

Preference-Graph Repair and Retrieval Effectiveness: An Empirical Decomposition

Soroush Vahidi

sv96@njit.edu

New Jersey Institute of Technology

Research Article

Keywords: preference graphs, feedback arc set, graph repair, rank aggregation, information retrieval, retrieval evaluation

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11518675/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Preference-Graph Repair and Retrieval Effectiveness: An Empirical Decomposition

Soroush Vahidi^{1*}

^{1*}Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA

ORCID: 0000-0003-1934-6282.

Corresponding author(s). E-mail(s): sv96@njit.edu;

Abstract

Preference-graph repair removes contradictory comparisons, but structural consistency and retrieval effectiveness are different objectives. Their separation is well understood in principle; less clear is where the effect of repair stops between the graph, the extracted ranking and its evaluated relevance profile. We investigate this question through matched unrepaired, greedy and exact-repair comparisons that hold candidates, preference evidence, extraction and evaluation fixed. The central analysis decomposes responses on 213 selected queries across four datasets. Greedy and exact repair change 180 and 160 rankings, respectively, yet measured utility remains unchanged in 79.4% and 84.4% of those changed rankings. Two mechanisms account for these zeros: pools without a judged-positive candidate and positive-containing pools whose evaluated top-ten relevance profile is preserved. The latter occurs in 74 cases per repair branch and across every dataset and provider-availability stratum. Thus missing judged opportunity alone does not explain the observed inactivity. Query-block inference supplies no Holm-adjusted rejection in the 35 exact-versus-greedy comparisons, while a complete small-instance diagnostic shows that equal-cost repairs can yield different utility. These findings distinguish an intervention that changes rankings from one that changes measured relevance. They extend earlier construction-dependent evidence through conditional frequencies and relevance-profile attribution, rather than a new repair algorithm. Nonrandom selection, incomplete judgments, sparse informative support and an unreliable compact-model comparison instrument limit generalization; the results establish neither equivalence nor a universal absence of benefit.

Keywords: preference graphs, feedback arc set, graph repair, rank aggregation, information retrieval, retrieval evaluation

1 Introduction

Pairwise comparisons offer a flexible way to rank documents using evidence from several retrieval models or language-model judges. A comparison favoring document u over v becomes a directed edge $u \rightarrow v$. Aggregated comparisons can form cycles, making it impossible to satisfy every preference in one order. Removing a feedback arc

set resolves that structural contradiction. It does not, however, determine whether a relevant document moves to a better position. The graph objective concerns agreement with the comparisons; retrieval effectiveness concerns relevance to the query.

This distinction creates an evaluation problem. A small average retrieval effect can arise because

the graph requires no repair, because deletion leaves the extracted order unchanged, because a changed order preserves the relevance grades seen by the metric, or because gains and harms offset one another. These outcomes have different implications. The first offers little opportunity to assess deletion, the second depends on the extractor’s response, and the third concerns which ranking changes the evaluation can distinguish. A mean effect alone does not locate that separation.

Existing research establishes effective pairwise and graph-based ranking methods, and our public preprint already reports construction-dependent cyclicity and conditional retrieval effects [30]. The unresolved empirical question is how often structural interventions reach the ranking and measured-relevance stages, and how the resulting zeros are distributed across the observed query populations. Equal evaluated grade sequences necessarily give equal nDCG under fixed normalization; that identity does not tell us whether such cases are common, whether they contain judged-positive candidates, or whether a few datasets dominate them.

We ask: *when preference-graph repair improves structural consistency, under what observed conditions does it change the extracted ranking and measured retrieval effectiveness?* We compare unrepaired, greedy and exact-repair branches while holding candidates, comparison evidence, extraction, tie handling and evaluation fixed. Four score-derived benchmarks provide controlled construction contrasts; a selected multi-provider cohort supports complete ranking and relevance-profile attribution. A six-query diagnostic examines all minimum-cost repairs, and a separate compact-model cohort tests the limits of the elicited evidence. These cohorts have different selection and measurement conditions and are not interchangeable replications.

1.1 Scientific contributions

The principal contribution is an empirical decomposition of the response to graph repair:

1. **From graph modification to ranking response.** We distinguish acyclic inputs from cyclic graphs whose repair leaves the fixed extractor’s ranking unchanged (Section 5).

2. **From ranking response to measured relevance.** On 213 selected queries, complete relevance-profile attribution separates changed rankings without judged-positive candidates from changed positive-pool rankings with fixed evaluated profiles. Dataset and provider-availability composition shows where these mechanisms occur, with both gains and harms retained (Section 6).
3. **From structural optimality to utility ambiguity.** Complete optimum sets for six graphs quantify the utility variation left unidentified by minimum deletion cost under fixed extraction (Section 8).

These are analyses of reused observations, not new independent datasets. Graph-repair algorithms and ranking-MWFAS theory are established methods. Corrected query-block inference supports the interpretation of the results; computational verification checks their arithmetic and finite completeness. Neither correction nor verification is presented as algorithmic novelty. The scientific increment is the observed composition of the intervention’s responses, which distinguishes missing evaluated opportunity from active ranking changes that measured relevance does not distinguish.

2 Related Work

2.1 Preference aggregation and structural objectives

Classical rank aggregation reconciles inconsistent preferences through minimum-disagreement ordering and related objectives [2, 12]. Feedback arc set optimization and linear-ordering formulations provide established methods for removing cycles [3, 14]. The author’s earlier training-free ranking pipeline develops the ranking–feedback-arc relationship and combines pruning, cycle-safe add-back and refinement [31]. Those methods, theory and ranking-recovery results are prior work. Related structural graph-benchmark research concerns graph identity and comparisons among existing optimizers; its benchmark outcomes are not new retrieval observations in the present study. Our question instead concerns deletion’s effect against external relevance judgments.

Alternative aggregation methods infer an order without solving a deletion problem. Bradley–Terry models estimate latent strengths [5, 17]; Rank Centrality uses a comparison Markov chain [24]; combinatorial Hodge theory separates ranking and cyclic components of pairwise flow [19]. SpringRank estimates a hierarchy through a weighted quadratic energy [11], while reciprocal rank fusion combines source ranks without constructing a preference graph [10]. These methods motivate complementary controls, but changing the aggregator as well as repairing the graph does not isolate deletion’s contribution. Nor do their assumptions make arbitrary edge weights calibrated correctness probabilities.

2.2 Language-model ranking and retrieval evaluation

Pairwise ranking prompting uses language-model comparisons to rerank documents [25]. PRP-Graph incorporates target-label probabilities into graph construction and aggregation [21]; a winner-only cache cannot recover that probability information. Listwise and setwise prompting elicit different evidence by ordering or selecting among several documents jointly [27, 36]. Our adapted PRP-Graph comparison therefore provides pipeline context rather than an exact reproduction of published effectiveness or an isolated test of edge deletion.

LLM-RankFusion addresses presentation-order and transitive inconsistency through prompting, calibration and aggregation [35]. Preference Graph Ensemble and Denoise uses multiple evaluators to obtain acyclic judgments for model evaluation, response selection and training-data selection [16]. TourRank combines tournament-style grouping with a points-based ensemble for document ranking [8]. Xu et al. show that non-transitive judgments can make model rankings sensitive to the baseline, and study tournament acquisition with Bradley–Terry aggregation [33]. These studies establish the relevance of consistency and its mitigation under their own tasks and acquisition procedures. They do not determine whether minimum-weight deletion improves document relevance when evidence and extraction are held fixed.

