

Online Resource 1: 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;

International Journal of Data Science and Analytics

This Online Resource provides the experimental protocols, complete comparison results and supplementary diagnostics for the main paper. It specifies candidate selection, graph construction, repair and extraction, relevance normalization, statistical families and instrument reliability. Source identities and computational verification are documented separately from the scientific results. The final section describes which evidence is available and the limits on independent reproduction.

Reading guide

Main question	Location in this Online Resource
Cohorts and measurement	Section S1: pools, construction, model/corpus identities and prior resource
Repair/extraction controls	Section S2: fixed methods, ties and optimum completeness
Endpoints and inference	Section S3: endpoint contracts, family map and corrected comparisons
Structural opportunities	Section S4: cost, cycles and repeated construction views
Ranking/grade attribution	Section S5: selected-cache profiles and descriptive cases
External retrieval controls	Section S6: matched rankers and candidate coverage
Instrument validity	Section S7: invalidity, abstentions, repeats and unsuccessful prompt development
Evidence reuse and access	Section S8: evidence reuse, historical limits and data access

S1 Detailed Protocols and Provenance

S1.1 Canonical score-derived cohorts

The canonical score sources are BM25 lexical retrieval, term frequency–inverse document frequency (TF-IDF) and the all-MiniLM-L6-v2 sentence encoder. Relevance judgments are abbreviated qrels; a directed acyclic graph is abbreviated DAG. The principal endpoint is normalized discounted cumulative gain (nDCG), whose ideal discounted cumulative gain (IDCG) denominator follows the original study contract. BM25 uses $k_1 = 1.5$, $b = 0.75$; TF-IDF uses cosine logarithmic term weighting. Source scores are held fixed and represent a controlled multi-source substrate, not a current retrieval leaderboard. Candidates are the top- P reciprocal rank fusion (RRF) order over the union of native source lists. Canonical cutoff equals pool size; the larger-pool study includes both $P > k$ cells permitting membership changes and $P = k$ reference cells (Table S1).

Table S1 Datasets and evaluation cells for the offline reranking study. Usable: queries retaining at least one relevant candidate after qrels filtering. Depth: common complete stored score depth across BM25, TF-IDF, and MiniLM. Canonical cells set pool size P equal to metric cutoff k ; the larger-pool study includes $P > k$ cells permitting top- k membership changes and $P = k$ reference cells.

Dataset	Usable	Depth	Evaluated (P, k) cells
SciDocs	120	50	canonical (20, 20); larger-pool (20, 5), (20, 10), (20, 20), (50, 5), (50, 10), (50, 20)
FiQA	120	50	same as SciDocs
HotpotQA	52	35	canonical (10, 10); larger-pool (10, 5), (35, 5), (35, 10), (35, 20)
BRIGHT	50	50	same as SciDocs

For a query-ranker score vector on available documents $D_{q,s}$, normalization is

$$\bar{x}_{q,s}(d) = \frac{x_{q,s}(d) - \min_{d' \in D_{q,s}} x_{q,s}(d')}{\max_{d' \in D_{q,s}} x_{q,s}(d') - \min_{d' \in D_{q,s}} x_{q,s}(d')}. \quad (\text{S1})$$

A range below 10^{-12} maps to a zero vector; missing scores remain unavailable. Retention matching first records raw-margin per-ranker retained-vote rates and regime edge-count targets. The normalized source thresholds use empirical margin quantiles; a discrete aggregate-threshold search minimizes the absolute retained-edge-count difference, breaking ties toward smaller thresholds. Vote thresholds are per dataset/ranker and shared across regimes; aggregate thresholds are per dataset/regime. This unsupervised calibration uses the same evaluated usable query list, without qrel tuning or a held-out calibration split. Table S2 and Algorithm S1 retain the original contracts.

Table S2 Retention-matched thresholds for the primary protocol (`primary_minmax_retention_matched`). Vote thresholds are shared across regimes within a dataset; aggregate thresholds and drop-mutual flags are regime-specific. Matching target: raw-margin retained-vote rates (per ranker) and raw retained edge count (per regime). Calibration: unsupervised collection-level protocol thresholds on the same usable queries later evaluated; qrels/nDCG are not used (transductive query list; the main internal-validity discussion). Regime abbreviations: `ms1`, `ms1dm=ms1_drop_mutual`, `ms2`.

Dataset	Reg.	BM25	MiniLM	TF-IDF	Agg.	Sup.	Drop
SciDocs	<code>ms1</code>	0.00468	0.274	0.340	0.0108	1	no
SciDocs	<code>ms1dm</code>	0.00468	0.274	0.340	0	1	yes
SciDocs	<code>ms2</code>	0.00468	0.274	0.340	0	2	no
FiQA	<code>ms1</code>	0.00621	0.276	0.375	0.0120	1	no
FiQA	<code>ms1dm</code>	0.00621	0.276	0.375	0	1	yes
FiQA	<code>ms2</code>	0.00621	0.276	0.375	0	2	no
HotpotQA	<code>ms1</code>	0.00317	0.272	0.228	0	1	no
HotpotQA	<code>ms1dm</code>	0.00317	0.272	0.228	0.00341	1	yes
HotpotQA	<code>ms2</code>	0.00317	0.272	0.228	0.295	2	no
BRIGHT	<code>ms1</code>	0.00108	0.240	0.555	0.0114	1	no
BRIGHT	<code>ms1dm</code>	0.00108	0.240	0.555	0.0119	1	yes
BRIGHT	<code>ms2</code>	0.00108	0.240	0.555	0	2	no

Alternative pool policies use equal per-ranker depth, round-robin interleaving, a single-ranker pool or CombSUM. These are construction sensitivities; baseline comparisons in the canonical pool do not constitute an interaction test proving pool-policy neutrality.

S1.2 Prospective observations and construction rules

Twenty queries and BM25 top-12 pools were frozen before relevance evaluation: seven FiQA, seven SciDocs and six BRIGHT biology. Each of 66 unordered pairs per query receives two orientations from each of two pinned models, giving 5,280 first observations. Seven pairs per query are independently repeated in both orientations, adding 560 responses; 40 separate listwise attempts bring the total to 5,880. Repeats estimate reliability and never add primary votes.

Displayed positions are mapped back to document identity before reconciliation. Valid TIE and INSUFFICIENT_INFORMATION choices are abstentions. Invalid schema, missing fields, unsupported choices/confidence labels and capped incomplete responses are unavailable judgments, never default A votes. The strict primary rule and four exploratory construction rules are:

1. Strict agreement: retain a unit edge only if both orientations are valid, directional and agree on the same document.

Algorithm S1 Canonical preference-graph construction for one query

Require: Query q ; candidate set D_q ; base rankers $S = \{\text{BM25}, \text{TF-IDF}, \text{MiniLM}\}$ with scores $x_{q,s}$; construction regime $r \in \{\text{ms2}, \text{ms1}, \text{ms1_drop_mutual}\}$

Ensure: Weighted directed preference graph $G_q = (V_q, E_q, w_q)$

```
1:  $V_q \leftarrow D_q$ ; initialize an empty weighted edge accumulator
2: for each ranker  $s \in S$  do
3:   Min-max normalize  $x_{q,s}$  to  $\bar{x}_{q,s}$ 
4:   for each unordered candidate pair  $\{u, v\} \subseteq D_q$  scored by  $s$  do
5:     if  $|\bar{x}_{q,s}(u) - \bar{x}_{q,s}(v)|$  is below the vote-margin threshold, or  $s$  does not score
       both  $u$  and  $v$  then
6:       Abstain on  $\{u, v\}$  for ranker  $s$ 
7:     else
8:       Accumulate directional support for  $u \succ v$  or  $v \succ u$  (whichever  $\bar{x}_{q,s}$ 
       supports), weighted by the normalized margin
9:     end if
10:  end for
11: end for
12: Apply regime  $r$ 's support rule and aggregate weight threshold to every accumu-
    lated directed pair
13: if  $r = \text{ms1\_drop\_mutual}$  then
14:   Remove every unordered pair retained in both directions
15: end if
16:  $E_q \leftarrow$  every retained directed pair with positive aggregated weight
17: return  $G_q = (V_q, E_q, w_q)$ 
```

2. AB only: use the first orientation's valid directional winner.
3. BA only: use the reverse orientation after restoring document identity.
4. Symmetric, both valid: require both parsed responses, count directional votes as $+1/-1$ and abstentions as zero, and retain a unit edge for a nonzero net winner. Opposed votes cancel; agreement does not double the weight.
5. Symmetric, available: apply the same sum to available valid responses, permitting one valid directional response when the other is invalid.

All variants reuse 40 model/query cells and 20 query identities. Agreement and confidence categories are retention metadata, not calibrated correctness.

S1.3 Dataset, model and software identity

Official query/document IDs, text hashes, frozen pool identities and qrel snapshots are retained. New FiQA/SciDocs cohorts use BEIR-hosted snapshots; BRIGHT uses original biology examples, positive IDs and excluded-document rules. Historical HotpotQA remains a retained QA-to-retrieval export with incompletely documented conversion. Tables S3 and S4 report exact model/corpus pins. Hosted license tags do not supersede underlying source rights.

Table S3 Pinned corpus snapshots for the new cohorts. Hub license tags describe the hosted artifact; they do not override underlying source rights. Original provenance records retain qrel snapshot IDs and input hashes.

Source	Exact revision	Recorded hub tag
BeIR/fiqa	979c07a7cb5ccc6ca009792241fa12 50b98055dd	cc-by-sa-4.0
BeIR/scidocs	acebceec772628c6679d76d6cff30a d517d17f4e	cc-by-sa-4.0
xlangai/BRIGHT	3066d29c9651a576c8aba4832d2498 07b181ecae	cc-by-4.0

Table S4 Model pins. Official model cards identify Apache 2.0 for these checkpoints; archived manifests bind cached file hashes and implementation versions.

