shrinkage and tie-aware metrics.
Goal is accuracy, AU-ROC, or pruning utility	Validate directly; do not infer utility from stability.

Table 5: Deployment guideline implied by the matched and diagnostic analyses. The structure-aware stage is optional and conditional, not a universal replacement for seed-only aggregation.

seed-median gives the best AUROC in the audited SST-2 setting, improving from 0.6804 for bare to 0.7194, while full SCARV reaches 0.6919. Full SCARV does not dominate AUROC, but it produces the most reproducible suspicious-example sets: the overlap of flagged examples rises from 0.2636 for bare to 0.4270 for seed-median and 0.4664 for full SCARV. Direct audit-queue diagnostics further separate item-level and group-level behavior. Hard structure can nearly eliminate split groups at the audit cutoff, but it also increases the burden of reviewing multiple members of the same redundancy group. Appendix Table 8 quantifies this trade-off: structure improves group consistency but reduces item-level overlap relative to matched Seed-Median and increases redundant-member review. The supported use case is therefore not universal detection quality; it is group-consistent auditing when a trusted redundancy map already exists and repeated review of related examples is operationally acceptable.

Overall, the experiments support five conclusions. First, stability-aware aggregation substantially improves over bare proxy-induced rankings under redundancy. Second, the dominant generic

mechanism is robust cross-seed aggregation rather than structure alone. Third, the matched structural increment is small, regime dependent, and sometimes negative; the submitted QQP full-pipeline comparison is positive, but matched structural effects remain small and protocol-dependent. Fourth, low seed count should not be equated with low total compute, because map construction and caching state matter. Fifth, the strongest practical benefit of the structural stage is conditional group-consistent auditing, with a trade-off between group consistency and review diversity.

5 Conclusion

SCARV targets stable sample-level ranking under redundancy. Rather than redefining sample value or replacing a chosen proxy, it regularizes how proxy-induced scores behave when duplicates and near-duplicates interact with stochastic training. Across synthetic and naturally mined redundancy, multiple proxy families, classification and sentence-pair tasks, and an end-to-end DistilBERT suite, stability-aware aggregation improves over bare proxy rankings and supports more reproducible ranking-based decisions.

Our mechanism analysis clarifies scope: robust cross-seed aggregation is the dominant generic stabilizer, while the incremental effect of structural regularization is small, map-dependent, dataset-dependent, and sometimes negative. Structure can be considered when the redundancy map is meaningful and group-consistent treatment is operationally desired; a small seed budget alone does not guarantee a benefit, especially once map-construction cost is included. These findings motivate a modular data-centric workflow: select a proxy for the downstream purpose, use seed-only aggregation as the default stabilizer, and add structure only when its operational group-consistency trade-off is justified. SCARV is not a universal data selector and is not claimed to uniformly improve downstream utility metrics; its practical value is repeatability and, in trusted-map settings, auditability of group-level decisions. Our empirical evidence is limited to classification and sentence-pair tasks and does not establish transfer to generative, long-document, or pretraining settings.

Limitations

Stability is not utility. The goal of this work is to stabilize proxy-induced sample rankings and

improve the reproducibility of ranking-based decisions; it does not guarantee better downstream selection accuracy, AUROC, or pruning performance. The structure-aware stage depends on redundancy maps: if clusters are noisy, biased, or semantically invalid, structural aggregation can over-smooth or stabilize the wrong signal. Matched comparisons show that structure is not a universal improvement over seed-only aggregation, and hard replacement can create ties that require tie-aware reporting. The post-hoc best seed-only upper baseline is a conservative stress-test reference for pure stability and is not necessarily a deployable rule known in advance. Our experiments focus on classification and sentence-pair datasets; large-scale pretraining, instruction tuning, multilingual corpora, and production-scale data curation remain future work. Current implementations use relatively simple redundancy maps and allocation rules; soft clusters, uncertainty-aware maps, and adaptive shrinkage are not fully explored. Valuation alignment, such as LOO, is used diagnostically and may vary by proxy and dataset.

Ethical Considerations

Automated data filtering and suspicious-example retrieval can incorrectly flag valid samples. Stable rankings can still encode biases inherited from the base proxy or dataset, and removing or down-weighting examples may disproportionately affect minority, rare, ambiguous, or difficult cases. Human review is necessary before high-stakes data deletion. The public repository includes the core implementation, tests, configuration examples, and a synthetic demo, but does not redistribute raw benchmark text when license or redistribution terms are unclear. Experiments use public NLP benchmarks and do not involve new human-subject data collection.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62441226) and the 111 Project (No. B16037).

References

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. 2024. Training data attribution via approximate unrolling. In *Advances in Neural Information Processing Systems*.

- Angelo Basile, Marc Franco-Salvador, and Paolo Rosso. 2023. Zero-shot data maps: Efficient dataset cartography without model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*.
- Junwei Deng, Weijing Tang, and Jiaqi W. Ma. 2025. A versatile influence function for data attribution with non-decomposable loss. In *Proceedings of the 42nd International Conference on Machine Learning*.
- Amirata Ghorbani, Michael Kim, and James Zou. 2020. A distributional framework for data valuation. In *Proceedings of the 37th International Conference on Machine Learning*.
- Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*.
- Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. 2023. Measuring forgetting of memorized training examples. In *International Conference on Learning Representations*.
- William Lopez Jaramillo and Evgueni Smirnov. 2021. Shapley-value based inductive conformal prediction. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*.
- Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gürel, Bo Li, Ce Zhang, Costas J. Spanos, and Dawn Song. 2019a. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*.
- Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019b. Towards efficient data valuation based on the shapley value. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.
- Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*.
- Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*.
- Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. DataInf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *International Conference on Learning Representations*.
- Yongchan Kwon and James Zou. 2022. Beta shapley: A unified and noise-reduced data valuation framework for machine learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Danny D. Leybzon and Corentin Kervadec. 2024. Learning, forgetting, remembering: Insights from tracking llm memorization during training. In *Proceedings of the 7th BlackboxNLP Workshop*.
- Weida Li and Yaoliang Yu. 2023. Robust data valuation with weighted banzhaf values. In *Advances in Neural Information Processing Systems*.
- Mathieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. 2024. Copyright traps for large language models. In *Proceedings of the 41st International Conference on Machine Learning*.
- Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*.
- Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*.
- Victor Quintas-Martinez, Mohammad Taha Bahadori, Eduardo Santiago, Jeff Mu, and David Heckerman. 2024. Multiply-robust causal change attribution. In *Proceedings of the 41st International Conference on Machine Learning*.
- Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. 2022. CS-Shapley: Class-wise shapley values for data valuation in classification. In *Advances in Neural Information Processing Systems*.
- Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.
- Mariya Toneva, Alessandro Sordoni, Rémi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*.

Jiachen T. Wang and Ruoxi Jia. 2023. Data banzhaf: A robust data valuation framework for machine learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*.

Jiachen T. Wang, Prateek Mittal, and Ruoxi Jia. 2024a. Efficient data shapley for weighted nearest neighbor algorithms. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*.