The distinction between intrinsic consistency and task-dependent quality is established in data-quality research [32]. Likewise, nDCG measures relevance at ranking positions rather than agreement with comparison edges [18]. Buckley and Voorhees examine retrieval evaluation with incomplete relevance information [6], motivating the distinction between an evaluation zero and semantic irrelevance. Their work does not validate our selected cohorts or recover their unobserved relevance. We retain the original metric definitions rather than select a different metric after observing zeros.

The public retrieval preprint already distinguishes construction, repair and extraction [30]. The present analysis adds complete selected-cohort attribution, its dataset and provider-availability composition, and the finite all-optimum diagnostic. Online Resource 1 identifies reused records and construction/normalization differences; the legacy-to-current execution bridge remains incomplete. Corrected paired inference respects independent query units and multiplicity [7, 29]. A failure to reject still does not establish equivalence [20]. These inferential qualifications support the empirical contribution without turning reused evidence into independent replication.

3 Problem Formulation

3.1 Fixed-evidence repair contrast

For query q , let D_q be a fixed candidate set and $G_q = (D_q, E_q, w_q)$ its directed preference graph. Edge $u \rightarrow v$ favors placing u ahead of v ; its nonnegative weight represents recorded evidence strength, not ground-truth relevance. A feedback arc set F makes $G_q - F$ acyclic. Minimum-weight repair solves

$$F_q^* \in \arg \min_{F \subseteq E_q} \sum_{e \in F} w_q(e) \quad (1)$$

subject to $G_q - F$ being acyclic.

The empty deletion is optimal for an acyclic positive-cost graph. Acyclicity need not imply a unique total order or relevance-optimal preferences. A structural optimum need not identify a unique deletion set either.

Let $\mathcal{R}$ denote a fixed deterministic ranking extractor and U_q the retrieval metric with fixed

relevance judgments and normalization. For a returned repair F , the observed contrast is

$$\Delta_q(F) = U_q(\mathcal{R}(G_q - F)) - U_q(\mathcal{R}(G_q)). \quad (2)$$

Greedy, exact and unrepaired branches share these inputs. An exact certificate concerns Eq. (1), not the maximum of Eq. (2). It also does not characterize every equally optimal repair or all rankings obtainable from a repaired graph.

3.2 Rankings, relevance and opportunity stages

The principal extractor for cached judgments is unweighted Copeland: outgoing degree minus incoming degree, sorted in descending order with ascending document-ID ties. The score-derived studies also hold other extractors fixed within their repair contrasts. Some extractors can be applied to cyclic graphs; topological extraction requires a directed acyclic graph (DAG) and appears only in suitable repaired-graph comparisons. Acyclicity alone does not guarantee that a score-based extracted order respects every retained edge.

Let $\pi_q = (d_1, \dots, d_{P_q})$ rank the fixed candidates, and let $r_q(d)$ be the recorded relevance grade, with an unjudged candidate assigned evaluation grade zero. Define $k_q = \min(k, P_q)$. Normalized discounted cumulative gain is

$$\text{nDCG}_q @k = \frac{\sum_{i=1}^{k_q} (2^{r_q(d_i)} - 1) / \log_2(i+1)}{\text{IDCG}_q @k}. \quad (3)$$

For an ideal-reference set B_q , let $r_{q,B_q,i}^\downarrow$ denote its grades sorted in descending order. Ideal discounted cumulative gain (IDCG) is

$$\text{IDCG}_q @k = \sum_{i=1}^{\min(k, |B_q|)} \frac{2^{r_{q,B_q,i}^\downarrow} - 1}{\log_2(i+1)}. \quad (4)$$

The implementation returns zero when IDCG is nonpositive. Candidate-conditioned IDCG uses $B_q = D_q$; full-query IDCG uses the complete recorded query qrels, including judged documents outside the candidate pool. Thus a short candidate ranking does not truncate the full-query ideal to the pool size. Evaluation zeros for unjudged documents do not establish irrelevance.

For binary relevance $z_{qi} = \mathbf{1}\{r_q(d_i) > 0\}$ and $R_q = \sum_{i=1}^{P_q} z_{qi}$, the historical exact-versus-greedy endpoints are

$$\begin{aligned} \text{AP}_q &= \frac{1}{R_q} \sum_{i=1}^{P_q} z_{qi} \frac{\sum_{j=1}^i z_{qj}}{i}, \\ \text{RR}_q &= \frac{1}{\min\{i : z_{qi} = 1\}}. \end{aligned} \quad (5)$$

Both are defined as zero when $R_q = 0$. Average precision (AP) and reciprocal rank (RR) use the complete recorded candidate ranking, not the full corpus. Within a single cell, mean average precision (MAP) and mean reciprocal rank (MRR) are the arithmetic means of AP and RR over eligible queries; the historical MAP label does not change AP’s candidate-positive denominator. Within a fixed comparison, methods share candidates, labels, cutoff and IDCG. Across families, changing IDCG can rescale query effects and their aggregate weighting; the original contrasts are retained.

We distinguish graph modification, full-ranking response and evaluated grade-position response. If a changed ranking has the same first- k grade sequence and fixed IDCG, Eq. (3) is unchanged. This is a sufficient explanation for a recorded zero; different grade sequences can also yield equal metric values. Top- k membership is a separate diagnostic: reordering a fixed set can change nDCG, while exchanging equal-gain documents can preserve it. We therefore do not use membership change as a necessary gate for retrieval activity.

4 Experimental Design

4.1 Datasets and query populations

Table 1 separates the evidence populations. We use relevance judgments (qrels) as the external evaluation reference. The canonical score-derived study uses SciDocs, FiQA, HotpotQA and BRIGHT [9, 23, 26, 34]. These benchmarks provide heterogeneous relevance supervision, including citation- and question-answering-derived labels. They are not uniformly direct human pairwise gold standards. Newer FiQA and SciDocs cohorts use BEIR-hosted snapshots

[28]; the historical HotpotQA retrieval conversion remains incompletely pinned.

Canonical counts are 120 SciDocs, 120 FiQA, 52 HotpotQA and 50 BRIGHT queries. Eighteen of 70 stored HotpotQA queries fail the canonical relevant-candidate eligibility condition. The six-query primary cache is disjoint from the selected 213-query cache; the latter comprises 77 SciDocs, 76 FiQA, 42 BRIGHT and 18 HotpotQA identities. It is selected and heterogeneous rather than a random prospective replication. Seventy-nine selected-cache pools have no judged-positive candidate; they remain in principal means.

The three newer 20-query cohorts are mutually disjoint. BGE reuses the PRP and prospective pools; graph-construction variants and prompt development reuse the prospective queries. The prospective collection contains 5,880 responses, but only 20 query units. Online Resource 1 preserves cohort selection, snapshot and reuse details.

4.2 Construction, repair and extraction contracts

The canonical graph uses stored BM25 lexical-retrieval scores, term frequency-inverse document frequency (TF-IDF) scores and MiniLM sentence-encoder scores as a controlled multi-source substrate. Per-query, per-ranker min-max normalization precedes margin votes. Retention-matched thresholds depend on stored score distributions, not relevance labels, but are calibrated on the evaluated query list and are therefore transductive. The permissive `ms1` rule retains ranker-supported directions; `ms2` requires support from at least two of three rankers; `ms1_drop_mutual` removes contested pairs during construction. This last rule differs from deleting reciprocal edges from an already constructed graph. Candidates come from the fixed reciprocal-rank-fusion pool policy; full thresholds and alternative policies are in Online Resource 1.