Model	Revision
Qwen2.5-7B-Instruct	acbd96531cda22292a3ceaa67e984955d3965282
SmolLM3-3B	a07cc9a04f16550a088caea529712d1d335b0ac1
BGE-reranker-v2-m3	953dc6f6f85a1b2dbfca4c34a2796e7dde08d41e
FLAN-T5-large	0613663d0d48ea86ba8cb3d7a44f0f65dc596a2a

The recorded PRP-Graph source commit is 968e45ee8178f5d3ac44a2b8f11f091af69206d7; the SpringRank source commit is f97a489e85e44c553a437e7c6f9e40a800a5ade4. The latter has inconsistent license-file/package metadata, so its third-party implementation is not redistributed in the review bundle.

The original prospective runtime records PyTorch 2.13.0+cu130, Transformers 5.16.1 and CUDA 13.0. Pair prompts use query/document truncation of 128/256 tokens, a 48-token output cap, temperature 0.7, top- $p = 0.9$ and seed 20261025162122; listwise output cap is 128 and context ceiling 4096. Model-specific chat/reasoning settings are frozen. Historical provider aliases identify observed serving names rather than immutable provider snapshots.

S1.4 Previously released derived judgment resource

Table S5 Consistency-Aware Judgments v1 configurations. These are record classes, not independent query samples

Configuration	Records	Interpretation
Pairwise judgments	48,777	Parsed provider/orientation observations
Repeated measurements	7,140	Summaries of repeated observations
Selector training	139	Historical derived feature records
Provider diagnostics	276	Provider/run diagnostic records
Total	56,332	Four configurations of one resource

Table S5 distinguishes the four record classes in the resource released on 13 August 2026. They describe different observations or summaries within one resource, not independent datasets. Scientific values and counts refer to that release; the current analyses do not create a second publication of those observations.

The resource includes SciDocs, HotpotQA and BRIGHT and excludes FiQA/Cohere and raw benchmark text. It does not contain the October prospective collection or supply every input used by the paper. Pair, provider, orientation and repeated-measurement fields must remain distinguishable. Some historical confidence/probability fields are absent; they must not be manufactured from winner labels. The exact independent-query sample cannot be inferred from the aggregate release receipt. Its CC BY-SA 4.0 license and original attribution remain in the confidential provenance record; its payload is not bundled or anonymously relicensed.

The separate numerical ranking-pool release contains 93,504 derived method measurements over 66 configurations and 342 query identities, not raw pairwise judgments or a new independent retrieval corpus. Structural MWFAS resources concern related graph-benchmark research, not additional retrieval observations. None of these releases is pooled as extra independent query support.

S2 Repair, Extraction and Optimal-Envelope Details

S2.1 Repair and extraction instantiations

For distinct nodes, binary b_{uv} means u precedes v . The established linear-ordering model is

$$\min \sum_{(u,v) \in E} w(u,v)b_{vu}, \quad (\text{S2})$$

$$\text{subject to } b_{uv} + b_{vu} = 1 \quad (u \neq v), \quad (\text{S3})$$

$$b_{ab} + b_{bc} + b_{ca} \leq 2 \quad (a, b, c \text{ distinct}), \quad (\text{S4})$$

$$b_{uv} \in \{0, 1\}. \quad (\text{S5})$$

An edge is deleted when its reverse-order variable is one. The formulation is established, not a new minimum-weight feedback arc set (MWFAS) algorithm [2, 6]. All 1,025 historical reported instances have SCIP proven-optimal status and zero reported gap. The empty BRIGHT/biology:0 ms2 graph was historically excluded; empty graphs have trivial repair, not infeasibility. Mean/median/maximum saved solve times are 7.4/0.25/236 ms under a 300-second safety limit. The frozen canonical SCIP objective matches independent brute-force results on 49 saved small cases; these are correctness fixtures, not 49 additional retrieval queries.

The computational analysis comprises 2,405 graph tasks and 56,335 metric rows across repeated methods and conditions. These are workload counts, not additional independent queries.

The offline replay uses PySCIPOpt 6.2.1/SCIP 10.0.2 and a 60-second exact safety limit. The complete-envelope diagnostic uses 30 seconds per canonical solve and 120

seconds for independent MIP enumeration. These contracts differ from the historical 300-second limit.

Table S6 records the method roles; Table S7 gives the corresponding extraction parameters. Copeland is the fixed primary extractor for the profile attribution because its complete branch permutations are available; this choice does not establish extractor-general attribution. The primary family has graph-free controls and five paired graph rules; the expanded exact-versus-unrepaired family adds PageRank, Rank Centrality, Markov hybrid and Bradley–Terry-style aggregation. Topological and prior-priority topological extraction require a DAG and enter exact-versus-greedy checks rather than an unrepaired cyclic-graph contrast.

Table S6 Controlled comparison set for isolating graph construction and repair (not a retrieval leaderboard). Family: primary = greedy repaired-vs-unrepaired; expanded = exact repaired-vs-unrepaired; struct. = exact-vs-greedy only. Repair variants: U = unrepaired, G = greedy, E = exact.

Method	Family	Graph?	Prior?	Repair variants compared
Prior-only	primary, expanded	no	yes	n/a (no graph)
RRF	primary, expanded	no	no	n/a (no graph)
CombSUM	primary, expanded	no	no	n/a (no graph)
Borda fusion	primary, expanded	no	no	n/a (no graph)
Copeland	primary	yes	no	U, G, E
Balance	primary	yes	no	U, G, E
Markov (damped)	primary	yes	no	U, G, E
Copeland hybrid	primary	yes	yes ($\alpha=0.3$)	U, G, E
Balance hybrid	primary	yes	yes ($\alpha=0.3$)	U, G, E
PageRank	expanded	yes	no	U, E
Rank Centrality	expanded	yes	no	U, E
Markov hybrid	expanded	yes	yes ($\alpha=0.3$)	U, E
Bradley–Terry	expanded	yes	no	U, E
Topological	struct. only	yes	no	G vs. E only
Priority-topological	struct. only	yes	yes (tie-break)	G vs. E only

Table S7 Extraction-method implementation parameters matching the canonical pipeline. Direction: edge $u \rightarrow v$ means u preferred over v . Final ranking sorts by descending score then ascending document id unless noted.

Method	Main parameter(s) / ties	Implementation note
Copeland	$\text{deg}^+ - \text{deg}^-$; id tie-break	Isolates score 0
Balance	weighted out-in; id tie-break	Uses edge weights
Markov (damped)	restart 0.15; tol 10^{-10} ; 10^4 iter	Loser→winner + uniform restart
Rank Centrality	self-loops; tol 10^{-8} ; 200 iter	No teleport; normalize π
PageRank	damping 0.85; tol 10^{-6} ; 100 iter	Reversed graph; weighted
Bradley-Terry-style	MM-style; tol 10^{-6} ; 100 iter	Uniform init; isolates 0
Topological	NetworkX <code>topological_sort</code>	DAG-only; construction-order ties
Priority-topo	RRF prior ($k=60$) among sources	Higher prior, then larger id
Hybrids	$\alpha=0.3$; min-max each side	$h = \hat{s}^{\text{prior}} + \alpha \hat{s}^{\text{graph}}$; const. $\rightarrow 0$

For hybrids, the original normalized graph score is combined with the normalized reciprocal-rank-fusion prior:

$$h_q(d) = \hat{s}_q^{\text{prior}}(d) + \alpha \hat{s}_q^{\text{graph}}(d), \quad \alpha = 0.3. \quad (\text{S6})$$

Graph-free CombSUM sums per-query/source normalized scores; Borda fusion sums rank points based on the query union size, with absent documents contributing zero. RRF sums reciprocal rank contributions, while prior-only uses the fixed upstream fusion score. These are distinct from graph out-degree or edge-weight sums. The prior uses pool-local ranks and reciprocal-rank constant 60; standalone fusion uses each full stored ranking before restricting to the pool. These can differ without a tie. The sensitivity sweep uses 0.1/0.3/0.5/1.0, not test-label selection of a best alpha.

Direct SpringRank minimizes the attributed energy [5]

$$H(s) = \frac{1}{2} \sum_{ij} A_{ij} (s_i - s_j - 1)^2 + \frac{\lambda}{2} \sum_i s_i^2. \quad (\text{S7})$$

The primary setting is $\lambda = 0$; 0.01 and 1 are sensitivities. Each weak component is centered at zero and isolates receive zero. Absolute differences up to 10^{-10} form numerical tie groups ordered by document ID. Component offsets are not identified by comparisons. Fractional weights are not independent win counts. Bradley-Terry-style aggregation is the saved finite-iteration implementation, without a likelihood-optimality certificate.

S2.2 Complete optimum sets and utility envelopes

For the six primary aggregate graphs, hold candidates, judgments, Copeland and ties fixed. At optimum cost c_q^* , let $\mathcal{F}_q^*$ contain every feasible deletion set of that cost. With the same fixed extractor $\mathcal{R}$ as in the main paper, the conditional utility envelope is

$$[U_q^-, U_q^+] = \left[\min_{F \in \mathcal{F}_q^*} U_q(\mathcal{R}(G_q - F)), \max_{F \in \mathcal{F}_q^*} U_q(\mathcal{R}(G_q - F)) \right]. \quad (\text{S8})$$

This is an envelope over repaired graphs under fixed extraction, not all topological ordering policies for one DAG. Canonical SCIP establishes the optimum; exhaustive enumeration visits all $8! = 40,320$ node orders. Duplicate backward-edge deletion sets are collapsed. With positive costs, any optimum F admits a retained-DAG topological order whose backward set $B \subseteq F$ already breaks all cycles; optimality implies $B = F$. With nonnegative costs, additional deleted forward edges can only have zero cost and their subsets must also be enumerated. Primary graphs have no zero-cost arcs. This diagnostic argument does not introduce a new ranking theorem.