Jiachen T. Wang, Tianji Yang, James Zou, Yongchan Kwon, and Ruoxi Jia. 2024b. Rethinking data shapley for data selection tasks: Misleads and merits. In *Proceedings of the 41st International Conference on Machine Learning*.

Jiachen T. Wang, Yuqing Zhu, Yu-Xiang Wang, Ruoxi Jia, and Prateek Mittal. 2023. [Threshold knn-shapley: A linear-time and privacy-friendly approach to data valuation](#). In *Advances in Neural Information Processing Systems*, volume 36.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K. Ravikumar. 2018. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*.

Jinsung Yoon, Serkan Arik, and Tomas Pfister. 2020. Data valuation using reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*.

A Artifact and Reproducibility Details

The public code repository is available at <https://github.com/Sn0wm1an/SCARV>. It provides the core SCARV implementation, tests, configuration examples, and a synthetic demonstration. Full benchmark reproduction requires independently obtaining the datasets and rerunning the experiments under the protocol described in this paper.

A.1 Contents

The repository includes code for global score normalization, optional diversity preparation, structure-aware cluster allocation, cross-seed aggregation, ranking utilities, and a small synthetic demo. It does not include raw benchmark text, saved per-example score banks, model checkpoints, or paper-table result archives.

A.2 Data Access

We do not redistribute raw dataset text. Users obtain datasets from official sources subject to their licenses. The scripts expect standard paths or download via common libraries that require the user to accept the relevant dataset terms.

A.3 Randomness and Determinism

Exact bitwise determinism is not guaranteed. We fix seeds where appropriate and report stability across outer runs. The key objective is the reproducibility of ranking-based decisions rather than identical seeds.

A.4 Environment Notes

Experiments rely on common Python/NLP stacks, including tokenizers, linear models, and transformer libraries for the fine-tuning suite. The public repository keeps the core dependency set lightweight; full benchmark reruns require additional experiment dependencies.

A.5 Manifest

The repository README documents installation, unit tests, and the toy-demo command. The benchmark configurations and reporting details in this appendix specify the full experimental protocol, but the public repository is a core-code release rather than a saved-result archive.

B Data and License Notes

We use public NLP benchmarks, including GLUE tasks and AG News. We do not redistribute raw text; users obtain each dataset from its official source under the corresponding license.

B.1 Licenses and Redistribution

To avoid ambiguity, the public repository contains the core implementation, tests, configuration examples, and a synthetic demonstration. It does not contain raw benchmark text, saved per-example score banks, model checkpoints, or paper-table result archives.

B.2 Dataset-Specific Notes

GLUE tasks follow the official GLUE license terms. AG News follows the dataset provider’s terms. The artifact README and third-party notices describe the expected loader paths and redistribution caveats.

Component	Logged / estimated runtime
TF-IDF/logistic audited extension	1.65 wall-clock hours
Seed-budget frontier experiment	2.23 logged wall-clock hours
LOO diagnostic	0.03 wall-clock hours
DistilBERT FT training	3.16 GPU-hours lower bound
DistilBERT FT postprocessing	0.016 wall-clock hours
Figure/table generation	< 0.1 CPU-hours estimate

Table 6: Conservative compute accounting. Values are logged execution-time sums or lower-bound estimates rather than scheduler-level CPU/GPU accounting. The seed-budget frontier entry records the total experiment log time, not a per-method end-to-end Pareto cost.

B.3 Responsible Use

Ranking-based filtering can remove or down-weight examples that are ambiguous or minority. Human review is advised for high-stakes decisions.

C Compute Budget and Environment

C.1 Budget Accounting

We report compute using conservative accounting of logged training time aggregated across outer runs and internal aggregation seeds. The budget table reports logged experiment execution time or lower-bound estimates; it is not the per-configuration timing used in the separate wall-clock Pareto diagnostic.

C.2 Environment

Controlled TF-IDF/logistic experiments run primarily on CPU. The end-to-end DistilBERT suite uses the available CUDA GPU environment recorded in the artifact logs. The final wall-clock diagnostic is measured separately under a whole-machine-idle protocol on one RTX 3090 device, with CUDA, PyTorch, thread counts, CPU information, and MiniLM cache state recorded in the timing manifest.

C.3 Reproducibility Scope

The goal of this paper is to stabilize ranking under redundancy. We therefore report stability across outer runs and focus on reproducibility of ranking-based decisions rather than exact bitwise determinism.

D Responsible NLP Checklist Mapping

Typical checklist items map to the artifact and paper as follows: datasets and sources are discussed in this appendix; models and pretraining are described in Appendix E; hyperparameters and budgets are described in the compute notes and Appendix E; metrics and statistical testing are described in Appendix E; limitations and ethics appear in the main text; and artifact availability is described in Appendix A.

E Additional Experimental Details

E.1 Result Sources and Reporting Scope

The paper keeps result scopes separate. Legacy reproduction outputs, submitted-pipeline tables, and later controlled diagnostics are labeled by their protocol rather than merged into a single summary. Generated tables and figures are produced from saved CSV/JSON outputs by the accompanying scripts.

The default synthetic near-duplicate operating point is redundancy rate $r = 0.3$ and perturbation strength $s = 0.6$. Sequential head-slice runs are treated as reproduction checks, while robustness claims are based on runs that vary subset draws and redundancy layouts. Tables and figures use programmatic generators rather than manual cell entry.

E.2 Datasets, Models, and Seeds

We use SST-2, MRPC, MNLI, and QQP from GLUE, together with AG News. The default setting uses TF-IDF features with logistic regression. We additionally report a small MLP and frozen DistilBERT feature settings in the extended experiments. For the strongest appendix-only feature setting, DistilBERT embeddings are extracted once and a linear classifier is trained on top.

End-to-end DistilBERT fine-tuning. For the transformer fine-tuning experiment, we fine-tune `distilbert-base-uncased` end-to-end on SST-2, MRPC, QQP, and MNLI subsets. The requested subset sizes are SST-2 train/validation 5000/1000, MRPC 3500/600, QQP 10000/1500, and MNLI 10000/2000; because the underlying validation splits are smaller for SST-2 and MRPC, the realized validation sizes are 872 and 408, respectively. Across all four tasks, we use 3 training epochs, batch size 16, learning rate 2×10^{-5} , maximum sequence length 128, AdamW with weight

decay 0.01, and a linear warmup-plus-decay schedule with warmup steps set to one tenth of the total training steps. We use three outer runs with seeds {11, 23, 37}, and each outer run forms a final aggregated ranking from five internal scoring seeds {42, 123, 456, 789, 1024}, i.e., aggregation budget $R = 5$. Checkpoints are selected by best validation accuracy, and the reported per-example rankings are extracted from that best checkpoint using loss, margin, and confidence proxy families.

We distinguish between outer experimental runs and internal aggregation seeds. Unless otherwise noted, the main stability experiments use $R = 5$ aggregation seeds. For significance-critical comparisons, we additionally run 10 outer experimental runs and report paired confidence intervals and paired nonparametric tests over these outer runs.