For cached judgments, strict parsing accepts complete A/B choices, maps reversed presentations back to document identity, and retains a provider edge only when both valid orientations agree. AB denotes the original document presentation and BA its reversal. Invalidity, abstention and reversal conflict never become a default winner. Individual-provider edges have unit weight; the primary aggregate sums contributions of one

quarter per provider, preserving opposing directions. All candidates, including isolates, remain nodes. Prospective Qwen2.5-7B-Instruct (7 billion parameters) and SmolLM3-3B (3 billion parameters) judgments use a separately frozen JSON protocol; strict construction similarly requires agreement of valid directional responses. Confidence labels are not calibrated probabilities. Alternative constructions are exploratory repeated views, not selected replacements for the strict protocol.

Greedy repair repeatedly deletes a minimum-weight edge from a deterministic depth-first-search cycle until acyclic. It is not the Eades-Lin-Smyth ordering heuristic [13]. Exact repair uses a linear-ordering mixed-integer program solved by SCIP through PySCIPOpt [1, 22], requiring proven optimality with zero reported gap. Pool, evidence, extractor, ties and cutoff remain fixed within a contrast, as shown in Fig. 1. The complete six-query diagnostic enumerates all distinct optimal deletion sets with independent completeness checks; it is conditional on six eight-node graphs, not a new optimizer.

External comparisons include original-graph SpringRank, direct Bradley-Terry-style aggregation, graph-free fusion, adapted PRP-Graph and a trained BGE cross-encoder. BGE follows the official raw-relevance-logit procedure on matched documents [4]; it is not budget- or training-matched to a generative comparator. The adapted PRP pilot retains the published implementation’s Swiss comparisons and personalized PageRank, but changes local pooling and runtime orchestration. It is not an exact reproduction of the published retrieval experiment. Missing or invalid listwise outputs remain missing controls. These distinctions prevent an acquisition or extractor change from being attributed to edge deletion.

4.3 Evaluation and query-level inference

The greedy-versus-unrepaired 20/60/110 and exact-versus-unrepaired 36/56 families use candidate-conditioned nDCG. Canonical cut-offs equal pool size: 20 for SciDocs, FiQA and BRIGHT, and 10 for HotpotQA. The larger-pool study retains its recorded (P, k) cells, including $P = k$ reference cells. The historical exact-versus-greedy 35/399 families instead use full-query IDCG at eligible cutoffs 5, 10 and 20; their

Table 1 Evidence cohorts and comparison contracts. Counts are unique query identities within each row; rows are not an additive independent sample. C/F denote candidate-conditioned/full-query IDCG.

Cohort	Queries	Pool / IDCG	Selection and role
Score-derived	342	20 or 10; larger 50 or 35 / C, F	Stored multi-ranker scores; 1,025 repeated query–regime exact instances; C for repair–unrepaired, F for pooled exact–greedy; details in Section 4.3
Cached primary	6	8 / F	Qrel-enriched four-provider diagnostic; complete optimum envelope
Selected cache	213	15 / F	Retrospective provider-availability strata: 110 two-provider and 103 four-provider aggregates
PRP-Graph pilot	20	20 / F	10 FiQA, 10 SciDocs; adapted external acquisition/aggregation comparison
Prospective	20	12 / F	7 FiQA, 7 SciDocs, 6 BRIGHT; two compact models, strict repair contrasts
Candidate diagnostic	20	12 per arm / F	Further queries; BM25 versus BGE selection from BM25 top 100

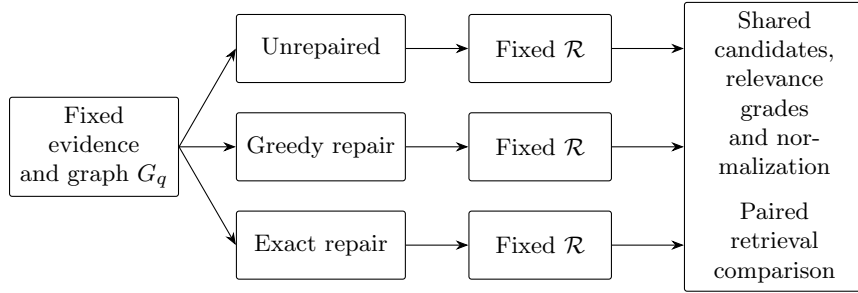


Fig. 1 Matched repair comparison. Each branch uses the same deterministic extractor $\mathcal{R}$, candidate set, tie handling, relevance grades, evaluation cutoff and normalization. Only graph repair changes; alternative aggregation methods are separate controls

MAP and MRR use complete candidate rankings. Cached/prospective LLM, SpringRank, PRP-Graph, BGE and candidate-coverage results use full-query nDCG@10. These families’ absolute means are not directly interchangeable. Online Resource 1 provides the endpoint/family map.

The corrected 35-test family crosses seven fixed extractors—balance, balance hybrid, Copeland, Copeland hybrid, damped Markov, topological and priority-topological—with MAP, MRR and the three nDCG cutoffs. The independent unit is the dataset–query identity, conditional on the benchmark sampling assumptions. Within a repair-versus-unrepaired cell, inference uses one paired query effect. In the separate pooled exact-versus-greedy diagnostic, repeated construction

regimes form one dependent query block. Ordinary single-cell means weight eligible queries equally. The pooled diagnostic retains its original regime-row-weighted effect rather than silently substituting an equal-query average. For query sums S_q and eligible regime counts m_q , its statistic is

$$T = \frac{\sum_q S_q}{\sum_q m_q}, \quad S_q = \sum_{r \text{ eligible}} (U_{qr}^{\text{exact}} - U_{qr}^{\text{greedy}}). \quad (6)$$

There are 342 query units for 28 comparisons and 290 for seven nDCG@20 comparisons, which exclude the original ineligible HotpotQA cutoff. Their 1,025 and 869 eligible regime rows, respectively, are not independent observations. The

historically omitted empty BRIGHT/ms2 record remains excluded, without zero imputation.

The query-block analysis uses 10,000 Monte Carlo whole-query sign assignments, seed 17, add-one two-sided p-values, and Holm correction over the unchanged 35-test family [15]. Bootstrap sensitivity uses 10,000 within-dataset query-cluster draws, seed 13, with percentile and basic 95% intervals. This corrected family does not use exact enumeration or bias-corrected and accelerated (BCa) intervals. All original effect estimates are preserved; historical row-based p-values are superseded.

Sign-flip inference assumes independent query units and exchangeability of the entire query’s exact/greedy outcome vector under label swap. Mean zero alone does not justify that null, and the deterministic methods were not randomly assigned. Bootstrap intervals are approximate and pointwise, especially with sparse, skewed differences. Separate canonical, larger-pool, 399-cell and exploratory families retain their own procedures and error scope; their juxtaposition provides no global error guarantee. In model-averaged prospective contrasts, fixed-model effects are averaged within query before inference; model-specific comparisons retain one effect per query. Non-significance is not equivalence, and a pointwise interval does not override a Holm decision. Online Resource 1 provides the complete comparison families and inference details.

4.4 Computational methods and AI assistance

The analyses use fixed comparison evidence and deterministic extraction within each repair contrast. Online Resource 1 specifies runtime versions, thresholds, model settings and numerical verification procedures. Historical hosted-model aliases and incomplete scorer reconstruction constrain computational reproducibility; data and code access are described in the declarations.

Anthropic Claude and OpenAI ChatGPT/Codex assisted with software development, analysis utilities, figures and drafting. The author retains responsibility for verifying the text, code, analyses and citations against the underlying evidence. These tools are not authors.