An independent deletion model uses binary d_e and positions $p_v \in \{0, \dots, n-1\}$:

$$p_u + 1 \leq p_v + nd_{uv} \quad ((u, v) \in E), \quad \sum_e w_e d_e = c_q^*. \quad (\text{S9})$$

After a solution D is found, it is excluded with the no-good constraint

$$\sum_{e \in D} (1 - d_e) + \sum_{e \notin D} d_e \geq 1. \quad (\text{S10})$$

Final infeasibility certifies completeness. All six independent enumerations terminated this way and agree with order enumeration; 33 tiny adversarial fixtures cover unequal/zero costs, isolates and disconnected graphs. A limited enumeration would provide only a lower bound on utility variation. The main paper reports all six utility ranges; the private study archive retains complete query identities and optimal deletion sets. Those complete inputs are not included in the available processed statistical evidence. The best utility is an oracle endpoint, not an available qrel-free selector.

Numerical verification.

An independent standard-library implementation enumerates all $6 \times 8! = 241,920$ orders from the processed graphs, minimizes exact scaled-integer backward-edge cost, deduplicates deletion sets, applies Copeland with the original tie order and recomputes full-query nDCG. All eight optimum repairs and all six utility ranges agree with the reference results, with maximum difference 1.11×10^{-16} . Reference objectives and repairs are read only after the computed outputs are written. This verifies the finite diagnostic conditional on its inputs, without regenerating provider judgments or producing a new MIP certificate. Strictly positive weights are essential to the completeness argument.

S3 Statistical Families and Complete Results

S3.1 Metric definitions and normalization contracts

For the complete candidate order $\pi_q = (d_1, \dots, d_{P_q})$, the numerator of nDCG@k sums $(2^{r_q(d_i)} - 1) / \log_2(i + 1)$ over $i \leq \min(k, P_q)$. IDCG sorts the grades in its reference set and sums the first $\min(k, |B_q|)$ ideal positions. Candidate-conditioned IDCG restricts B_q to the candidates; full-query IDCG uses all recorded query qrels. A full-query ideal is not truncated to the candidate count when $P_q < k$. Nonpositive IDCG returns zero. Unjudged candidates have evaluation grade zero; this convention does not establish irrelevance. Retrieval evaluation under incomplete information is an established problem [4]; that literature does not independently validate our cohorts. The canonical qrel helper retains the highest grade for duplicate document IDs.

For $z_{qi} = \mathbf{1}\{r_q(d_i) > 0\}$, $R_q = \sum_i z_{qi}$ and precision $p_q(i) = i^{-1} \sum_{j=1}^i z_{qj}$, the historical endpoints are

$$\text{AP}_q = \frac{\sum_{i=1}^{P_q} z_{qi} p_q(i)}{R_q}, \quad \text{RR}_q = \frac{1}{\min\{i : z_{qi} = 1\}}. \quad (\text{S11})$$

Both return zero when no candidate is judged positive. The MAP label aggregates this full-candidate AP; its denominator is candidate positives, not all positives in the corpus. MRR aggregates RR over the complete candidate ranking. Single-cell means give each eligible query one value. The corrected pooled family retains regime-row weighting while resampling complete query blocks, as defined in the main paper. Missing/ineligible records are excluded, never scored zero.

Table S8 maps the retained contracts. The historical pooled generator passes all query grades, with zero grades added for unjudged candidates, to the full-query nDCG routine at every reported cutoff; its AP/MRR remain candidate-ranking endpoints. These stored results are not renormalized using a later helper implementation.

Table S8 Retained endpoint and normalization contracts. C/F denote candidate-conditioned/full-query IDCG. Query means within a single cell have equal weights; the corrected pooled exact-minus-greedy statistic retains eligible regime-row weights with query-block inference. Family sizes count hypotheses, not samples. Cutoffs and missing eligibility remain as recorded.

Comparison / family	Endpoint / ideal	Cutoff and aggregation scope
Greedy-unrepaired, 20/60 canonical	nDCG / C	$k = P$: 20 except HotpotQA 10; dataset/regime/method cells
Greedy-unrepaired, 110 larger-pool	nDCG / C	Recorded (P, k) in Table S1, including reference $P = k$ cells
Exact-unrepaired, 36/56	nDCG / C	Canonical/larger-pool cells; nine fixed extraction pairs
Exact-greedy, pooled 35; separate 399	nDCG / F; MAP, MRR	$k = 5, 10, 20$ when $k \leq P$; $P = 20$ except HotpotQA 10. AP/RR use full candidate order. Pooled 35: 28 tests on 342 queries, seven nDCG@20 on 290
Cached/prospective LLM and external controls	nDCG@10 / F	Fixed candidates within each contrast; six/213-query cache and disjoint 20-query cohorts retain separate families
Candidate diagnostic	Oracle nDCG@10 / F	Qrel-ordered candidate ceiling on 20 further queries; not achieved reranking utility

S3.2 Family contracts and the pooled correction

Table S9 preserves the original canonical family boundaries. The eight alternative-pool baseline comparisons use the canonical pool; they are not a pool-policy interaction test. The corrected 35 and separate 399 families compare exact against greedy; the 36/56 families compare exact against unrepaired graphs. Table S8 identifies their denominators and cutoffs. The historical family inventories are retained; no public preregistration is asserted. The query-cluster correction protocol was frozen before its reanalysis, rather than before the original data collection.

Table S9 Retained Holm families (claim-to-inference map). Primary and robustness families retain separate error scopes; failure to reject does not establish a null or equivalence. Counts match the complete results and sensitivity tabulations below.

Family	Question	Scope	Methods	Tests
Active canonical	Greedy repair gain?	$4 \times \text{ms1}$, $P=k$	Prim. 5 pairs	20
Full canonical	All regimes?	4×3 regimes, $P=k$	Prim. 5 pairs	60
Active larger-pool	Larger-pool study?	ms1 , larger-pool	Prim. 5 pairs	110
Exact-repair canonical	Exact vs unrepaired?	$4 \times \text{ms1}$, $P=k$	Exp. 9 pairs	36
Exact-repair larger-pool	Exact, larger-pool?	larger-pool cells	Exp. 9 pairs	56
Exact vs greedy	Exact vs greedy ret.?	query-block pooled / per-cell	Prim.+topo	35/399
Alternative-pool	Baseline contrasts?	canonical pool	vs RRF/CombSUM	8
Equivalence (TOST)	Narrow equivalence?	active larger-pool	Prim. pairs	110

For corrected 35, the dataset–query is the sign/bootstrap unit. Repeated regime effects retain their original weighting in the mean. Ordinary endpoints have 1,025 eligible rows on 342 query identities; nDCG@20 has 869 rows on 290 identities. One BRIGHT/ms2 graph was historically omitted and original HotpotQA nDCG@20 was ineligible. Missing cells are not scored zero. All 35 effect estimates are unchanged by the correction, within recorded numerical tolerance.

The completed analysis uses 10,000 Monte Carlo whole-query signs, seed 17, add-one two-sided p-values and Holm across 35. The same sign multiplies every regime effect in the query. All 35 use Monte Carlo, including sparse outcomes; they are not exact enumeration. The 10,000-draw, seed 13 bootstrap resamples whole query blocks within each dataset and retains the ratio estimator, with percentile/basic point-wise 95% intervals. It does not use bias-corrected and accelerated (BCa) intervals for this correction. Query-level label-exchangeability and between-query independence remain assumptions; a weak mean-zero null alone is insufficient. No distribution-free or population-causal guarantee is inferred. Nonsignificance does not establish equivalence or a general absence of repair benefit.

Table S10 reports all 35 comparisons, with effects and adjusted p-values rounded for display. The underlying processed-observation results retain full-precision effects, raw/adjusted p-values, percentile/basic intervals and Monte Carlo diagnostics. Access conditions are described in Section S8.2.

Table S10: All completed query-block exact-minus-greedy comparisons. Effects retain their original regime-row weighting; adjusted p-values use the unchanged 35-test Holm family. Values are typeset from saved outputs, rounded for display

Extractor	Metric	Queries	Mean effect	Holm p
Balance	MAP	342	-0.000257	1.00000
Balance	MRR	342	-0.000028	1.00000
Balance	nDCG@10	342	-0.001038	1.00000
Balance	nDCG@20	290	-0.000286	1.00000
Balance	nDCG@5	342	-0.001043	1.00000
Balance hybrid	MAP	342	+0.000414	1.00000
Balance hybrid	MRR	342	+0.000932	1.00000
Balance hybrid	nDCG@10	342	+0.000201	1.00000
Balance hybrid	nDCG@20	290	+0.000390	1.00000
Balance hybrid	nDCG@5	342	+0.000408	1.00000
Copeland	MAP	342	-0.000687	1.00000
Copeland	MRR	342	-0.000198	1.00000
Copeland	nDCG@10	342	-0.001785	1.00000
Copeland	nDCG@20	290	-0.000265	1.00000
Copeland	nDCG@5	342	-0.000223	1.00000
Copeland hybrid	MAP	342	+0.000907	1.00000
Copeland hybrid	MRR	342	+0.001965	1.00000
Copeland hybrid	nDCG@10	342	+0.000005	1.00000
Copeland hybrid	nDCG@20	290	+0.000657	1.00000
Copeland hybrid	nDCG@5	342	+0.000302	1.00000
Markov	MAP	342	-0.000107	1.00000
Markov	MRR	342	+0.000592	1.00000
Markov	nDCG@10	342	-0.001111	1.00000
Markov	nDCG@20	290	-0.000016	1.00000
Markov	nDCG@5	342	-0.000621	1.00000
Priority-topological	MAP	342	-0.001648	1.00000
Priority-topological	MRR	342	-0.001864	1.00000
Priority-topological	nDCG@10	342	-0.002230	0.85041
Priority-topological	nDCG@20	290	-0.001055	1.00000
Priority-topological	nDCG@5	342	-0.001214	1.00000
Topological	MAP	342	-0.000841	1.00000
Topological	MRR	342	-0.000409	1.00000
Topological	nDCG@10	342	-0.001452	1.00000
Topological	nDCG@20	290	-0.000094	1.00000
Topological	nDCG@5	342	-0.000413	1.00000

Every corrected Holm decision is false; minimum raw/adjusted p-values are 0.02429757/0.85041496. For priority-topological nDCG@10, exact-minus-greedy effect is -0.002229579922 over 342 query units, with 30 positive, 44 negative and 268 zero query sums. Percentile interval is $[-0.004356623955, -0.000334221075]$ and basic interval $[-0.004124938769, -0.000102535889]$. The raw-p Monte Carlo standard error is approximately 0.00154; variation at that scale is not new evidence of a retrieval effect.