E.3 Redundancy Construction

Exact duplicates. For exact-duplicate experiments, we sample source examples from the clean training set and append duplicate copies at a target redundancy rate r . In the robustness experiments, both the subset draw and the redundancy layout are varied explicitly.

Synthetic near-duplicates. Synthetic near-duplicates follow the same source-sampling rule as exact duplicates, but each copied text is perturbed before re-featurization. The perturbation pipeline applies a stochastic combination of word dropout, local word swaps, word repetition, intra-sentence replacement, and segment shuffle. The default audited operating point is $r = 0.3$ and $s = 0.6$, and the appendix further reports sweeps over redundancy rate, perturbation strength, cluster size, cluster noise, and aggregation budget.

Natural redundancy. The main natural-redundancy showcase is QQP. We consider three cluster-discovery procedures: TF-IDF cosine thresholding, sentence-embedding similarity, and lexical character n-gram similarity. The resulting clusters are approximate rather than oracle-quality, which makes this setting a more realistic test of redundancy-aware ranking.

E.4 Proxies and SCARV Instantiations

The base proxies are a loss-based proxy, a TracIn-style proxy, a hybrid proxy used in the method-decomposition study, and a margin proxy used for proxy-generalizability checks. The default full SCARV instantiation follows the main-text method section:

within-seed global min-max normalization over the complete scored subset for each seed, the frozen diversity default $(w, k) = (0.2, 10)$, hard cluster mean replacement, and cross-seed median aggregation. To clarify the mechanism, we separate submitted-pipeline alternatives from matched operator comparisons. Structural-increment claims are based on Struct-Mean versus Seed-Mean, Struct-Median versus Seed-Median, and Struct-Borda versus Seed-Borda under identical preprocessing and seeds.

E.5 Evaluation Metrics and Statistical Testing

We report four families of metrics. Alignment is measured against LOO with Spearman correlation. Global ranking stability is computed from pairwise Spearman correlations across seed-specific rankings for single-seed methods and across final outer-run rankings for aggregated methods. Local stability is measured analogously with top- k and bottom- k Jaccard overlap. Decision reproducibility is measured with selected-subset overlap, selection gap, and overlap of suspicious-example sets. Downstream utility is reported as retained-subset test accuracy for subset selection and AU-ROC/AUPRC/precision/recall for noisy-label retrieval.

All significance tests are paired over outer runs. Pairwise seed-pair Spearman values are first summarized within each outer run and are not treated as independent samples. For matched structural comparisons, the accompanying diagnostic artifacts compute paired intervals and paired nonparametric tests over outer runs. The Full-SCARV versus Seed-median table is retained as an overall-pipeline comparison only; it is not used as evidence that the structure module alone has an independent effect.

For ranking metrics with ties, the reported diagnostic correlations use average ranks; deterministic example-id secondary sorting is reserved only for operational exported queues. We also keep protocol scopes separate in the appendix. The seed-budget frontier varies the number of proxy seeds and does not charge fresh map construction. The wall-clock diagnostic, by contrast, charges map construction or loading and is reported as a separate timing study. Smoke runs, blocked wall-clock attempts, and recovery-only manifests are not used as paper evidence.

F Additional Results and Sensitivity Analyses

This section reports appendix-level diagnostics that stress-test the interpretation established in the main text. Across the analyses below, the same picture persists: stability-aware aggregation is robust to subset and layout variation, transfers beyond the default SST-2 configuration, and improves the reproducibility of ranking-based decisions. At the same time, the appendix sharpens the mechanism story: robust multi-seed aggregation is the dominant generic stabilizer, while the incremental structural effect is small, regime dependent, and sometimes negative under matched comparisons.

F.1 Alignment to Leave-One-Out

The audited leave-one-out checks are diagnostic rather than central evidence. They support two limited conclusions. First, proxy choice matters: TracIn-style scores are generally more positively aligned with LOO than loss or margin scores, which are often weakly or negatively aligned. Second, stabilization and valuation alignment are not the same objective. Full SCARV substantially improves ranking stability, but its effect on LOO alignment is mixed across datasets and proxies. We therefore interpret SCARV as a stability-oriented aggregation layer rather than as a replacement valuation functional.

F.2 Mechanism Decomposition and Stress Tests

The main text already reports the central mechanistic result: robust multi-seed aggregation is the primary source of stability gain in the default synthetic regime. Here we expand that comparison to include cluster-only variants, dedup-first baselines, and multiple synthetic stress regimes.

F.3 Matched Structural-Increment Diagnostic

The matched factorial is included as a diagnostic rather than as a replacement for the full-pipeline decomposition. Structure OFF and Structure ON share the same preprocessing, score bank, seeds, rank direction, and cross-seed operator; only hard cluster allocation changes. This diagnostic answers a different question from the default Full-SCARV versus seed-only pipeline comparisons. It supports the qualitative conclusion that most generic stability gain should not be attributed to structural allocation alone, while exact numeric values depend

on the controlled score bank and sample-universe scope.

F.4 Map Quality, Diversity, and Tie Diagnostics

Map-degradation experiments are designed to avoid a common false reassurance: a nominally corrupted map may leave structured scores and final rankings nearly unchanged. The diagnostic pipeline therefore records map hashes, changed pair relations, cluster-size distributions, coverage, pairwise overlap, ARI/NMI where appropriate, structured-score change rates, and final-rank change rates before reporting stability. Structure-OFF baselines are required to remain invariant to corruption. The key qualitative result is that all-example metrics can look flat when most examples are singletons, especially for QQP natural redundancy; covered-only stability, proxy fidelity, and group-level metrics are more sensitive. The current evidence does not support a universal clustering-purity threshold.

The diversity grid treats $(w, k) = (0.2, 10)$ as a frozen default rather than a tuned optimum. Increasing w can improve stability, but it also moves the prepared scores farther from the raw proxy and can reduce retained utility. The k choice has a smaller effect in the tested grid. Thus the diversity term is best interpreted as an optional stability–fidelity regularizer, not a necessary component justified by stability alone.

Hard replacement creates exact within-cluster ties. Tie-aware Spearman ranks, Kendall τ_b , strict/inclusive top- k , and random cutoff tie-breaking diagnostics are therefore reported separately from deterministic operational rankings. Soft shrinkage with $\lambda < 1$ is used as a diagnostic to test whether near-hard stability can be obtained with fewer ties and less proxy distortion. These analyses prevent the paper from interpreting stability gains caused by ties as evidence of a strictly better item-level ranking.

Across these diagnostics, the appendix reports negative and mixed outcomes in the same format as positive outcomes. Settings are not removed because structure is weak, because a map is noisy, or because a stability gain is accompanied by lower fidelity. Diagnostic results qualify the mechanism claim, while submitted-pipeline tables remain separate overall-pipeline evidence.