5 Structural and Ranking Opportunities

5.1 Construction-dependent cyclicity

Table 2 distinguishes graph workload from query support. Of 1,025 canonical graph records, 646 are acyclic and 379 cyclic. The records reuse 342 queries across three regimes; 316 identities are cyclic under at least one regime. Conservative ms2 graphs are acyclic throughout the observed cohort. This is an empirical property of these constructed inputs, not a theorem that majority-supported preferences are always acyclic.

In the active ms1 regime, cyclic-query rates are 99.2% for SciDocs, 98.3% for FiQA, 63.5% for HotpotQA and 92.0% for BRIGHT. Construction-time mutual dropping reduces them to 11.7%, 30.8%, 1.9% and 22.0%. Post-hoc reciprocal deletion from the SciDocs ms1 graph instead gives 10.8%; the different operation and threshold contract matter. Source-scale sensitivity is also substantial: conditional BM25 edge-weight share is 0.988 under raw margins and 0.512 after normalization. These measurements establish construction dependence, not the semantic accuracy of the resulting votes.

The genuine-judgment cohorts demonstrate a separate aggregation effect. All 24 individual-provider graphs in the six-query corrected pilot are acyclic, but all six four-provider aggregate graphs are cyclic. In the selected cache, the two disjoint provider-availability strata contain 89/110 and 102/103 cyclic aggregate graphs. Absence of structural opportunity therefore cannot explain the selected-cache outcomes generally.

The prospective strict protocol has no cyclic-positive case: only one of 40 model/query graphs is cyclic, and its candidate pool contains no judged-positive document. All 14 positive-containing model/query graphs are acyclic. The alternative constructions create more cycles but reuse the same 20 queries. Qwen cyclic counts for strict, AB-only, BA-only, symmetric-both-valid and symmetric-available rules are 0/8/8/5/5; Smol counts are 1/13/12/1/5. More retained edges change the intervention opportunity, without establishing more reliable preference evidence.

Table 2 Structural and ranking opportunity. Greedy/exact outcomes share queries. A dash denotes an unavailable complete ranking decomposition, not zero activity

Cohort	Query IDs	Graph records	Cyclic	Cyclic, same rank	G/E
Canonical	342	1,025	379		—
Selected cache	213	213	191		11 / 31
Strict prospective	20	40	1		0 / 0

Canonical records repeat construction regimes. Prospective records repeat two models; its sole cyclic graph has no judged-positive candidate. Both repairs modify every reported cyclic input graph; the extracted ranking can remain unchanged.

5.2 Repair activity and ranking response

Both repairs modify all 379 cyclic canonical records and all 191 selected-cache cyclic aggregates. Exact repair has a different deletion set from greedy in 87.9% of canonical cyclic records. It strictly lowers removed weight in 333 of those 379 records; lower dataset-mean removal costs are shown in Fig. 2. Mean reductions relative to greedy are 25.1%, 29.2%, 13.4% and 22.3% for SciDocs, FiQA, HotpotQA and BRIGHT. These are structural quantities; retrieval effect sizes are evaluated separately.

Graph activity does not imply a full-ranking response. Among the 191 cyclic selected-cache graphs, greedy leaves 11 Copeland rankings unchanged and exact leaves 31 unchanged. Including 22 acyclic graphs, unchanged-ranking totals are 33 and 53 of 213. This difference shows that choosing a structurally cheaper deletion can change the amount of extracted-order activity; it is not evidence that activity itself is beneficial. In the six-query aggregate diagnostic, each repair changes two rankings and leaves four unchanged.

The prospective construction workload contains 400 greedy/exact branch records across 200 model/query/construction graphs. Existing diagnostics identify 116 graph changes, 106 ranking changes and nine nonzero utility records, covering six query identities. These workload counts are descriptive repeated measurements of 20 queries. Complete historical permutations are unavailable for the 1,025 historical exact canonical scalar records, so a corresponding full canonical rank-response decomposition is not reconstructed. The selected-cache and prospective decompositions are

therefore reported with their own explicit support rather than extrapolated to all canonical graphs.

6 Relevance-Profile Attribution and Its Composition

Table 3 decomposes the responses on all 213 selected query identities using complete Copeland rankings and fixed relevance judgments. Greedy repair changes 180 rankings, of which 37 change measured utility; the remaining 143 preserve it. Exact repair changes 160 rankings, with 25 measured utility changes and 135 unchanged values. Thus measured retrieval utility remains unchanged in 79.4% (143/180) of changed greedy rankings and 84.4% (135/160) of changed exact rankings. These are descriptive frequencies in the selected cohort, not estimates of population prevalence or comparisons between independent samples.

The changed-ranking zeros have two distinct compositions. Greedy’s 143 cases comprise 69 pools without a judged-positive candidate and 74 positive-containing pools with an unchanged evaluated top-ten grade sequence. Exact’s 135 cases split into 61 and 74. No positive-pool changed-ranking zero has a different evaluated top-ten profile in this cohort. Missing judged positives therefore account for only part of the observed inactivity: fixed profiles also account for zeros when a positive candidate is available. The two branches share queries, and their 74-case subsets must not be added as independent observations.

Within the 74 positive-pool equal-profile cases, all moved documents are unjudged in 55 greedy and 58 exact records. Positive documents move in

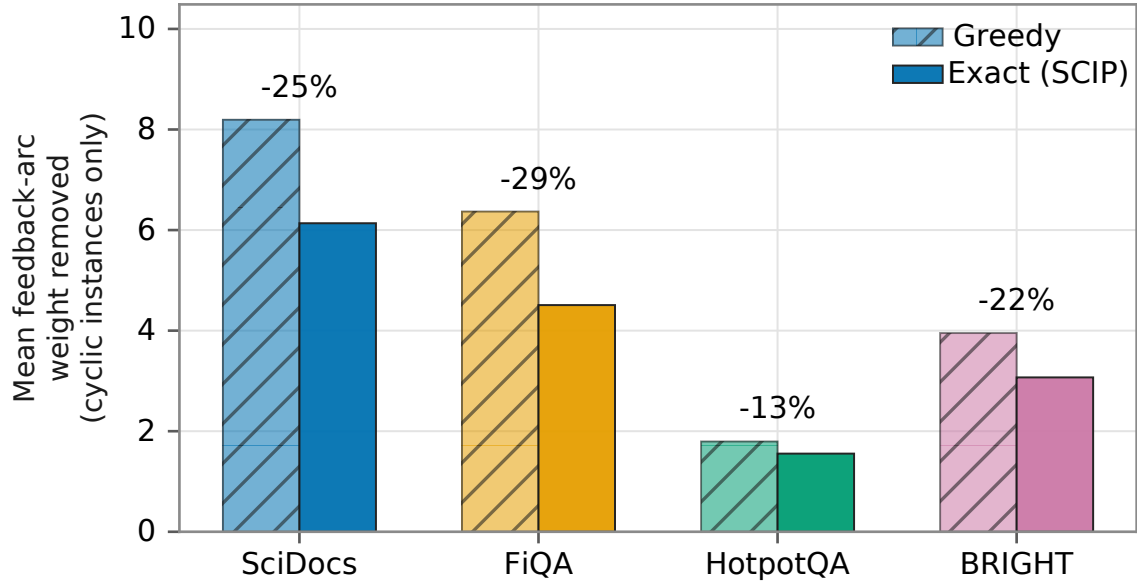


Fig. 2 Mean deletion weight on 379 cyclic canonical query–regime records, pooled over ms1, ms2 and construction-time mutual dropping. Hatched/solid bars denote greedy/exact repair. Percentage labels show the original rounded reduction from the greedy dataset mean; individual records can tie. These are structural costs, not retrieval effects

Table 3 Relevance-profile attribution on 213 selected query identities. Mutually exclusive rows sum to 213 in each branch. The two columns share queries; the 74-case subsets are not 148 independent observations

Observed outcome	Greedy	Exact
Full ranking unchanged	33	53
Ranking changed; no judged positive in pool	69	61
Ranking changed; positive pool; same evaluated top-ten grade sequence and zero metric change	74	74
Ranking and recorded metric changed	37	25

Metric-active cases contain 15 gains/22 harms for greedy and 16 gains/9 harms for exact. Unchanged rankings include 22 acyclic graphs and 11/31 cyclic graph-changing cases.

six and five records, respectively, without changing the evaluated grade sequence. Movement outside the prefix or between equal-grade positions can preserve that sequence; those subtypes were not separately classified. The attribution explains the realized metric under its qrel contract, not the semantic reason for the comparisons or the best utility obtainable from a different intervention.