Two pointwise percentile intervals exclude zero, for Copeland and priority-topological nDCG@10, but neither Holm test rejects. Copeland's basic interval includes zero, illustrating interval-method sensitivity. Depending on the endpoint, only 4–99 query sums are nonzero. Sparse outcomes and selected queries limit interpretation even when the roster contains 342 identities. The historical raw 0.026/adjusted 0.906 priority-topological values remain only in labeled historical-comparator columns of the data, not current inferential support.

S3.3 Other families and finite sensitivity

The separate 399 per-dataset/regime tests retain zero Holm rejections. Their cached counts/means and one-record-per-query-per-cell structure were checked in the completed correction; their tests were not rerun. Canonical exact-versus-unrepaired 36 and larger-pool 56 retain their original estimands and no-rejection decisions. The original greedy active 20/full 60 and larger-pool 110 results are reproduced as saved tabulations in Table S11; fusion macros there are descriptive, not paired superiority tests.

Table S11 Canonical retrieval evidence under greedy repair. Panel A: primary Holm-corrected repaired-vs-unrepaired nDCG families (five-pair primary method family, Section S2). Panel B: descriptive dataset-macro mean nDCG across the reported cells for selected anchors (graph-free methods require no repair); these macros are not a paired superiority test between CombSUM and the repaired hybrid. Values are candidate-conditioned nDCG and cannot be compared directly with the full-query LLM means. No cell in Panel A survives Holm correction.

Panel A. Primary Holm families (paired repaired vs. unrepaired)			
Comparison family	Cells	Holm-rejected	Smallest Holm-adj. p
Active canonical (ms1)	20	0/20	0.240
Full canonical (all regimes)	60	0/60	0.720
Active larger-pool (ms1 , reference cells incl.)	110	0/110	0.352

Panel B. Descriptive dataset-macro mean nDCG (not paired confirmatory tests)	
Method	Mean nDCG
CombSUM (graph-free fusion)	0.554
Reciprocal rank fusion (graph-free)	0.546
Copeland hybrid, repaired (ms1)	0.546
Prior-only (graph-free)	0.545
Borda fusion (graph-free)	0.545

Table S12 summarizes the associated robustness and sensitivity checks.

Table S12 Robustness of the canonical no-rejection finding (canonical evidence only). Each row probes a distinct artifact risk; none overturns the Holm-corrected repaired-vs-unrepaired finding. MDE/TOST characterize detectable scale and limited equivalence, not a general proof of zero effect.

Check	Result
Baseline fairness	Testing whether the RRF-centered primary candidate pool gives a structural advantage to RRF-family baselines: 0/8 Holm-significant cells (RRF and CombSUM vs. the fixed repaired Copeland hybrid, canonical pool, all four datasets). No superiority detected in those contrasts; this is not an interaction test proving the pool policy neutral.
Power and minimum detectable effect	In the active larger-pool family, median observed absolute mean delta 0.0036 vs. median Holm-adjusted 80%-power minimum detectable effect 0.0201 (107 finite MDE cells out of 110; undefined values excluded): the reported detectable scale exceeds typical observed effects. MDE is a design-sensitivity calculation, not a confidence bound ruling out effects.
Narrow equivalence	Two-one-sided-tests equivalence, active larger-pool family: 13/110 cells reject non-equivalence at a ± 0.005 margin, 32/110 at ± 0.010 , after Holm correction. Equivalence is established for a minority of cells, not the majority; the correct reading of the null result is absence of reliable positive evidence, not blanket practical equivalence.
Larger-pool ($P > k$) family	Already reported in Table S11: 0/110 Holm-significant despite a 10.6% mean top- k membership-change rate (reported in the main paper), so membership change alone is not sufficient to produce a corrected retrieval gain.

The 107 finite minimum detectable effect (MDE) cells out of 110 give median absolute mean effect 0.0036 and median corrected 80%-power minimum detectable effect 0.0201. Minimum detectable effect is a sensitivity calculation, not a confidence bound or practical-importance threshold. The Holm-corrected two one-sided tests (TOST) procedure supports 13/110 cells at ± 0.005 and 32/110 at ± 0.010 ; these minority cell/margin results do not establish blanket equivalence.

The corrected cached primary family contains five graph conditions times two repairs, on six query clusters. Exact two-sided sign enumeration has minimum $p = 2/64 = 0.03125$, above its first ten-test Holm threshold 0.005. The historical best-of-greedy/exact estimand is qrel-selected and differs from the fixed strict replay. The selected-cache exploratory family retains its saved 502-test inventory, rather than a newly favorable subset. SpringRank has eight retrospective contrasts; PRP-Graph eight acquisition/native controls; strict prospective four; construction sensitivity eight; prompt development two. These are separate analysis roles and multiplicity families, not additional independent replications.

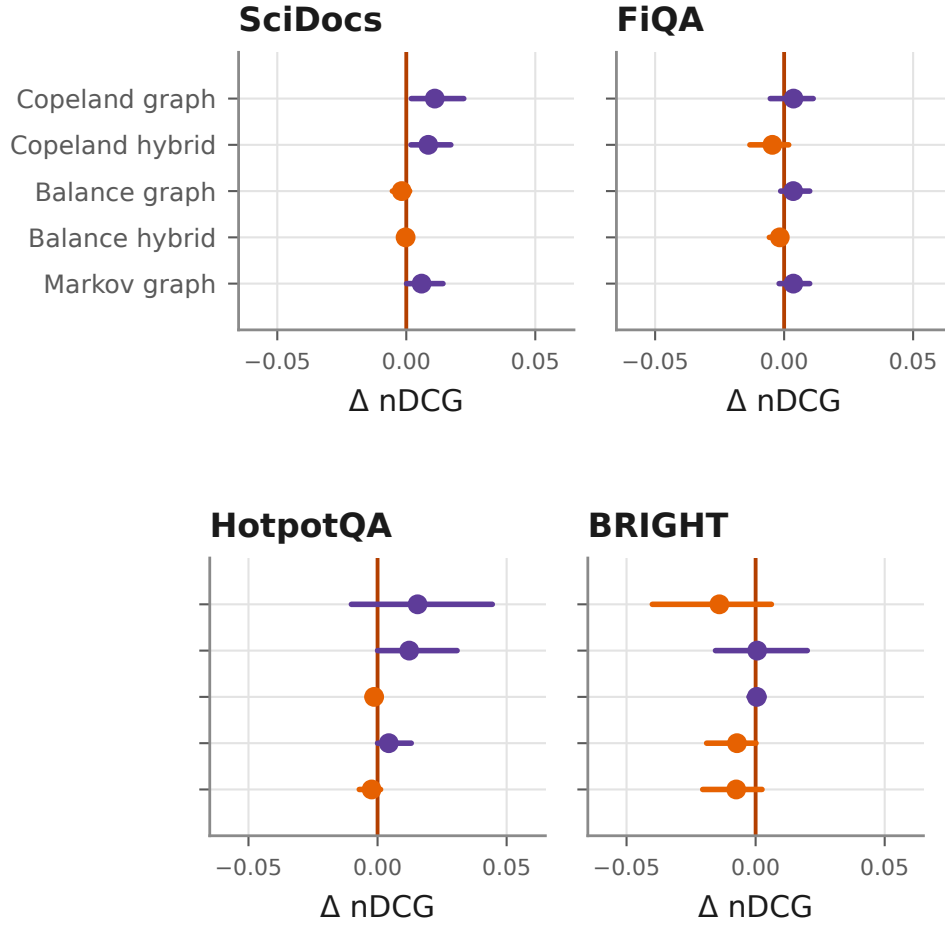


Fig. S1 Greedy-minus-unrepaired nDCG effects and pointwise 95% bootstrap intervals under ms1 and candidate-conditioned IDCG: nDCG@20 for SciDocs/FiQA/BRiGHT (120/120/50 queries), nDCG@10 for HotpotQA (52). Purple/orange marks positive/negative point estimates, not significance; the vertical line is zero. Upper panels: SciDocs and FiQA; lower panels: HotpotQA and BRiGHT, with the same method-row order. None of the active family's 20 comparisons survives Holm correction; interval positions are not corrected discovery decisions

S3.4 Retained active-regime effect display

Figure S1 retains the original pointwise effect display; its family-level decisions are unchanged.

S4 Structural and Ranking Sensitivity

S4.1 Canonical construction and cost

Figures S2 and S3 show source-scale sensitivity and cyclicity. Conditional BM25 edge-weight share is 0.988 raw/0.512 normalized. Normalization changes source influence; it does not establish calibrated correctness. Construction-time mutual dropping and post-hoc reciprocal deletion remain different interventions, with the SciDocs 11.7% versus 10.8% residual cyclicity distinction retained.

Tables S13 and S14 retain canonical cycle rates and deletion costs, respectively.

Table S13 Canonical structural rates. FAS fraction is the mean within-graph ratio of greedy-deleted weight to original total edge weight (zero for zero-weight graphs); displayed zeros may hide small positive values

Dataset	Regime	Cyclic%	Post-mutual%	FAS fraction
SciDocs	ms2	0.0	0.0	0.000
	ms1	99.2	10.8	0.074
	ms1_drop_mutual	11.7	11.7	0.002
FiQA	ms2	0.0	0.0	0.000
	ms1	98.3	30.8	0.080
	ms1_drop_mutual	30.8	30.8	0.004
HotpotQA	ms2	0.0	0.0	0.000
	ms1	63.5	1.9	0.029
	ms1_drop_mutual	1.9	1.9	0.000
BRIGHT	ms2	0.0	0.0	0.000
	ms1	92.0	22.0	0.055
	ms1_drop_mutual	22.0	22.0	0.003

Table S14 Structural gap on cyclic instances ($n = 379$, pooled regimes): mean feedback-arc weight removed by exact versus greedy repair. Relative reduction is from greedy to exact. Retrieval-level exact-versus-greedy and exact-versus-unrepaired Holm outcomes appear in the accompanying text (all nonsignificant after correction).