F.5 Numeric Diagnostic Summaries

The following tables give compact numeric summaries for the additional diagnostics. They are reported separately from the full-pipeline main tables because they use controlled diagnostic protocols with different score-bank or sample-universe scopes.

Protocol	Panel	Operator	Struct–Seed Δ	95% CI	Wilcoxon p	Holm p	n_{outer}
Final controlled	$w = 0$	Mean	+0.000802	[-0.000254, +0.001857]	0.312	0.938	5
Final controlled	$w = 0$	Median	+0.000858	[-0.000166, +0.001882]	0.438	0.938	5
Final controlled	$w = 0$	Borda	+0.000263	[-0.000240, +0.000771]	0.625	0.938	5
Final controlled	$w = .2, k = 10$	Mean	-0.016332	[-0.018914, -0.013475]	0.062	0.375	5
Final controlled	$w = .2, k = 10$	Median	-0.016432	[-0.018578, -0.014021]	0.062	0.375	5
Final controlled	$w = .2, k = 10$	Borda	-0.016943	[-0.018459, -0.015307]	0.062	0.375	5

Table 7: Operator-controlled matched structural diagnostic. This table reports an additional controlled protocol whose score bank and sample universe differ from the full-pipeline main results. Both protocols support the same qualitative conclusion of no general structural dominance. For each contrast, setting-level deltas are first macro-averaged within each outer run. A Wilcoxon signed-rank test is then applied to the five outer-run values, and Holm correction is applied across the six resulting p -values.

Contrast	Metric	Δ	95% CI	Unit
Struct-Median–Bare	Item queue overlap	+0.060639	[+0.051453, +0.069826]	Jaccard/overlap delta
Struct-Median–Seed-Median	Item queue overlap	-0.022607	[-0.025274, -0.019940]	Jaccard/overlap delta
Struct-Median–Seed-Median	Split-group rate	-0.107374	[-0.113988, -0.099570]	rate delta
Struct-Median–Seed-Median	Redundant burden	+26.275	[+25.244, +27.512]	extra selected examples
Struct-Median–Seed-Median	Burden rate	+0.048202	[+0.045720, +0.050725]	extra selected examples / queue size

Table 8: Audit-queue trade-off under the final controlled diagnostic. Queue metrics are macro-averaged over the predefined budgets $q \in \{1\%, 5\%, 10\%, 20\%\}$ and available dataset/proxy settings. Burden is reported both as extra selected examples and as a fraction of queue size.

Diagnostic	Contrast	Stability Δ	Fidelity Δ	Additional metric
Map degradation	corruption > 0 – level 0	-0.000195 [-0.000421, +0.000068]	-0.008220 [-0.008507, -0.007933]	covered-only diagnostics in artifact
Diversity	$w = .2 - w = 0$	+0.054892 [+0.047502, +0.060481]	-0.045798 [-0.049120, -0.041263]	accuracy +0.002696 [-0.001929, +0.010735]
Allocation	$\lambda = 1 - \lambda = .75$	-0.004012 [-0.004507, -0.003414]	-0.009893 [-0.010129, -0.009652]	tie fraction +0.247968 [+0.246762, +0.249113]

Table 9: Boundary diagnostics for map corruption, diversity, and hard allocation. Values are deltas relative to the stated baseline; each cell reports the mean effect and 95% paired bootstrap interval.



Figure 3: Sensitivity to aggregation rule. Adding multi-seed aggregation is the largest source of stability gain, while the choice among mean, median, and Borda still matters.

Regime	Proxy	Bare	Full	Seed-M	Seed-Med	Borda
base	hybrid	0.6845	0.8502	0.9224	0.8799	0.9207
base	loss	0.6040	0.8018	0.8920	0.8422	0.8880
base	TracIn	0.9564	0.9774	0.9922	0.9769	0.9914
noise25	hybrid	0.7412	0.8627	0.9370	0.8986	0.9344
noise25	loss	0.6724	0.8339	0.9187	0.8769	0.9138
noise25	TracIn	0.9642	0.9851	0.9930	0.9834	0.9923
high-rate	hybrid	0.7636	0.8727	0.9444	0.9117	0.9433
high-rate	loss	0.7024	0.8379	0.9266	0.8913	0.9245
high-rate	TracIn	0.9781	0.9873	0.9940	0.9868	0.9935
clust4	hybrid	0.7412	0.8626	0.9370	0.8986	0.9344
clust4	loss	0.6724	0.8326	0.9187	0.8769	0.9138
clust4	TracIn	0.9642	0.9849	0.9930	0.9834	0.9923
clust8	hybrid	0.7392	0.8615	0.9316	0.8924	0.9278
clust8	loss	0.6656	0.8229	0.9130	0.8665	0.9058
clust8	TracIn	0.9710	0.9806	0.9916	0.9799	0.9913
low-sim	hybrid	0.7279	0.8564	0.9342	0.8923	0.9315
low-sim	loss	0.6702	0.8300	0.9162	0.8731	0.9114
low-sim	TracIn	0.9654	0.9775	0.9909	0.9748	0.9901

Table 10: Method decomposition, aggregation panel (Spearman stability, higher is better).

Regime	Proxy	A.Dup	O.Dup	C.Coll	C.Only	C.Div
base	hybrid	0.6668	0.6508	0.6508	0.6804	0.6296
base	loss	0.5330	0.6008	0.6008	0.6000	0.5393
base	TracIn	0.9498	0.9739	0.9739	0.9567	0.9584
noise25	hybrid	0.6776	0.6701	0.6701	0.7425	0.6729
noise25	loss	0.5669	0.5588	0.5588	0.6740	0.5892
noise25	TracIn	0.9676	0.9585	0.9585	0.9642	0.9665
high-rate	hybrid	0.6556	0.7039	0.7039	0.7598	0.6935
high-rate	loss	0.5866	0.5517	0.5517	0.6969	0.6078
high-rate	TracIn	0.9762	0.9663	0.9663	0.9785	0.9759
clust4	hybrid	0.6776	0.6723	0.6723	0.7391	0.6731
clust4	loss	0.5669	0.5570	0.5570	0.6691	0.5873
clust4	TracIn	0.9676	0.9512	0.9512	0.9644	0.9660
clust8	hybrid	0.6960	0.6810	0.6810	0.7370	0.6778
clust8	loss	0.6136	0.5635	0.5635	0.6621	0.5839
clust8	TracIn	0.9749	0.9673	0.9673	0.9710	0.9701
low-sim	hybrid	0.6684	0.6725	0.6725	0.7237	0.6554
low-sim	loss	0.5699	0.5577	0.5577	0.6653	0.5897
low-sim	TracIn	0.9143	0.9512	0.9512	0.9659	0.9660

Table 10 continued. Deduplication and cluster-baseline panel. A.Dup/O.Dup are approximate/oracle dedup-then-rank variants.