Top-ten membership changes in 56 greedy and 38 exact cases, yet 45 and 32 still have zero measured effect. Conversely, 26 greedy and

19 exact nonzero effects occur without membership changes. In larger canonical pools, the mean membership-change rate is 10.6% over dependent query–regime–cell rows. Membership activity therefore supplies neither a necessary nor a sufficient indicator of relevance gain.

The four symmetric-available cyclic-positive prospective cells illustrate the same measured explanation on only three query identities. Both repairs change full rankings, but every moved document is unjudged and the top-ten grade

sequence remains fixed. These selected solutions are descriptive cases, not evidence that all equally optimal repairs or all reliable comparison settings are utility-inert. Directional constructions provide contrary cases: nine nonzero records include harms and one exact improvement, six within an unchanged top-ten set. All eight model-averaged exploratory construction contrasts retain Holm $p=1$.

6.1 Composition of observed opportunity losses

Table 4 describes the same query population by dataset and provider availability. Of the 79 zero-positive pools, 36 are SciDocs and 38 FiQA: these two datasets contribute 74/79, compared with four BRIGHT and one HotpotQA pool. Absence of judged positives is therefore composition-dependent. By contrast, positive-pool fixed-profile zeros occur in all four datasets and both availability strata. They cannot be attributed entirely to a single dataset or to missing positives. These are observed evaluation mechanisms, not explanations of semantic relevance or preference correctness.

Of 134 positive-containing pools, 120 are cyclic. Greedy changes 111 positive-pool rankings, with 37 measured changes; exact changes 99, with 25 measured changes. The remaining 74 per branch retain the evaluated profile. Among 79 zero-positive pools, greedy/exact change 69/61 rankings without changing measured utility; 10/18 rankings remain unchanged. Thus the all-query and changed-ranking denominators answer different questions. The corresponding all-query counts, including unfavorable cases, remain the principal description; conditioning does not create new evidence of efficacy.

The two-provider stratum has 57 cyclic-positive queries among 110, versus 63 among 103 in the four-provider stratum. Dataset, available providers and retained evidence vary together; these strata are not randomized treatments or isolated provider interventions. The complete dataset $\times$ availability $\times$ positive-availability table in Online Resource 1 includes an empty HotpotQA/two-provider/zero-positive cell and zero-positive cells of size one or two. These cells support counts only, without stable subgroup-effect claims. No stratum was selected for significance and no testing family is added.

Observed fixed-profile case (C1).

In a BRIGHT 2-provider selected-cohort query, greedy repair exchanges the documents at ranks 2 and 5. All moved documents are unjudged. The evaluated top-ten grade vector remains (1, 0, 0, 1, 0, 0, 0, 0, 0, 0), and full-query nDCG@10 remains 0.558508 (change zero at 10^{-12} tolerance). Unjudged is an evaluation status, not proven irrelevance. Section 8 gives the contrasting observed case in which an optimal-repair choice changes a positive document’s rank and measured utility.

7 Retrieval Effects and Uncertainty

7.1 Retrieval effects under matched repair

Table 5 summarizes the principal matched families. The canonical active and full greedy families have zero Holm rejections, as does the 110-cell larger-pool family. Online Resource 1 retains the original active-regime forest display with pointwise uncertainty; Table 6 reports effect estimates alongside the number of informative queries. Exact-versus-unrepaired families likewise retain zero rejections. These are no-demonstrated-benefit findings under the stated protocols and sensitivity, not evidence that all true effects vanish.

7.2 Informative support and pointwise uncertainty

The 35 corrected exact-versus-greedy comparisons contain 28 endpoints with 342 independent query units and seven with 290; only 4–99 query sums per endpoint are nonzero. Table 6 shows representative endpoints alongside both selected-cache strata and the six-query primary diagnostic. Its intervals are pointwise sensitivity summaries under their original family-specific procedures, not simultaneous confidence intervals. No pooled inferential family or altered weighting is introduced. Copeland’s percentile interval excludes zero while its basic interval includes zero; the respective Holm decision remains no rejection. Improving deletion cost does not demonstrate a retrieval advantage in the corrected 35-test diagnostic. It has zero Holm rejections;

Table 4 Composition of the 213 selected queries. $P+$ counts pools with a judged-positive candidate. Paired entries are greedy/exact (G/E) counts on the same queries. Same/changed refer to full rankings; Eq.+ counts changed-ranking, positive-pool cases with the same evaluated top-ten grade profile and zero metric change. ΔU counts measured utility changes. Dataset and availability panels are overlapping descriptions, not independent samples.

Group	N	$P+$	Same G/E	Changed G/E	Eq.+ G/E	ΔU G/E	Gains G/E	Harms G/E
SciDocs	77	41	10/15	67/62	16/23	19/9	7/7	12/2
FiQA	76	38	6/18	70/58	27/22	8/7	5/4	3/3
BRIGHT	42	38	14/16	28/26	18/18	8/6	2/3	6/3
HotpotQA	18	17	3/4	15/14	13/11	2/3	1/2	1/1
Two-provider	110	71	30/49	80/61	31/26	19/12	8/7	11/5
Four-provider	103	63	3/4	100/99	43/48	18/13	7/9	11/4
All queries	213	134	33/53	180/160	74/74	37/25	15/16	22/9

Table 5 Principal matched result families. Test counts are hypotheses, not query counts. Separate rows do not define a combined family-wise guarantee

Contrast / scope	Query support	Tests	Holm rejections
Greedy-unrepaired, active canonical	50–120 per dataset	20	0
Greedy-unrepaired, full canonical	342 roster, repeated cells	60	0
Greedy-unrepaired, larger pools	342 roster, repeated cells	110	0
Exact-unrepaired, canonical	342 roster	36	0
Exact-unrepaired, larger pools	342 roster	56	0
Exact-greedy, corrected pooled	342; 290 at nDCG@20	35	0

the minimum adjusted p is 0.85041496. Priority-topological nDCG@10 has mean exact-minus-greedy effect -0.002229580 across 342 query units, raw $p=0.02429757$ and Holm $p=0.85041496$. Its query sums are positive/negative/zero for 30/44/268 identities. A pointwise percentile interval $[-0.004356624, -0.000334221]$ does not establish a family-corrected discovery. The separate saved 399-cell per-dataset/regime family also reports zero rejections and was not rerun in the correction.

The corrected six-query primary aggregate has mean nDCG 0.881750 unrepaired and 0.882760 after either selected repair, a difference of $+0.001010$ with pointwise interval $[0, 0.003030]$.