Dataset	Wt. removed (exact)	Wt. removed (greedy)	Reduction
SciDocs	6.14	8.19	25.1%
FiQA	4.51	6.37	29.2%
HotpotQA	1.55	1.80	13.4%
BRIGHT	3.07	3.95	22.3%

On the 379 cyclic query-regime records, greedy/exact deletion sets differ in 87.9%, and 333 have strictly lower exact deletion cost. The average recorded deletion weights are 6.24 greedy/4.60 exact, a 26.3% mean reduction. These structural comparisons do not identify the sign of retrieval effects.

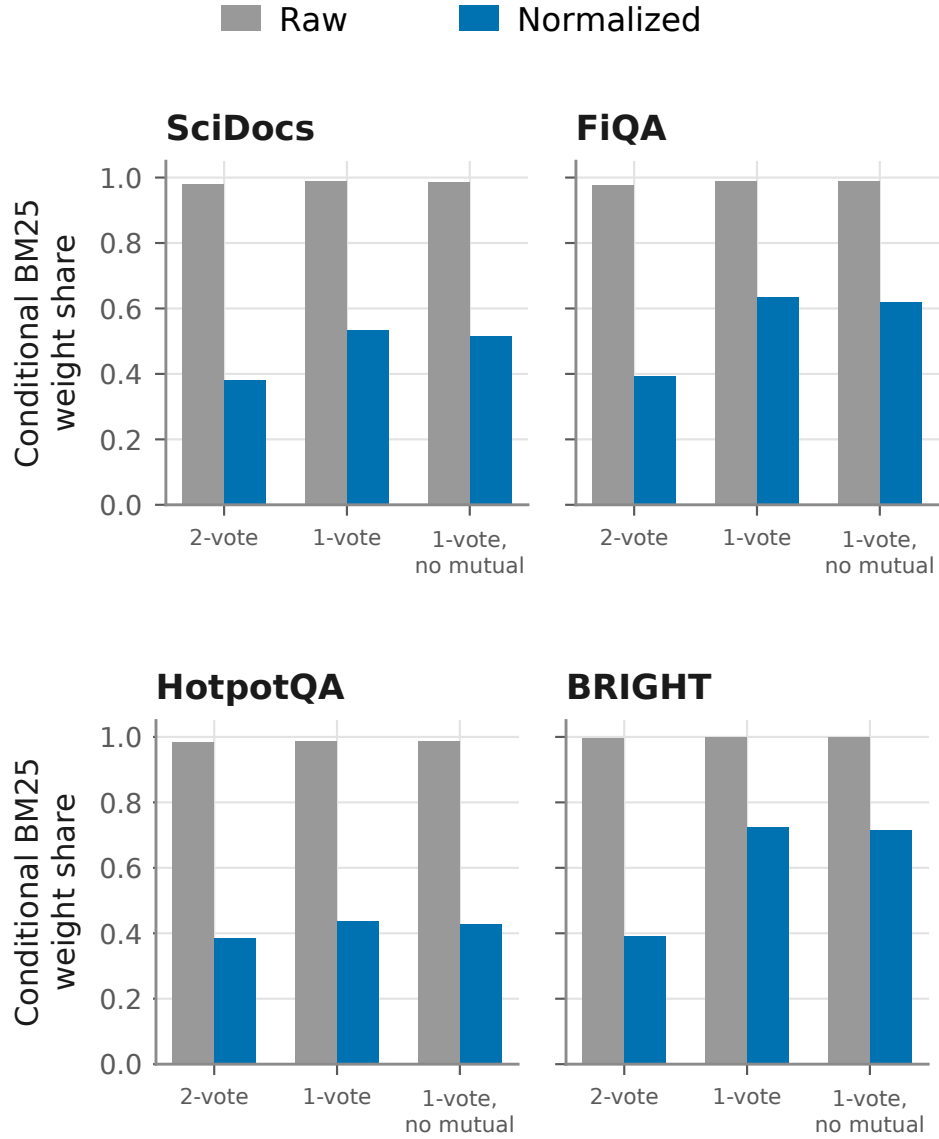


Fig. S2 Source-scale sensitivity on the fixed canonical cohorts (120 SciDocs, 120 FiQA, 52 HotpotQA and 50 BRIGHT queries). Within each dataset/regime, conditional share is the ratio of summed BM25 weight to summed total source weight over retained directed-edge support with positive BM25 contribution. Gray/blue bars use raw/normalized margins. Panels preserve the original plotted values; 2-vote, 1-vote and 1-vote no mutual denote ms2, ms1 and construction-time mutual dropping

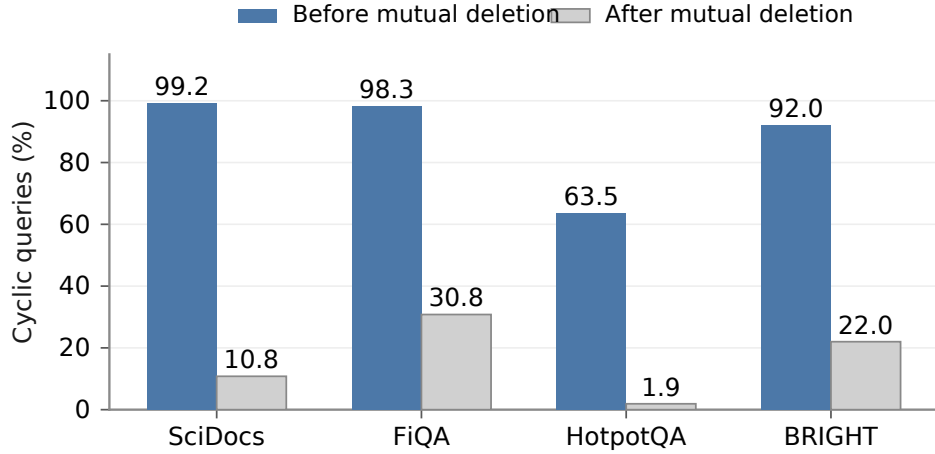


Fig. S3 Cyclic-query percentages under ms1 before and after post-hoc reciprocal-edge deletion (120 SciDocs, 120 FiQA, 52 HotpotQA and 50 BRIGHT queries). This post-hoc diagnostic is distinct from construction-time mutual dropping

S4.2 Prospective construction variants

All 200 model/query/construction graphs reuse 20 identities. Qwen’s two symmetric variants coincide, as do Smol’s strict and symmetric-both-valid graphs. Every cyclic graph has saved nonempty deletions and a repaired DAG. Tables S15 and S16 report the full structural and utility matrices; Table S17 separates domains. Figure S4 summarizes the repeated opportunity counts. No rule is selected by test relevance.

Table S15 Graph-construction sensitivity. E = retained edges; C = cyclic graphs; T = cyclic triads; SCC = maximum strongly connected component size; deletion and full-ranking change counts are G/E (greedy/exact). Q/S denote Qwen/Smol. A cyclic triad is a directed three-node cycle; a strongly connected component is a maximal mutually reachable node set. All counts cover twenty queries per row.

Model	Rule	E	C	T	SCC	Deleted	Rank changes
Q	Strict	289	0	0	1	0/0	0/0
Q	AB	728	8	37	9	33/16	8/8
Q	BA	656	8	46	12	55/18	7/6
Q	Sym. both	665	5	5	6	8/5	4/4
Q	Sym. avail.	665	5	5	6	8/5	4/4
S	Strict	144	1	1	3	1/1	1/1
S	AB	1250	13	108	12	133/38	13/12
S	BA	1253	12	212	12	82/58	11/11
S	Sym. both	144	1	1	3	1/1	1/1
S	Sym. avail.	257	5	11	10	14/8	5/5

Table S16 Exploratory sensitivity on the same 20 independent queries per model. Cyc = cyclic graphs; Opp = cyclic graphs with at least one judged-positive candidate; U/G/E = unrepaired/greedy/exact nDCG@10. Qwen’s two symmetric variants coincide. No alternative becomes a new primary protocol.

Model	Construction	Cyc	Opp	U	G	E
Qwen	Strict	0	0	0.126872	0.126872	0.126872
	AB only	8	3	0.109908	0.109908	0.109908
	BA only	8	4	0.110179	0.109416	0.110179
	Symmetric (both/available)	5	2	0.121881	0.121881	0.121881
Smol	Strict	1	0	0.097435	0.097435	0.097435
	AB only	13	6	0.105196	0.085103	0.105196
	BA only	12	4	0.055740	0.047407	0.051396
	Symmetric (both valid)	1	0	0.097435	0.097435	0.097435
	Symmetric (available)	5	2	0.088779	0.088779	0.088779

Table S17: Exploratory full-query nDCG@10 by model, construction and dataset. U/G/E/SR = unrepaired/greedy/exact/SpringRank. Opp = cyclic-positive queries. Qwen/Smol identify the models; FiQA/SciD/BR denote FiQA/SciDocs/BRIGHT, with 7/7/6 queries. All-query variant means are in Table S16.