Regime	Proxy	Medoid	Uniform	Soft	Random
base	hybrid	0.6987	0.6804	0.6820	0.0041
base	loss	0.6231	0.6000	0.6018	0.0041
base	TracIn	0.9556	0.9567	0.9566	0.0041
noise25	hybrid	0.7816	0.7425	0.7424	0.0071
noise25	loss	0.7239	0.6740	0.6748	0.0071
noise25	TracIn	0.9691	0.9642	0.9642	0.0071
high-rate	hybrid	0.7773	0.7598	0.7616	0.0063
high-rate	loss	0.7259	0.6969	0.6995	0.0063
high-rate	TracIn	0.9775	0.9785	0.9783	0.0063
clust4	hybrid	0.7508	0.7391	0.7401	0.0070
clust4	loss	0.6888	0.6691	0.6707	0.0070
clust4	TracIn	0.9645	0.9644	0.9644	0.0070
clust8	hybrid	0.7436	0.7370	0.7381	0.0093
clust8	loss	0.6720	0.6621	0.6639	0.0093
clust8	TracIn	0.9709	0.9710	0.9710	0.0093
low-sim	hybrid	0.7501	0.7237	0.7261	0.0070
low-sim	loss	0.6991	0.6653	0.6678	0.0070
low-sim	TracIn	0.9658	0.9659	0.9656	0.0070

Table 10 continued. Structural-control panel.

The decomposition is informative in both positive and negative ways. Negatively, it rules out the claim that the structure-aware stage is the universally dominant component. Positively, it shows that full SCARV is still meaningfully above bare, cluster-only, and dedup-first baselines in the submitted-

pipeline comparison. This is an overall-pipeline result; structural-increment claims require separate operator-controlled diagnostics.

F.6 Natural Redundancy and Seed-Budget Frontier

The natural-redundancy results are interpreted together with the matched and covered-only diagnostics. The submitted full-pipeline QQP comparison is positive, whereas the operator-controlled matched structural effect is small and protocol-dependent. QQP’s non-singleton coverage remains modest, so global stability can hide whether the structure stage is actually improving the covered examples. The compute frontier below is a seed-budget frontier; the separate wall-clock diagnostic distinguishes fresh-map, amortized-map, and existing-map regimes.

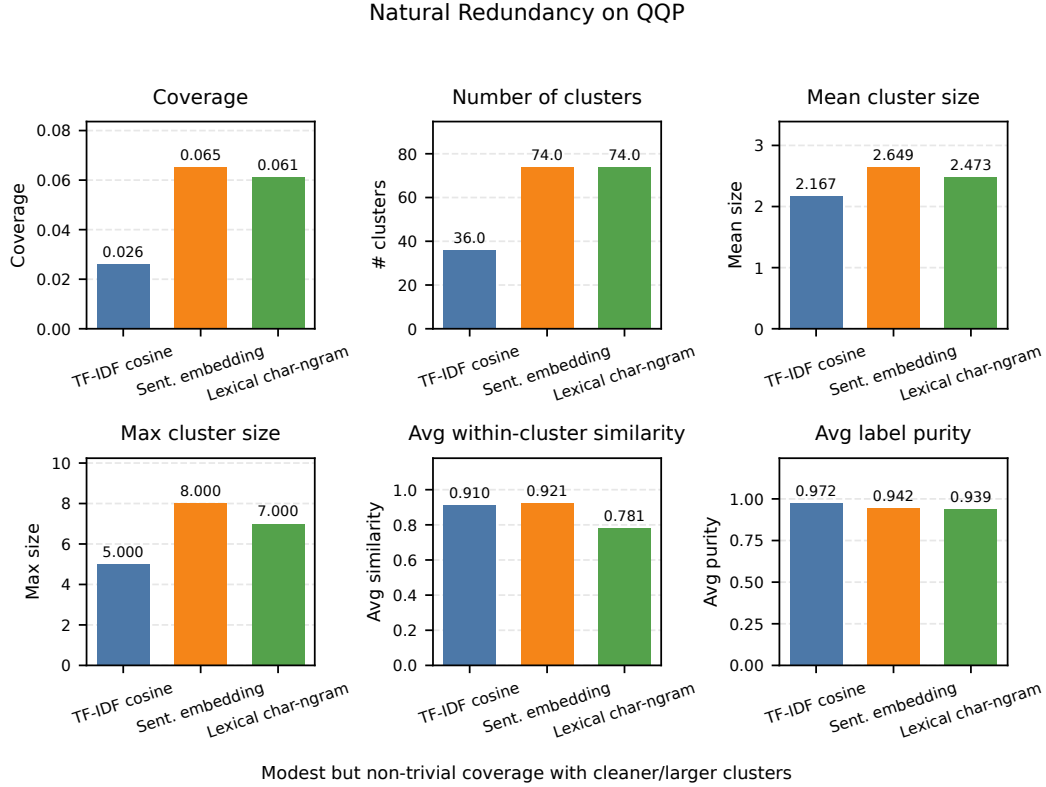


Figure 4: Natural redundancy on QQP. Coverage is modest but non-trivial, with cleaner and larger clusters than legacy pair-only settings.

Method	Cov.	#Clust.	Mean	Max	Sim. / Pur.
TF-IDF cosine	0.026	36	2.17	5	0.910 / 0.972
Sent. embedding	0.065	74	2.65	8	0.921 / 0.942
Lexical char-ng	0.061	74	2.47	7	0.781 / 0.939

Table 11: Natural redundancy on QQP. Cluster statistics for three approximate discovery procedures.

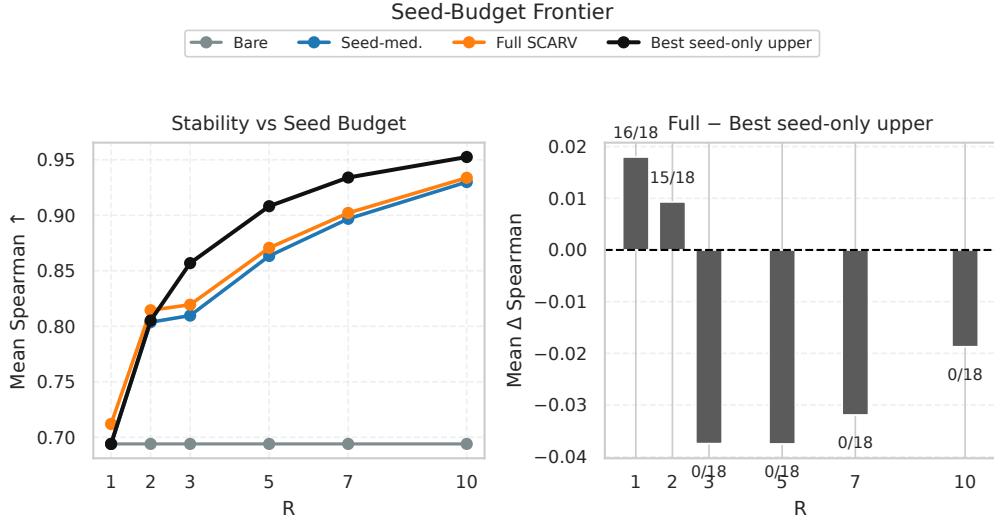


Figure 5: Seed-budget frontier. Full SCARV is competitive at small seed counts in this diagnostic, while the post-hoc best seed-only upper baseline becomes strongest for pure stability as seed count increases. This figure does not by itself charge fresh map-construction cost.