One query improves and five tie; raw and adjusted p -values equal one. This ten-test family cannot attain its first Holm threshold with only six queries even if every query changes: the smallest attainable two-sided exact sign-flip p -value is $2/2^6 = 0.03125$, exceeding $0.05/10 = 0.005$. In the selected cache, exact means improve by $+0.003164$ and $+0.004840$ in the 110/103 strata, each with Holm $p=1$ in the existing exploratory family. Greedy differences are -0.000950 and -0.002248 . Positive selected point estimates do not establish a general benefit.

Complementary external controls help assess whether the repaired pipeline earns its complexity. On 213 selected queries, SpringRank, unrepaired

Table 6 nDCG@10 effects and pointwise 95% intervals. N is distinct queries; $n_{\neq 0}$ counts nonzero query sums for the pooled corrected family and nonzero query differences otherwise. The pooled exact-minus-greedy estimand retains regime-row weighting. Other rows use their original query means. Intervals are percentile except where basic is explicitly shown; Explor. denotes the original exploratory family; each Holm decision belongs only to the stated original family. These are selected informative displays, not a new testing family.

Comparison	Family	N	$n_{\neq 0}$	Mean Δ	Pointwise interval	Holm p
Copeland E-G	Corrected 35	342	63	-0.001785	$[-0.003584, -0.000112]$	1
basic sensitivity	same endpoint	342	63	same	$[-0.003459, 0.000013]$	1
Priority-topological E-G	Corrected 35	342	74	-0.002230	$[-0.004357, -0.000334]$	0.850415
G-U, two-provider	Explor. 502	110	19	-0.000950	$[-0.007574, 0.005823]$	1
E-U, two-provider	Explor. 502	110	12	0.003164	$[0.000060, 0.007053]$	1
G-U, four-provider	Explor. 502	103	18	-0.002248	$[-0.014272, 0.009574]$	1
E-U, four-provider	Explor. 502	103	13	0.004840	$[-0.001427, 0.014088]$	1
G/E-U, primary aggregate	Primary 10	6	1	0.001010	$[0, 0.003030]$	1

Copeland, greedy Copeland and exact Copeland means are 0.357436, 0.353838, 0.352261 and 0.357813; exact exceeds SpringRank by 0.000377. None of the eight matched SpringRank contrasts survives Holm correction. Table 7 adds the trained BGE and adapted PRP-Graph controls. Training, model family, acquisition and compute differ; the table is matched-document effectiveness context, not an isolated deletion test or a confirmatory ranking of model families.

BGE’s exploratory paired differences on the PRP pool are +0.075682 versus BM25, interval $[-0.004973, 0.158395]$, and +0.084332 versus adapted PRP-Graph, interval $[0.008773, 0.180777]$. These pointwise, unregistered comparisons are not confirmatory superiority tests. The adapted PRP comparison’s eight primary native/random-acquisition contrasts all have Holm $p=1$. The strict prospective repair effects are zero for every query; the inactive opportunity documented above limits that null’s interpretation. Invalid direct-listwise outputs, zero valid Qwen and ten valid Smol rankings among 20 attempts each, cannot serve as complete baselines or be scored zero.

7.3 Limits of the completed prospective instrument

The strict prospective null must be read with the instrument’s observed dispositions. Qwen abstains on 1,378 of 2,920 valid pairwise outputs. Smol selects displayed A on 2,591 of 2,772 directional responses, despite reversed presentation. There are 148 capped invalid outputs; these belong to invalid-response accounting, not an additional independent sample. Abstentions, invalids and reversal conflicts do not become directional votes. Repeated observations assess reliability without adding primary votes or query units. Parsing success alone therefore does not establish an informative instrument.

Strict construction produces no cyclic-positive graph: its sole cyclic model/query graph has no judged-positive candidate, and all 14 positive-containing model/query graphs are acyclic. The resulting zero repair effects test an almost inactive intervention under a compromised instrument, not equivalence in reliable cyclic-positive evidence. Invalid listwise attempts remain missing controls

Table 7 Completed external controls: full-query nDCG@10 on two disjoint 20-query pools. Shared BM25/BGE prospective values are not replicated observations. U/G/E denote unrepaired/greedy/exact Copeland

Method	PRP pool	Prospective Qwen	Prospective Smol
BM25	0.278865	0.080112	0.080112
BGE cross-encoder	0.354547	0.108610	0.108610
Adapted PRP-Graph	0.270215	—	—
U-Copeland	0.271771	0.126872	0.097435
G-Copeland	0.271771	0.126872	0.097435
E-Copeland	0.271526	0.126872	0.097435
SpringRank	0.259999	0.134091	0.085859
Bradley-Terry-style	0.254006	0.134091	0.089653

(20/20 Qwen and 10/20 Smol), without zero imputation. Full dispositions, model-specific repeats and the unsuccessful neutral-schema development are retained in Online Resource 1. These observations do not establish how repair would behave with reliable contradictory preferences.

8 Finite Optimal-Repair Diagnostic

The six-query complete envelope supplies a different diagnostic: structural cost does not identify external utility even among optimal deletion sets. Table 8 retains all six instances. Four have a unique optimum. FiQA/2377 has two optimal repairs exchanging grade-one documents at ranks two and three, with unchanged utility. One SciDocs graph has two cost-0.75 repairs that exchange a grade-one and a grade-zero document: placing the positive at rank seven gives 0.956591277, while rank eight gives 0.950531008.

This finite range over a complete optimum set is not a confidence interval. The certified mean utility range is [0.881749864, 0.882759909], corresponding to improvement [0, +0.001010045] over unrepaired Copeland. The selected SCIP optimum is already utility-best on the affected query. An equally optimal choice removes the pilot’s sole improvement and leaves six ties; no certified choice produces a negative effect or strict sign reversal. The 0.006060269 per-query width demonstrates conditional non-identification, not a population effect, new algorithm or deployable relevance-aware selector.

9 Discussion and Limitations

9.1 Where structural changes cease to affect retrieval

The main finding is a difference in the composition of measured zeros. Many score-derived graphs and almost all strict compact-model graphs are acyclic, leaving little opportunity for deletion. That explanation does not account for the selected cache, where most aggregates are cyclic and both repairs frequently change the ranking. Relevance-profile attribution identifies what happens next: some changed rankings contain no judged-positive candidate, while others rearrange documents without changing the evaluated gain positions despite positive candidates being available. These observations locate different points at which structural activity ceases to affect the measured outcome.

The mathematical invariance of nDCG under a fixed grade sequence explains an individual zero but does not establish its frequency or composition. The selected cohort shows that fixed-profile cases with positive candidates occur across every dataset and both availability strata; zero-positive pools are instead concentrated in SciDocs and FiQA. The absence of judged-positive candidates is therefore insufficient to account for all observed inactivity. Conversely, an unchanged measured profile does not establish that the new ordering is semantically equivalent: many moved documents are unjudged. The attribution describes what the evaluation observes, not the correctness of each comparison.