Model	Rule/domain	Opp	U	G	E	SR
Qwen	Strict/FiQA	0	0.2267	0.2267	0.2267	0.2474
Qwen	Strict/SciD	0	0.0242	0.0242	0.0242	0.0242
Qwen	Strict/BR	0	0.1301	0.1301	0.1301	0.1301
Qwen	AB/FiQA	1	0.2267	0.2267	0.2267	0.2226
Qwen	AB/SciD	0	0.0242	0.0242	0.0242	0.0242
Qwen	AB/BR	2	0.0736	0.0736	0.0736	0.0736
Qwen	BA/FiQA	2	0.2051	0.2029	0.2051	0.1706
Qwen	BA/SciD	0	0.0187	0.0187	0.0187	0.0173
Qwen	BA/BR	2	0.1061	0.1061	0.1061	0.1061
Qwen	Sym. both/FiQA	1	0.2180	0.2180	0.2180	0.2226
Qwen	Sym. both/SciD	0	0.0187	0.0187	0.0187	0.0173
Qwen	Sym. both/BR	1	0.1301	0.1301	0.1301	0.1301
Qwen	Sym. avail./FiQA	1	0.2180	0.2180	0.2180	0.2226
Qwen	Sym. avail./SciD	0	0.0187	0.0187	0.0187	0.0173
Qwen	Sym. avail./BR	1	0.1301	0.1301	0.1301	0.1301
Smol	Strict/FiQA	0	0.1705	0.1705	0.1705	0.1374
Smol	Strict/SciD	0	0.0242	0.0242	0.0242	0.0242
Smol	Strict/BR	0	0.0976	0.0976	0.0976	0.0976
Smol	AB/FiQA	4	0.2226	0.1788	0.2226	0.1903
Smol	AB/SciD	0	0.0187	0.0187	0.0187	0.0187
Smol	AB/BR	2	0.0691	0.0532	0.0691	0.0560
Smol	BA/FiQA	2	0.0530	0.0498	0.0612	0.0530
Smol	BA/SciD	0	0.0153	0.0153	0.0153	0.0153
Smol	BA/BR	2	0.1061	0.0821	0.0821	0.1061
Smol	Sym. both/FiQA	0	0.1705	0.1705	0.1705	0.1374
Smol	Sym. both/SciD	0	0.0242	0.0242	0.0242	0.0242
Smol	Sym. both/BR	0	0.0976	0.0976	0.0976	0.0976
Smol	Sym. avail./FiQA	0	0.1491	0.1491	0.1491	0.1229
Smol	Sym. avail./SciD	0	0.0209	0.0209	0.0209	0.0209
Smol	Sym. avail./BR	2	0.0976	0.0976	0.0976	0.0736

Model-averaged AB-only greedy effect is -0.010046 , pointwise interval $[-0.027757, 0]$, raw $p=0.5$; AB-only exact is zero. BA-only greedy gives -0.004548 , interval $[-0.012177, -0.000140]$, raw $p=0.125$; BA-only exact gives -0.002172 , interval $[-0.010946, 0.004430]$, raw $p=0.75$. Both symmetric constructions have zero observed

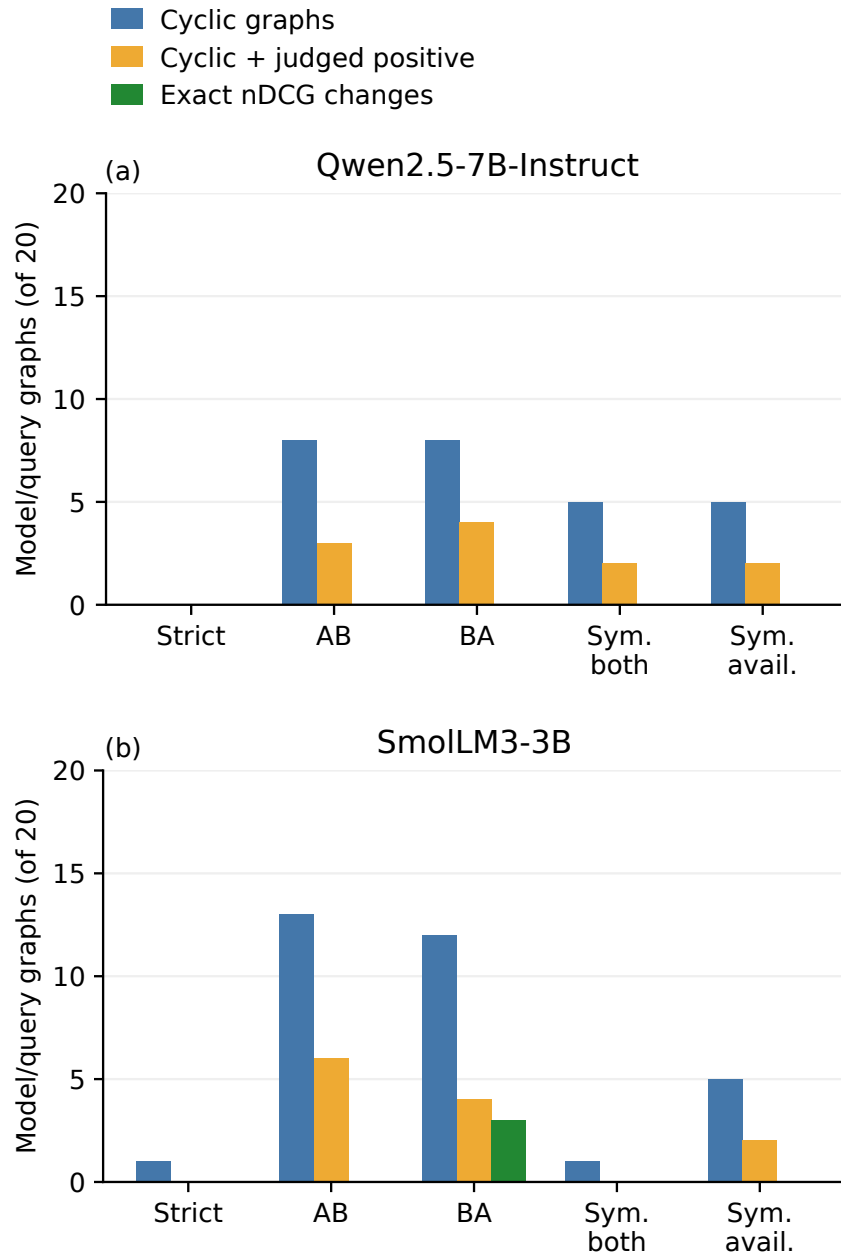


Fig. S4 Exploratory opportunity counts on the same 20 query identities across all five rules, separately for (a) Qwen2.5-7B-Instruct and (b) SmolLM3-3B. Blue counts cyclic graphs, orange cyclic graphs containing a judged-positive candidate, and green nonzero exact-minus-unrepaired full-query nDCG@10 records. Counts are descriptive repeated views; cyclic-positive eligibility alone does not imply utility improvement

repair effects. All eight Holm p-values are one. Six of nine nonzero branch records change order within a fixed top-ten set; these records cover six query identities. Smol BA-only exact improves FiQA 3705, harms BRIGHT biology 57 and FiQA 9349 and ties 17 queries. These are conditional outcome descriptions, not a selection rule.

The judged-positive subset contains seven of 20 prospective pools. Conditional strict Qwen/Smol means are 0.362491/0.278386, compared with principal all-query means 0.126872/0.097435. The 13 zero-positive pools remain in principal estimates. Neither conditioning nor treating unjudged documents as evaluation zero restores missing relevant candidates.

S5 Cached LLM Evidence and Relevance Attribution

S5.1 Primary and selected-cache contracts

The corrected primary cache uses six independent queries, three SciDocs/three FiQA, eight candidates and four historical provider identities: Azure GPT-4.1-mini, Cohere command-r-plus-08-2024, Fireworks gpt-oss-120b and Gemini2.5Flash. These serving names are historical aliases. Each provider compares 168 unordered query–document pairs in both orientations, giving 1,344 responses across the four providers. Strict parsing yields 1,334 valid directional responses, six invalid responses and four abstentions. Every individual-provider graph is a DAG; all six aggregate graphs are cyclic.

The earlier cluster analysis used a legacy parser that could default invalid output to A and resolve reversal conflicts toward the first input. It also selected the better whole-graph greedy/exact utility using qrels before averaging 20 conditions per query. Its saved difference +0.00019, cluster interval $[-0.00072, 0.00140]$, exact $p=0.875$ concerns that historical estimand, not the corrected aggregate-only replay. Subtracting it from +0.001010 would not isolate a parsing effect. Both use only six independent queries.

Table S18 retains the primary and selected-cache repair effects.

Table S18 Exploratory cached-LLM repair effects with fixed Copeland and full-query nDCG@10. Primary: six independent queries; pool variants reuse them. Selected-cache strata: selected, disjoint provider-availability groups. Values are mean repaired-minus-unrepaired differences, not effects of changing pools.

Population / graph	Queries	Greedy Δ	SCIP Δ
Primary pool 8, four-provider	6	+0.001010	+0.001010
Complete pool 6, four-provider	6	+0.001732	+0.001732
Sparse pool 8, four-provider	6	−0.003920	−0.004930
Sparse pool 10, four-provider	6	−0.006396	−0.006396
Selected cache, two-provider	110	−0.000950	+0.003164
Selected cache, four-provider	103	−0.002248	+0.004840

The selected cache contains 213 disjoint-from-primary identities, with two-provider110 and four-provider103 availability strata. Fixed Copeland exact effects are +0.003164, interval $[0.000060, 0.007053]$, raw $p=0.09449$, and +0.004840, interval

$[-0.001427, 0.014088]$, raw $p=0.25627$. Both adjusted p -values equal one in their saved exploratory family. Only 18/213 queries have complete matched source-score maps for graph-free fusion. The 217 score-derived task records on these queries are not genuine-LLM observations.

S5.2 Relevance-profile attribution and descriptive cases

The attribution uses strict-agreement, real-LLM aggregate graphs from the selected cohort, with fixed Copeland extraction, empty insertion-order seed and one baseline task per dataset/query. Each query has the same complete 15-document candidate set in the none/greedy/exact permutations. Qrels define the evaluated top-ten grade vectors; missing labels receive evaluation grade zero. The inherited 10^{-12} tolerance defines a zero metric difference. Discounted gains and nDCG are independently recomputed from the permutations, judged masks and full-query ideal grades before comparison with reference classifications.