R	Full > best upper	Mean Δ_{best}	Median Δ_{best}
1	16	+0.0180	+0.0233
2	15	+0.0095	+0.0126
3	0	-0.0374	-0.0335
5	0	-0.0374	-0.0307
7	0	-0.0318	-0.0251
10	0	-0.0187	-0.0138

Table 12: Compact seed-budget frontier summary. Δ_{best} is Full SCARV Spearman minus the post-hoc strongest seed-only method under the same seed count.

The seed-budget frontier is therefore a diagnostic about the number of proxy evaluations, not a wall-clock claim. It shows where the submitted full pipeline is competitive under fixed seed-count accounting, while the wall-clock appendix separately charges construction and loading costs for the redundancy map.

F.7 End-to-End DistilBERT Fine-Tuning

To test whether the stability phenomenon is tied to shallow models, we additionally fine-tune distilbert-base-uncased end-to-end on SST-2, MRPC, QQP, and MNLI subsets. Across loss, confidence, and margin proxies, full SCARV consistently improves over bare rankings in Spearman stability, top-10 Jaccard overlap, and selected-subset reproducibility. The fine-tuning protocol and seed structure are described in Appendix E. The results are also consistent with the main mechanism: full SCARV is competitive with seed-median, while seed-mean and seed-Borda remain strong upper baselines for pure stability.

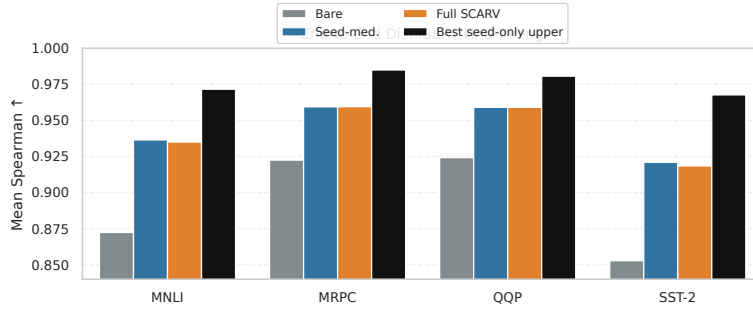


Figure 6: End-to-end DistilBERT fine-tuning summary. Full SCARV improves over bare and is competitive with seed-median; Best seed-only upper remains strongest for pure stability.

Dataset	Bare	Seed-med.	Full SCARV	Best upper
MNLI	0.872	0.937	0.935	0.970
MRPC	0.923	0.959	0.960	0.983
QQP	0.924	0.959	0.959	0.981
SST-2	0.853	0.921	0.919	0.965
Avg.	0.893	0.944	0.943	0.975

Table 13: End-to-end DistilBERT fine-tuning. Values are mean Spearman stability averaged over confidence, loss, and margin proxies.

Proxy	Method	Spearman	Top10 Jacc.	Subset overlap@0.3
Confidence	approx. dedup	0.8707	0.5922	0.6712
Confidence	bare	0.8732	0.5948	0.6759
Confidence	cluster-only	0.8705	0.5956	0.6715
Confidence	full SCARV	0.9356	0.6958	0.7662
Confidence	seed-Borda	0.9700	0.7636	0.8246
Confidence	seed-mean	0.9706	0.7674	0.8368
Confidence	seed-median	0.9371	0.6860	0.7708
Loss	approx. dedup	0.8707	0.5920	0.6712
Loss	bare	0.8732	0.5948	0.6759
Loss	cluster-only	0.8705	0.5959	0.6715
Loss	full SCARV	0.9351	0.6952	0.7644
Loss	seed-Borda	0.9700	0.7636	0.8246
Loss	seed-mean	0.9708	0.7702	0.8388
Loss	seed-median	0.9366	0.6854	0.7696
Margin	approx. dedup	0.8682	0.5827	0.6681
Margin	bare	0.8707	0.5839	0.6723
Margin	cluster-only	0.8680	0.5853	0.6690
Margin	full SCARV	0.9344	0.6837	0.7648
Margin	seed-Borda	0.9694	0.7583	0.8226
Margin	seed-mean	0.9699	0.7643	0.8365
Margin	seed-median	0.9359	0.6833	0.7686

Proxy	Method	Spearman	Top10 Jacc.	Subset overlap@0.3
Confidence	approx. dedup	0.9222	0.4575	0.7629
Confidence	bare	0.9225	0.4463	0.7489
Confidence	cluster-only	0.9213	0.4590	0.7590
Confidence	full SCARV	0.9598	0.5545	0.8159
Confidence	seed-Borda	0.9826	0.6490	0.8664
Confidence	seed-mean	0.9834	0.7120	0.8935
Confidence	seed-median	0.9597	0.5471	0.8044
Loss	approx. dedup	0.9221	0.4575	0.7628
Loss	bare	0.9225	0.4463	0.7489
Loss	cluster-only	0.9209	0.4590	0.7591
Loss	full SCARV	0.9589	0.5501	0.8139
Loss	seed-Borda	0.9826	0.6490	0.8664
Loss	seed-mean	0.9828	0.7176	0.8942
Loss	seed-median	0.9589	0.5447	0.8026
Margin	approx. dedup	0.9222	0.4575	0.7629
Margin	bare	0.9225	0.4463	0.7489
Margin	cluster-only	0.9213	0.4590	0.7590
Margin	full SCARV	0.9598	0.5545	0.8159
Margin	seed-Borda	0.9826	0.6490	0.8664
Margin	seed-mean	0.9834	0.7120	0.8935
Margin	seed-median	0.9597	0.5471	0.8044

Table 14: End-to-end DistilBERT fine-tuning on MNLI. Values are means over outer runs.

Proxy	Method	Spearman	Top10 Jacc.	Subset overlap@0.3
Confidence	approx. dedup	0.9241	0.6399	0.7982
Confidence	bare	0.9242	0.6380	0.7973
Confidence	cluster-only	0.9236	0.6398	0.7970
Confidence	full SCARV	0.9591	0.7619	0.8621
Confidence	seed-Borda	0.9807	0.8210	0.8987
Confidence	seed-mean	0.9775	0.8143	0.8951
Confidence	seed-median	0.9592	0.7607	0.8626
Loss	approx. dedup	0.9246	0.6400	0.7983
Loss	bare	0.9242	0.6380	0.7973
Loss	cluster-only	0.9241	0.6399	0.7970
Loss	full SCARV	0.9590	0.7625	0.8621
Loss	seed-Borda	0.9807	0.8210	0.8987
Loss	seed-mean	0.9763	0.8120	0.8942
Loss	seed-median	0.9589	0.7613	0.8619
Margin	approx. dedup	0.9241	0.6400	0.7982
Margin	bare	0.9242	0.6380	0.7973
Margin	cluster-only	0.9236	0.6399	0.7970
Margin	full SCARV	0.9591	0.7619	0.8621
Margin	seed-Borda	0.9807	0.8210	0.8987
Margin	seed-mean	0.9775	0.8143	0.8951
Margin	seed-median	0.9592	0.7607	0.8626

Table 16: End-to-end DistilBERT fine-tuning on QQP. Values are means over outer runs.