Table 8 Utility ranges over all optimal repairs of six primary graphs under fixed Copeland extraction. SciDocs identifiers are shortened. The selected SCIP repairs attain the maxima; utilities are rounded to six decimals

Dataset / query	Cost	Repairs	Min nDCG	Max nDCG
FiQA /11039	1.50	1	0.835891	0.835891
FiQA /2377	1.00	2	0.727330	0.727330
FiQA /4002	0.25	1	0.776747	0.776747
SciDocs /067405...	0.25	1	1.000000	1.000000
SciDocs /1b1aed...	0.50	1	1.000000	1.000000
SciDocs /970093...	0.75	2	0.950531	0.956591

The retained gains and harms rule out a uniformly inactive account of repair. Some rankings change measured relevance, in both directions. The difference between the branches’ gain/harm counts is descriptive, not evidence that exact repair generally selects better outcomes. Likewise, provider availability varies with dataset and retained evidence, so the stratum differences cannot isolate a provider effect. The cohorts reveal different experimental opportunities rather than contradictory estimates from interchangeable samples.

9.2 What exact optimization and evaluation controls establish

Exact repair controls one explanation for a weak retrieval result: failure to minimize the stated deletion cost. It does not remove the mismatch between that cost and external relevance. The finite diagnostic makes a second distinction concrete. Even after minimum cost is established, an equally optimal deletion can change the position of a judged-positive document and hence utility. Complete enumeration identifies that ambiguity for the six specified graphs; it establishes neither its prevalence nor a way to choose the better optimum without relevance labels. A single returned optimum and a complete utility envelope answer different questions.

These findings support evaluating graph repair along the path from comparisons to rankings to measured relevance. A matched unrepaired control isolates deletion; direct graph aggregators and trained relevance models answer complementary pipeline questions. Reporting cyclicity, ranking response, evaluated profiles and gains/harms alongside an average effect makes inactive and

informative comparisons distinguishable. Query identities determine inferential support, rather than the number of responses or graph variants. This is a methodological implication of the observed decomposition, not an experimentally validated rule for selecting repair on an unseen query. Earlier selector and mixed-protocol headroom studies have not established a generalizing utility selector, but do not prove one impossible.

9.3 Measurement, sampling and inferential limitations

The score-derived experiments use shallow candidate pools and a controlled multi-ranker substrate. The selected cache is heterogeneous and nonrandom, while the six-query diagnostic uses relevance-enriched pools. The prospective compact-model collection avoids that particular selection mechanism but supplies no strict cyclic-positive case under a compromised instrument. Its null therefore offers little evidence about repair with reliable contradictory preferences. Alternative constructions reuse the same queries and change evidential quality as well as cyclicity; format validity, positional stability and repeatability do not establish semantic correctness. The neutral-schema comparison did not validate a better instrument.

Incomplete judgments limit every relevance-based interpretation. A separate 20-query candidate diagnostic increases judged-positive coverage from 6 to 8 pools and the oracle ceiling by +0.069106, with interval $[-0.016958, 0.164250]$, but loses coverage in one domain. This is uncertain measured headroom, not an achieved ranking benefit. Neither that diagnostic nor the profile

attribution rules out better candidates, alternative ranking policies or useful preferences among unjudged documents.

The inferential conclusions depend on independent query units and whole-query sign-exchangeability assumptions, not randomized assignment of deterministic methods. Sparse nonzero differences and separate multiplicity families restrict what can be concluded from no rejection. The canonical finite-cell median minimum detectable effect is 0.0201 versus an observed absolute mean effect of 0.0036; this sensitivity calculation is neither a bound on the true effect nor a practical-importance threshold. Only a minority of equivalence cells meet their declared margins. Pointwise intervals and Holm-adjusted decisions have different roles; the evidence does not establish broad equivalence.

Finally, the profile attribution concerns fixed Copeland rankings, not every extractor. Complete historical exact-ranking permutations are unavailable for the canonical scalar outcomes, and the all-optimum envelope covers only six eight-node graphs. Computational verification of the supplied processed inputs cannot regenerate historical model serving, recover missing scorer revisions or broaden the sampling scope. Hosted-provider identities and some dataset conversions are incompletely pinned. External baselines differ in training, truncation and compute and do not constitute a comprehensive retrieval leaderboard. Central attribution and envelope inputs remain unavailable for independent external reproduction pending sharing clearance and a confirmed confidential access route; the declarations state the access limits.

10 Conclusion

Preference-graph repair can resolve cycles without changing the extracted ranking, and can change that ranking without changing measured retrieval effectiveness. The selected-cohort decomposition shows that missing judged positives and fixed positive-pool relevance profiles are distinct explanations of this separation. Their different dataset and availability composition adds empirical knowledge beyond the metric-invariance identity. Gains and harms coexist with these zeros; corrected comparisons supply no Holm-adjusted rejection, and the finite optimum diagnostic shows

that minimum structural cost need not identify utility. The findings support separate evaluation of structural, ranking and relevance responses under matched conditions. Nonrandom cohorts, incomplete judgments, limited informative support and inferential assumptions preclude a universal negative conclusion, equivalence claim or validated repair-selection policy.

Supplementary information

Online Resource 1 (ESM_1.pdf): protocols, query populations, complete comparison results, structural and retrieval diagnostics, source identities, reuse and reproducibility limits.

Acknowledgments

The author gratefully acknowledges Professor Ioannis Koutis for his academic guidance and support.

Statements and Declarations

Data availability.

The study uses SciDocs, FiQA, HotpotQA and BRIGHT; Online Resource 1 identifies their sources, versions and conversion limits. Public derived judgment and ranking-output resources provide only part of the evidence. Processed observations support reconstruction of all 35 corrected statistical comparisons, but do not supply the inputs for every central analysis. A private extension has been validated for the 213-query relevance-profile attribution and six-graph optimum envelope; it is *not cleared for distribution*, and independent reviewer access has not been established. Confidential delivery through a verified non-public editorial route remains unconfirmed. Raw benchmark text, provider responses, source qrels and complete ranking/graph projections are not publicly redistributed. Verification of processed inputs does not establish reproduction of historical scoring or model serving. Detailed access conditions and computational coverage are described in Online Resource 1.

Code availability.

Analysis code is available in the public repository <https://github.com/SoroushVahidi/consistency>

-aware-llm-rankin. Public code access does not provide the restricted inputs or authorize their redistribution. Data-free verification procedures and source documentation specify the required inputs; full external reproduction remains subject to the data-access and historical-provenance limitations above.

Funding.

This work received in-kind computational support through the Cohere Labs Catalyst Grant Program, the Google Cloud Research Credits Program, Microsoft Azure for Students, and Fireworks AI credits provided through the AMD AI Developer Program. These resources supported historical multi-provider model judgments. The score-derived experiments used stored BM25, TF-IDF and MiniLM scores and local open-source SCIP solving. The prospective open-weight study used NJIT Wulver high-performance computing resources. The supporting organizations had no role in study design, analysis, interpretation or manuscript preparation.

Competing interests.

The author declares no competing interests relevant to this work.

Author contributions.

Soroush Vahidi conceived the study, designed and implemented the experimental pipeline, conducted the original experiments, analysed the results and prepared the manuscript and figures. Anthropic Claude and OpenAI ChatGPT/Codex assisted with software development, auditing, figure preparation and drafting/revision, as disclosed in Experimental Design. The author retains responsibility for scientific verification, final content and approval.

Ethics and consent.

No new human participants or human relevance labels were collected. The study uses existing benchmarks and model outputs. Participant consent and consent for publication of individual personal data are not applicable.

AI-assisted tools.

Anthropic Claude and OpenAI ChatGPT/Codex assisted with development, audit utilities, figures

and drafting/revision, as described in Section 4.4. The author retains responsibility for verification and final approval.