The 426 branch records are paired observations on 213 queries. All 74 positive-pool changed-ranking/zero-nDCG cases per branch preserve the evaluated top-ten grade sequence. Greedy/exact zero-positive ranking-change counts are 69/61. Among positive-pool equal-profile cases, every moved document is unjudged in 55/58 records; positive-document movement occurs in 6/5 and explicitly zero-grade movement in 13/11. These indicators overlap and are descriptive, not additional population hypothesis tests.

Independent arithmetic checks reproduce all 426 branch records and 6,816 reference fields, with maximum difference 2.22×10^{-16} . Query-local integer relabeling preserves the permutations, and judged masks distinguish explicit zero grades from unjudged documents. This verifies attribution conditional on rankings and labels; it does not regenerate provider comparisons or original repair choices. The numerical inputs remain private pending sharing clearance.

S5.3 Dataset, availability and positive-pool composition

The composition analysis links each query’s two branch outcomes to dataset, provider availability and pre-repair cyclicity. Table S19 crosses dataset, two/four-provider availability and judged-positive availability, retaining empty cells. Both branches refer to the same 213 identities. The cohort is distinct from the canonical score-derived population; the strata are descriptive and introduce no new significance-testing family.

Across positive-containing pools ($N = 134$, 120 cyclic), greedy/exact leave 23/35 rankings unchanged, change 111/99 and change measured utility in 37/25 (15/16 gains and 22/9 harms). The other 74/74 changed rankings retain their evaluated profile. Across zero-positive pools ($N = 79$, 71 cyclic), 10/18 rankings are unchanged and 69/61 change without a measured change. There is no changed-profile/zero-metric positive-pool residual under the inherited 10^{-12} tolerance. The main composition table gives the overlapping dataset and availability marginals.

The zero-positive subset is 36 SciDocs, 38 FiQA, four BRIGHT and one HotpotQA. Positive-pool fixed-profile zeros occur in all four datasets and both strata. The

HotpotQA/two-provider/zero-positive cell is empty; HotpotQA/four-provider/zero-positive has one query and both BRIGHT zero-positive cells have two. These are sparse descriptive counts, not reliable estimates of subgroup efficacy. Availability is observational and confounded with datasets and evidence retention. Judged-positive absence is not true irrelevance.

Table S19 Complete dataset $\times$ provider-availability $\times$ judged-positive-availability decomposition. $P+$ is yes/no positive pool; C counts cyclic graphs (cyclic-positive cases in yes rows). Same/changed count full rankings; Equal counts changed rankings with identical evaluated grade profiles, including zero-positive cases. The last six columns contain paired greedy/exact counts. Empty cells have $N = 0$; they are not estimates of population behavior.

Dataset	Prov.	$P+$	N	C	Same	Changed	Equal	ΔU	Gain	Harm
SciDocs	2	no	18	15	4/6	14/12	14/12	0/0	0/0	0/0
SciDocs	2	yes	23	21	5/8	18/15	7/9	11/6	4/4	7/2
SciDocs	4	no	18	18	0/0	18/18	18/18	0/0	0/0	0/0
SciDocs	4	yes	18	18	1/1	17/17	9/14	8/3	3/3	5/0
FiQA	2	no	19	17	3/8	16/11	16/11	0/0	0/0	0/0
FiQA	2	yes	22	21	3/8	19/14	13/10	6/4	3/1	3/3
FiQA	4	no	19	19	0/1	19/18	19/18	0/0	0/0	0/0
FiQA	4	yes	16	16	0/1	16/15	14/12	2/3	2/3	0/0
BRIGHT	2	no	2	0	2/2	0/0	0/0	0/0	0/0	0/0
BRIGHT	2	yes	18	8	11/14	7/4	5/3	2/1	1/1	1/0
BRIGHT	4	no	2	2	0/0	2/2	2/2	0/0	0/0	0/0
BRIGHT	4	yes	20	20	1/0	19/20	13/15	6/5	1/2	5/3
HotpotQA	2	no	0	0	0/0	0/0	0/0	0/0	0/0	0/0
HotpotQA	2	yes	8	7	2/3	6/5	6/4	0/1	0/1	0/0
HotpotQA	4	no	1	0	1/1	0/0	0/0	0/0	0/0	0/0
HotpotQA	4	yes	9	9	0/0	9/9	7/7	2/2	1/1	1/1

The observed case C1 illustrates a greedy positive-pool fixed-profile response in which only unjudged documents move within the evaluated prefix. It is selected by minimizing the number of moved positions, followed by dataset/query identity; it is not a random or representative query. Its BRIGHT two-provider ranking exchanges positions 2/5 while preserving the evaluated profile and utility. Only this minimal derived description is shown, not benchmark passages, model responses or complete rankings.

All four symmetric-available cyclic-positive cells are shown in Table S20. They contain only three independent query identities. Each has one judged-positive candidate. Every moved document is unjudged and no judged-positive position changes. Exact top-ten membership is unchanged in all four; greedy Qwen/FiQA 7284 changes membership without utility. These are returned-solution cases, not certified complete optimum sets.

Table S20 Four descriptive cyclic-positive model/query cells on three queries. G/E denote greedy/exact; moved documents are all unjudged

Model / query	Deleted G/E	Moved G/E	Top10 set G/E
Qwen / BRIGHT biology 61	2/1	2/4	No/No
Qwen / FiQA 7284	3/1	7/2	Yes/No
Smol / BRIGHT biology 61	5/2	6/4	No/No
Smol / BRIGHT biology 57	1/1	2/3	No/No

Saved structural-utility associations are heterogeneous: exact removed-weight/effect Spearman association is $+0.3094$ for the Azure selected-cache graphs, but $-0.0441/-0.0682$ for the two aggregate strata. These are descriptive existing associations without predictive-generalization claims. Removed weight measures intervention magnitude, not necessarily better optimization. Neither a global absence of association nor a causal relationship follows.

S6 External Controls and Candidate Coverage

S6.1 Matched graph and trained relevance controls

The original-graph SpringRank comparison preserves exact candidates, directed weights and full-query qrels, including isolates. The six-query primary mean 0.882760 ties both selected repairs query by query. On 213 selected queries, means are 0.357436 SpringRank, 0.353838 unrepaired, 0.352261 greedy, 0.357813 exact and 0.352890 Bradley-Terry-style aggregation. Exact versus SpringRank has 23 wins, 12 losses and 178 ties. All 79 zero-positive pools remain; the 134 positive-containing subset has means $0.568163/0.568763$ for SpringRank/exact. These are coverage-conditioned descriptions, not improved candidate recall. Table S21 provides all saved matched SpringRank contrasts.

Table S21 Exploratory original-graph SpringRank comparison, primary $\lambda = 0$, full-query nDCG@10. U/G/E are unrepaired/greedy/exact Copeland. Positive Δ favors SpringRank; intervals are pointwise paired query-bootstrap 95% intervals. Holm covers these eight tests. Selected denotes the 213-query cache. Six-query comparisons with G/E tie on every query; displayed zeros suppress floating-point errors below 10^{-12} .

Cohort	Control	Δ	95% interval	Holm p
Six ($n = 6$)	U-Copeland	$+0.001010$	$[0, 0.003030]$	1
	G-Copeland	0	$[0, 0]$	1
	E-Copeland	0	$[0, 0]$	1
	Bradley-Terry	$+0.001010$	$[0, 0.003030]$	1
Selected ($n = 213$)	U-Copeland	$+0.003598$	$[-0.003052, 0.011223]$	1
	G-Copeland	$+0.005175$	$[-0.000692, 0.012694]$	1
	E-Copeland	-0.000377	$[-0.006577, 0.006980]$	1
	Bradley-Terry	$+0.004546$	$[-0.000582, 0.010750]$	1

The adapted PRP-Graph pilot uses frozen 20-document pools, pinned FLAN-T5-large and the published implementation’s probability-aware comparison/Swiss acquisition and personalized PageRank logic [7]. Runtime adapters call unchanged modules directly where the public command-line round handling is unsuitable; no test-qrel interpolation is tuned. Its published TREC-COVID run, original first-stage index and recorded graph are not reproduced. Local changes are documented rather than described as exact reproduction of published effectiveness.

The trained **BAAI/bge-reranker-v2-m3** cross-encoder follows its official model-card procedure [3]. It supplies exactly 400 PRP-pool and 240 prospective-pool query–document scores, with 32-bit floating-point (FP32) inference, batch size 8, joint maximum length 512, longest-first/right truncation, raw logits and document-ID ties. These 640 scores are not 640 independent queries. Training and compute differ from the generative models. Table S22 preserves the full absolute means, including PageRank.

Table S22 Matched full-query nDCG@10. Each column uses 20 independent queries. U/G/E = unrepaired/greedy/exact Copeland; SR = SpringRank; BT = Bradley–Terry. Shared BM25/BGE values in the prospective columns are one observation per query, not model replications. PRP-native graph methods use its Swiss observations; prospective methods use the strict primary graph, fixed before its repair evaluation.

Method	PRP pilot	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	—	—
Unrepaired Copeland	0.271771	0.126872	0.097435
Greedy repair	0.271771	0.126872	0.097435
SCIP exact repair	0.271526	0.126872	0.097435
SpringRank	0.259999	0.134091	0.085859
PageRank	0.232208	0.126872	0.105153
Bradley–Terry	0.254006	0.134091	0.089653

On the adapted PRP 20-query cohort, all eight original comparisons against four native Swiss/four random-acquisition controls have Holm $p=1$. BGE differences and intervals are exploratory pointwise comparisons, not an added confirmatory family. Graph-free canonical macros in Table S11 similarly provide context rather than paired superiority evidence.

S6.2 Qrels-blind candidate intervention

Twenty further queries exclude prior newer-cohort identities and exact query-text duplicates. Both arms are frozen before qrels: BM25 top12, versus BGE top12 selected from the same BM25 top100. Retrieval uses unchanged lowercase-alphanumeric BM25, $k_1 = 1.5$, $b = 0.75$, original IDs, title–newline–body text and deterministic document-ID ties. The oracle orders available candidates by qrel gain with full-query IDCg; it is a label-informed ceiling, not achieved reranker effectiveness. Table S23 separates overall and dataset-specific candidate coverage.