Table 15: End-to-end DistilBERT fine-tuning on MRPC. Values are means over outer runs.

Proxy	Method	Spearman	Top10 Jacc.	Subset overlap@0.3
Confidence	approx. dedup	0.8494	0.5937	0.7311
Confidence	bare	0.8529	0.5925	0.7356
Confidence	cluster-only	0.8483	0.5987	0.7310
Confidence	full SCARV	0.9184	0.7242	0.8201
Confidence	seed-Borda	0.9649	0.7653	0.8576
Confidence	seed-mean	0.9471	0.7840	0.8712
Confidence	seed-median	0.9209	0.7232	0.8237
Loss	approx. dedup	0.8494	0.5937	0.7311
Loss	bare	0.8529	0.5925	0.7356
Loss	cluster-only	0.8483	0.5987	0.7310
Loss	full SCARV	0.9186	0.7282	0.8200
Loss	seed-Borda	0.9649	0.7653	0.8576
Loss	seed-mean	0.9439	0.7875	0.8670
Loss	seed-median	0.9211	0.7236	0.8255
Margin	approx. dedup	0.8494	0.5937	0.7311
Margin	bare	0.8529	0.5923	0.7356
Margin	cluster-only	0.8483	0.5985	0.7310
Margin	full SCARV	0.9184	0.7242	0.8202
Margin	seed-Borda	0.9649	0.7647	0.8576
Margin	seed-mean	0.9471	0.7840	0.8712
Margin	seed-median	0.9209	0.7232	0.8237

Table 17: End-to-end DistilBERT fine-tuning on SST-2. Values are means over outer runs.

F.8 Practical Decisions, Proxy Generality, and Significance

Stability Helps Through Reproducible Decisions

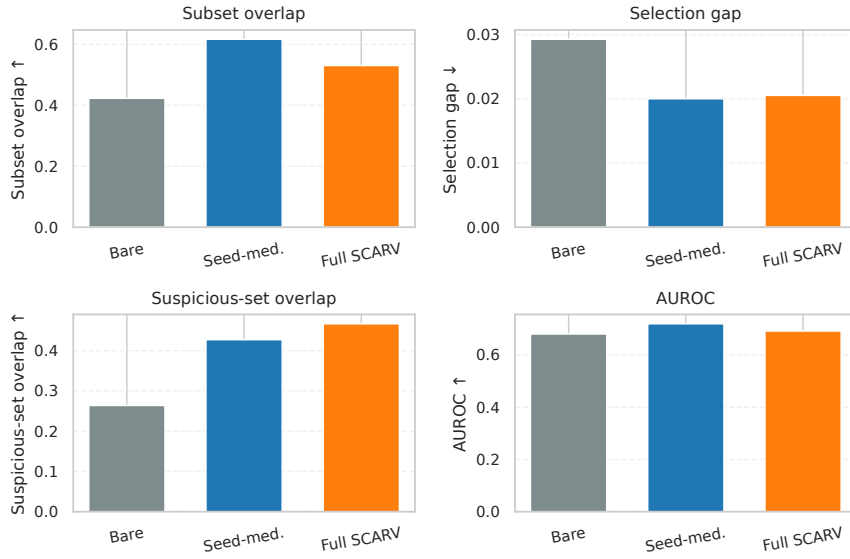


Figure 7: Reproducible decisions under stabilization. The four-panel layout separates subset overlap, selection gap, suspicious-set overlap, and AUROC to avoid dual-axis ambiguity. The results emphasize repeatability over universal utility gains.

Proxy	Noise	Bare	Seed-med.	Full SCARV
Loss	10%	0.7044	0.7442	0.7115
Loss	20%	0.6563	0.6946	0.6724
TracIn	10%	0.5863	0.5859	0.5601
TracIn	20%	0.6081	0.6117	0.5872

Table 18: Noisy-label retrieval AUROC on SST-2. Results are shown for loss and TracIn proxies at 10% and 20% noise rates.

Proxy	Bare	Seed-med.	Full SCARV
Loss	0.6040	0.8422	0.8018
Margin	0.5413	0.8028	0.7791
TracIn	0.9564	0.9769	0.9774

Table 19: Proxy generality on SST-2. Stability results across loss, TracIn, and margin proxies.

Dataset	Mean Δ	95% CI	p -value
SST-2	-0.0121	$[-0.0178, -0.0060]$	0.0004*
AG News	-0.0011	$[-0.0092, +0.0070]$	0.5609
MRPC	+0.0098	$[+0.0054, +0.0142]$	0.1250
MNLI	+0.0059	$[+0.0049, +0.0069]$	0.1250
QQP	+0.0140	$[+0.0127, +0.0154]$	0.0312*

Table 20: Overall-pipeline paired significance: Full SCARV vs. Seed-median. Mean Δ Spearman, 95% bootstrap CI, and Wilcoxon signed-rank p -value over outer configurations. This table does not isolate the structural module because preprocessing and operator choices may differ. Asterisks denote uncorrected $p < 0.05$ for this overall-pipeline comparison; Holm correction is applied separately to the fixed operator-matched structural contrasts.

The statistical results support a deliberately nuanced claim. On the audited SST-2 synthetic regime, seed-median is significantly stronger than full SCARV. On QQP natural redundancy, the sign reverses in favor of full SCARV. On MRPC and MNLI, full SCARV is slightly ahead but not significant under the current sample size. These are overall-pipeline comparisons. Operator-controlled matched diagnostics answer a separate Structure ON versus Structure OFF question and do not support a claim of universal structural improvement.

F.9 Final Wall-Clock Diagnostic

The seed-budget frontier varies the number of proxy seeds. The whole-machine-idle wall-clock diagnostic instead charges map construction/loading and post-processing costs. In this timing study, map construction is non-negligible for cheap TF-IDF/logistic proxies, and the fixed-method Pareto audit finds zero structure-aware Pareto-optimal rows in the tested pure-stability/time scope. Existing or amortized maps reduce overhead, but they do not offset the matched stability disadvantage. These results do not replace the seed-budget table; they limit compute-efficiency claims and reserve structure for group-consistent decisions.