References

- [1] Achterberg T (2009) SCIP: Solving constraint integer programs. *Mathematical Programming Computation* 1(1):1–41. <https://doi.org/10.1007/s12532-008-0001-1>
- [2] Ailon N, Charikar M, Newman A (2008) Aggregating inconsistent information: Ranking and clustering. *Journal of the ACM* 55(5):23:1–23:27. <https://doi.org/10.1145/1411509.1411513>
- [3] Baharev A, Schichl H, Neumaier A, et al (2021) An exact method for the minimum feedback arc set problem. *ACM Journal of Experimental Algorithmics* 26:1.4:1–1.4:28. <https://doi.org/10.1145/3446429>
- [4] Beijing Academy of Artificial Intelligence (2024) BGE-reranker-v2-m3: Official model card and inference procedure. URL <https://huggingface.co/BAAI/bge-reranker-v2-m3>, pinned revision 953dc6f6f85a1b2dbfca4c34a2796e7dde08d41e; accessed 25 October 2026
- [5] Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4):324–345. <https://doi.org/10.2307/2334029>
- [6] Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, pp 25–32, <https://doi.org/10.1145/1008992.1009000>
- [7] Carterette BA (2012) Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems* 30(1):4:1–4:34. <https://doi.org/10.1145/2094072.2094076>
- [8] Chen Y, Liu Q, Zhang Y, et al (2025) TourRank: Utilizing large language models

- for documents ranking with a tournament-inspired strategy. In: *Proceedings of the ACM on Web Conference 2025*. ACM, pp 1638–1652, <https://doi.org/10.1145/3696410.3714863>
- [9] Cohan A, Feldman S, Beltagy I, et al (2020) SPECTER: Document-level representation learning using citation-informed transformers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pp 2270–2282, <https://doi.org/10.18653/v1/2020.acl-main.207>, URL <https://aclanthology.org/2020.acl-main.207/>
- [10] Cormack GV, Clarke CLA, Büttcher S (2009) Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, pp 758–759, <https://doi.org/10.1145/1571941.1572114>
- [11] De Bacco C, Larremore DB, Moore C (2018) A physical model for efficient ranking in networks. *Science Advances* 4(7):eaar8260. <http://doi.org/10.1126/sciadv.aar8260>
- [12] Dwork C, Kumar R, Naor M, et al (2001) Rank aggregation methods for the web. In: *Proceedings of the 10th International Conference on World Wide Web*. ACM, Hong Kong, Hong Kong, pp 613–622, <https://doi.org/10.1145/371920.372165>
- [13] Eades P, Lin X, Smyth WF (1993) A fast and effective heuristic for the feedback arc set problem. *Information Processing Letters* 47(6):319–323. [https://doi.org/10.1016/0020-0190\(93\)90079-O](https://doi.org/10.1016/0020-0190(93)90079-O)
- [14] Grötschel M, Jünger M, Reinelt G (1984) A cutting plane algorithm for the linear ordering problem. *Operations Research* 32(6):1195–1220. <https://doi.org/10.1287/opre.32.6.1195>
- [15] Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70. URL <https://www.jstor.org/stable/4615733>
- [16] Hu Z, Zhang J, Xiong Z, et al (2026) Towards acyclic preference evaluation of language models via multiple evaluators. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 21903–21911, <https://doi.org/10.1609/aaai.v40i26.39343>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/39343>
- [17] Hunter DR (2004) MM algorithms for generalized bradley–terry models. *The Annals of Statistics* 32(1):384–406. <https://doi.org/10.1214/aos/1079120141>
- [18] Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20(4):422–446. <https://doi.org/10.1145/582415.582418>
- [19] Jiang X, Lim LH, Yao Y, et al (2011) Statistical ranking and combinatorial hodge theory. *Mathematical Programming* 127(1):203–244. <https://doi.org/10.1007/s10107-010-0419-x>
- [20] Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* 1(2):259–269. <https://doi.org/10.1177/2515245918770963>
- [21] Luo J, Chen X, He B, et al (2024) PRP-graph: Pairwise ranking prompting to LLMs with graph aggregation for effective text re-ranking. In: Ku LW, Martins A, Srikumar V (eds) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, pp 5766–5776, <https://doi.org/10.18653/v1/2024.acl-long.313>
- [22] Maher S, Miltenberger M, Pedroso JP, et al (2016) PySCIPOpt: Mathematical programming in python with the SCIP optimization suite. In: *Mathematical Software – ICMS 2016*. Springer International Publishing, p 301–307, https://doi.org/10.1007/978-3-319-42432-3_37

- [23] Maia M, Handschuh S, Freitas A, et al (2018) WWW’18 open challenge: Financial opinion mining and question answering. In: Companion of the The Web Conference 2018 on The Web Conference 2018. ACM, Lyon, France, pp 1941–1942, <https://doi.org/10.1145/3184558.3192301>
- [24] Negahban S, Oh S, Shah D (2017) Rank centrality: Ranking from pairwise comparisons. *Operations Research* 65(1):266–287. <https://doi.org/10.1287/opre.2016.1534>
- [25] Qin Z, Jagerman R, Hui K, et al (2024) Large language models are effective text rankers with pairwise ranking prompting. In: Duh K, Gomez H, Bethard S (eds) Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, Mexico City, Mexico, pp 1504–1518, <https://doi.org/10.18653/v1/2024.findings-naacl.97>
- [26] Su H, Yen H, Xia M, et al (2025) BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In: The Thirteenth International Conference on Learning Representations (ICLR), URL <https://openreview.net/forum?id=ykuc5q381b>, spotlight
- [27] Sun W, Yan L, Ma X, et al (2023) Is ChatGPT good at search? investigating large language models as re-ranking agents. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 14918–14937, <https://doi.org/10.18653/v1/2023.emnlp-main.923>
- [28] Thakur N, Reimers N, Rücklé A, et al (2021) BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. Curran Associates, Inc., Red Hook, NY, USA, URL <https://openreview.net/forum?id=wCu6T5xFjeJ>
- [29] Urbano J, Lima H, Hanjalic A (2019) Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, USA, pp 505–514, <https://doi.org/10.1145/3331184.3331259>
- [30] Vahidi S (2026) Consistency-aware reranking via preference graph repair: Structural gains and conditional retrieval effects. Research Square, version 1, posted 17 June 2026, <https://doi.org/10.21203/rs.3.rs-9335700/v1>, public preprint; reused evidence and version differences are disclosed in Online Resource 1
- [31] Vahidi S (2026) Training-free ranking from pairwise comparisons via acyclic graph construction. *The Journal of Supercomputing* 82:705. <https://doi.org/10.1007/s11227-026-08852-4>
- [32] Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12(4):5–33. <https://doi.org/10.1080/07421222.1996.11518099>
- [33] Xu Y, Ruis L, Rocktäschel T, et al (2025) Investigating non-transitivity in LLM-as-a-judge. In: Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 267. PMLR, pp 69583–69612, URL <https://proceedings.mlr.press/v267/xu25w.html>
- [34] Yang Z, Qi P, Zhang S, et al (2018) HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 2369–2380, <https://doi.org/10.18653/v1/D18-1259>, URL <https://aclanthology.org/D18-1259/>
- [35] Zeng Y, Tendolkar O, Baartmans R, et al (2026) LLM-RankFusion: Mitigating intrinsic inconsistency in LLM-based ranking. *Transactions on Machine Learning Research* URL <https://openreview.net/forum?id=VUY0j74Yes>

- [36] Zhuang S, Zhuang H, Koopman B, et al (2024) A setwise approach for effective and highly efficient zero-shot ranking with large language models. In: Yang GH, Wang H, Han S, et al (eds) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Washington DC, USA, pp 38–47, <https://doi.org/10.1145/3626772.3657813>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.pdf](#)