Map	Cache state	Median (s)	IQR	Min	Max	Fresh/proxy
lexical_chargram	existing_map_load	0.124	0.002	0.121	0.171	
lexical_chargram	fresh_map	2.852	0.030	2.807	3.024	11.02
semantic_minilm	cached_embeddings_rebuild_clusters	0.923	0.008	0.921	0.929	
semantic_minilm	cached_weights_fresh_embeddings	3.991	0.038	3.975	4.045	15.42
semantic_minilm	existing_map_load	0.121	0.001	0.121	0.126	
tfidf_cosine	existing_map_load	0.123	0.004	0.121	0.230	
tfidf_cosine	fresh_map	1.242	0.006	1.235	1.248	4.80

Table 21: Wall-clock cost components from the whole-machine-idle timing study on the QQP $n = 10,000$ controlled universe. Fresh/proxy divides map-build time by one training-inclusive TF-IDF/logistic proxy seed (0.2587 s). MiniLM uses cached local weights with freshly computed embeddings, not a true cold start. These timings are separate from the seed-budget frontier.

Cache scenario	Map source(s)	Struct. rows
amortized_M10	all 3 maps	0/3
amortized_M2	all 3 maps	0/3
amortized_M5	all 3 maps	0/3
cached_embeddings	semantic_minilm	0/1
existing_map	all 3 maps	0/3
fresh_map	all 3 maps	0/3

Table 22: Fixed-method Pareto audit. Four fixed seed-only rows form the Pareto set in each comparison; no structure-aware row is Pareto-optimal for pure stability per unit time in the tested scope.

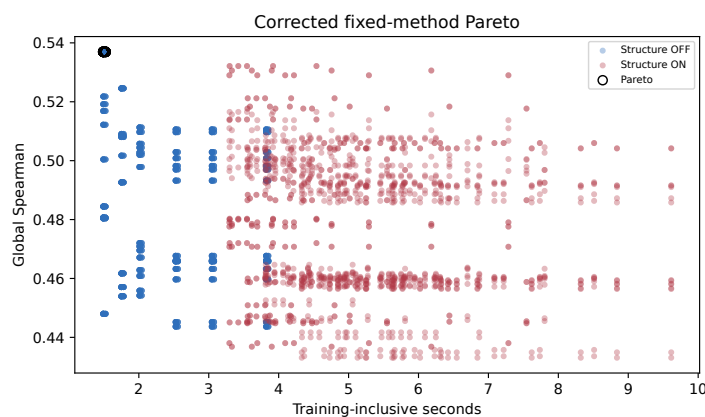


Figure 8: Final wall-clock Pareto diagnostic. The seed-budget and wall-clock analyses answer different questions; the wall-clock frontier includes map costs and shows no structure-aware Pareto-optimal rows under the tested pure-stability/time criterion.

G Additional Related Discussion

G.1 Attribution beyond the Main-Text Baselines

The main paper focuses on loss-based and TracIn-style proxies, but the broader attribution literature now includes scalable or generalized formulations such as DataInf, approximate-unrolling-based attribution, multiply-robust causal change attribution, and generalized influence-style methods for non-decomposable losses (Kwon et al., 2024; Bae et al., 2024; Quintas-Martinez et al., 2024; Deng et al., 2025). These methods make example-level attribution increasingly practical in larger or more structured settings. Their relevance to SCARV is complementary rather than competitive: they provide richer proxy mechanisms, while SCARV addresses the separate question of how those proxy-induced rankings can be aggregated under repeated or near-repeated structure.

Semivalue and nearest-neighbor valuation methods have also developed task-specific and scalable variants, including efficient task-specific valuation for nearest-neighbor models and subsequent weighted or thresholded neighborhood-based extensions (Jia et al., 2019a; Wang et al., 2024a, 2023). Class-wise value decompositions such as CS-Shapley and Shapley-style reasoning in conformal prediction further show that “value” itself is not monolithic (Schoch et al., 2022; Jaramillo and Smirnov, 2021). Our paper is orthogonal to that axis: SCARV does not redefine the value notion, but stabilizes the ranking induced by a chosen proxy when redundancy and stochastic training make pointwise scores unreliable.

G.2 Diagnostics, Memorization, and Repeated Structure

A further neighboring literature studies ranking-like signals for dataset diagnostics, memorization, and instability. In addition to example forgetting and dataset cartography, zero-shot data maps and recent memorization-tracking work show that example-level behavior can reflect ambiguity, noise, recency, and repeated exposure in ways that are consequential but not reducible to a single scalar notion of value (Basile et al., 2023; Jagielski et al., 2023; Leybzon and Kervadec, 2024). At the corpus level, repeated content also interacts with privacy and copyright concerns in large language models (Meeus et al., 2024). This broader view supports the motivation for our paper: under redundancy, reproducibility of a proxy-induced ranking is itself a meaningful property, even when the downstream use of the ranking varies across attribution, debugging, pruning, or curation.

G.3 Proxy–Objective Mismatch

One of the clearest empirical patterns in the paper is that influence-oriented and utility-oriented proxies can disagree sharply. In our diagnostic checks, TracIn-style scores are generally more positively aligned with LOO-style valuation than loss or margin scores, but the effect of stabilization on LOO alignment is mixed. TracIn-style scores are also often weak direct selectors under aggressive pruning. Loss-based scores exhibit the opposite pattern more often: they can be weaker valuation signals while remaining competitive or stronger for subset utility under some budgets. This mismatch is not a failure of the framework. It reflects a genuine difference between what different proxy families are designed to measure.

From this perspective, SCARV does not attempt to collapse all ranking signals into one universal ordering. Rather, it stabilizes the behavior of a chosen proxy under redundancy and stochastic training. If the underlying proxy is attribution-oriented, SCARV makes that attribution-oriented ranking more reproducible. If the underlying proxy is utility-oriented, SCARV makes that utility-oriented ranking more reproducible. This reinforces the paper’s main separation between valuation alignment and ranking stability.

G.4 Mechanism, Scope, and Future Directions

The appendix-level decomposition results materially sharpen how the method is interpreted. In the default synthetic regime, the strongest methods are seed-only aggregators such as seed-mean, seed-Borda, and seed-median. This makes the contribution narrower than a claim of universal dominance, but stronger as an empirical statement about what actually causes ranking instability under redundancy. The main source of gain is repeated robust aggregation across seeds. The submitted full-pipeline QQP comparison is positive, whereas the operator-controlled matched structural effect is small and protocol-dependent;

SST-2 near-duplicate settings can be negative. A small seed count alone is not a deployment rule, because fresh map construction can dominate the total cost. The structure-quality diagnostics add a further caveat: high stability under a structural regularizer does not by itself certify that the clusters are semantically meaningful.

Several limitations remain. First, the strongest support in the paper concerns ranking stability and decision reproducibility, not universal gains in downstream selection utility. Second, the natural-redundancy evidence still stops well short of large-scale end-to-end pretraining or instruction-tuning workflows. Third, the current framework assumes that some usable redundancy structure can be injected, mined, or approximately recovered; in many realistic settings, redundancy is softer, noisier, or hierarchical. Future work should move from hard clusters to soft or uncertain redundancy maps, study stability-oriented aggregation in larger fine-tuning pipelines, and design structure-aware aggregators that adapt more explicitly to cluster size, cluster confidence, or proxy family.