Table S23 Candidate intervention: frozen BM25 top-12 (A) versus BGE-selected top-12 from BM25 top-100 (B). Coverage counts pools with judged-positive documents; oracle ceilings use the full-query ideal. Both arms were frozen before qrels access.

Dataset (n)	Coverage A/B	Positive docs A/B	Oracle A	Oracle B
FiQA (7)	2/4	2/4	0.230450	0.364529
SciDocs (7)	4/3	5/5	0.224375	0.200150
BRIGHT biology (6)	0/1	0/1	0.000000	0.102191
All (20)	6/8	7/10	0.159189	0.228295

Coverage rises 6 $\rightarrow$ 8/20, difference +0.10 with interval $[-0.10, 0.30]$. Mean candidate recall rises 0.125000 $\rightarrow$ 0.183333 and oracle nDCG 0.159189 $\rightarrow$ 0.228295, difference +0.069106 with interval $[-0.016958, 0.164250]$, four improvements, one harm and 15 ties. SciDocs loses one positive-containing pool. Of 240 candidates per arm, 233 BM25/230 BGE are unjudged. BM25 top100 has positive coverage 11/20 and oracle 0.318911; that is a diagnostic ceiling, not a third tuned arm. The inspected relevance labels are not an untouched method-selection set.

S7 Measurement Validity and Prompt Development

S7.1 Prospective reliability and missing controls

The original JSON prompt includes an A/high illustrative response. It is a plausible source of positional sensitivity, not an isolated cause established by the collection. Qwen has 2,920 valid pairwise outputs with 1,378 abstentions; Smol has 2,772 valid of 2,920, with 148 invalid, all token-capped. Invalidity and capping overlap rather than add. Initial opposite-winner/abstention and repeat denominators are retained in Table S24. A repeated positional preference can be stable without being correct.

Table S24 Prospective reliability denominators, separately by model. Pairwise totals include independent repeats; initial orientation denominators exclude repeats. A/B denote the first/second displayed document; AB/BA reverse their presentation and restore actual document identity. Displayed-A rates condition on valid directional observations. Invalid and capped counts overlap; valid abstentions remain separate from invalid or missing responses.

Endpoint	Qwen	Smol
Pairwise valid / attempted	2920/2920	2772/2920
Valid abstentions	1378	0
Directional A / valid directional	747/1542	2591/2772
Initial AB/BA discordance / comparable	591/1320	1051/1195
Initial opposite-direction winners	215	1051
Initial one-abstention pairs	376	0
Initial both-abstention pairs	440	0
Initial pairs with invalid response(s)	0	125
Initial single-invalid / double-invalid pairs	0/0	113/12
Invalid pairwise outputs / capped invalid	0/0	148/148
Repeat disagreement / comparable	22/280	14/262
Repeat comparisons involving invalidity	0	18
Valid listwise / attempted	0/20	10/20

Qwen abstentions comprise 1,376 insufficient-information and two ties. Smol invalid pairwise outputs comprise 137 first observations and 11 repeats; initial pairs include 113 single-invalid and 12 double-invalid cases. The 30 invalid listwise outputs terminate normally but are non-JSON, unlike the capped pairwise failures. Missing listwise rankings are never assigned zero utility.

S7.2 Unsuccessful neutral-schema development

Thirty deterministic repeated pairs across the same 20 development queries, ten pairs per dataset, yield 240 historical controls and 240 neutral-schema responses from two models, two orientations and two independently executed replicates. Only the A/high example is replaced; other rendered text, model settings, truncation, temperature, top- p , output cap and per-observation seeds are matched. Controls are historical, not contemporaneously randomized; numerical/hardware effects are not fully excluded. Table S25 retains the descriptive prompt-validity counts.

Table S25 Prompt-validity development: descriptive pooled counts. Original/neutral prompts share exact pair provenance; neither responses nor pairs are independent queries. A/B denote first/second displayed documents and AB/BA their two orientations. A rate conditions on valid directional outputs; the neutral Smol denominator is only two. “Corroborated” requires valid AB/BA choices for the same actual document, with all sixty pair/replicate cases in the denominator. Capping and invalidity overlap; abstentions are valid responses, distinct from missing or invalid outputs.

Model	Prompt	Valid	Capped	A/directional	Corroborated
Qwen	Original	120/120	0	40/75	13/60
	Neutral	120/120	0	16/40	13/60
Smol	Original	119/120	1	114/119	5/60
	Neutral	3/120	74	1/2	0/60

Qwen abstentions rise 45→80/120. Orientation disagreement falls 36→11/60 while both-abstention rises 11→36 and corroborated directional cases remain 13/60. The primary query-averaged usable-direction difference is zero, interval $[-0.1625, 0.1750]$, Holm $p=1$. Secondary abstention difference is $+0.29375$, interval $[0.19375, 0.40000]$. Conditional displayed-A change uses 11 defined query cells and is not unconditional debiasing.

Smol neutral outputs contain 117 invalid responses: 74 capped and 43 invalid despite normal termination. Of three valid outputs, one abstains. The primary usable-direction difference is -0.0875 , interval $[-0.2000, 0]$, Holm $p=0.5$; secondary validity difference is -0.96875 , interval $[-0.99375, -0.93750]$. All neutral orientation/repeat cases involve invalidity, so no repeat-disagreement rate is estimable. Qwen repeat disagreement changes 4→2/60; Smol original is 1/59 and neutral unavailable, not zero. Two surviving neutral Smol directional responses do not validate debiasing.

Primary inference weights 20 query means equally, separately by model, with Holm over two corroboration contrasts. Secondary intervals are pointwise. The frozen code derives endpoint/model seeds from root 20261025; literal-root sensitivity leaves Qwen’s primary interval unchanged and changes Smol’s lower endpoint to -0.2125 . This disclosed seed-prose ambiguity does not replace the frozen implementation result. The unsuccessful variant neither revises primary retrieval inference nor validates a better instrument.

S8 Related-Work and Reuse Detail

The published acyclic-ranking pipeline concerns graph disagreement and ranking recovery; pruning, exact add-back, min-cut exchange and ranking-MWFAS theory remain prior work [12]. The present study evaluates deletion against external retrieval qrels. Related graph-benchmark research studies graph identity and structural optimizer comparisons. Shared mathematical methods and research infrastructure do not make those benchmark outcomes new retrieval observations or independent replications here.

PRP-Graph evaluates probability-aware graph reranking [7]; direct SpringRank estimates hierarchy on original graphs [5]. These are different interventions from fixed-extractor edge deletion. Pairwise prompting and listwise prompting have established

effectiveness evidence under their own evaluation contracts [9, 10]; the present qualified null does not refute those results. Nor does a graph-internal disagreement optimum imply external nDCG optimality.

S8.1 Version-specific public-preprint comparison

The public Research Square v1 preprint (posted 17 June 2026) already reports the construction/repair/extraction distinction and conditional retrieval effects [11]. Those are not newly discovered claims. Table S26 distinguishes reused evidence from the present explanatory increment. It is a version-specific content map, not a claim that all differences are independent new experiments.

Table S26 Contributions relative to public preprint v1. Section/table pointers in the middle column refer to the preprint. Reused observations are not independent replications.

Component	Already present in v1	Present analysis / difference
Algorithms and theory	Repair/extraction pipeline, established graph formulations	Used as controls; no new algorithm or ranking–MWFAS theorem
Benchmarks / records	Four datasets, regime names and stored-score evidence (Sections 4–5)	Reused; cohort and normalization contracts explicit; previously released resources not new datasets
Construction	Raw-margin/fixed-threshold legacy results (Section 5.2, Table 6)	Current min–max / retention-matched pipeline; version bridge incomplete; see text
Broad retrieval message	Structural gains, conditional retrieval effects (abstract, discussion)	Retained and narrowed; broad consistency/utility distinction is not new
Statistical analysis	Earlier saved comparison summaries	Query-block correction preserves effects and 35-test family; original row-independent significance interpretations superseded
Opportunity composition	No complete selected-cache branch decomposition	213 queries / 426 dependent branches locate ranking/profile/utility activity; fixed dataset, availability and positive-pool strata
Relevance attribution	No complete saved-profile accounting	74 positive-pool fixed-profile cases per repair; observed case and all gains/harms retained; identity alone not novelty
Exact-optimum choice	A returned optimum does not enumerate all optima	Complete six-graph/eight-repair envelope; one graph utility-varying; finite diagnostic only
Reproducibility	Earlier source/results provenance	Independent processed-statistic, profile and finite-envelope checks; input access and historical pins remain limited

The legacy saved publication table exactly matches the preprint’s permissive Sci-Docs/BRIGHT cyclicity summaries, 87.5%/60.0%. It is explicitly marked superseded by the later calibrated pipeline. The current primary per-query/ranker min-max and retention-matched construction reports 99.2%/92.0%. Both versions retain 120 Sci-Docs and 50 BRIGHT query counts, but equal counts do not establish identical execution inputs or candidate pools.

Importantly, the current pipeline’s raw/fixed reference already reports 97.5%/88.0%, rather than the legacy 87.5%/60.0%. Its transition to the current primary construction therefore cannot explain the whole historical difference by normalization alone. The legacy script records raw-margin voting and fixed support/margin rules; current manifests record per-query/ranker normalization and dataset/ranker quantile vote thresholds, with aggregate thresholds matching raw retained-edge counts. Calibration uses the evaluated query list without qrel tuning. These establish a pipeline/version and documented construction-contract change, not an isolated causal effect of normalization between legacy and current tables. The exact legacy execution-to-pool/input bridge is not fully reconstructed and remains a disclosed reproducibility limit. Current score/qrel hash matches do not fill that legacy gap.

The current protocol’s paired normalization/threshold ablations are the appropriate within-pipeline contrasts. The 87.5%/60.0% values are historical context, not a second measurement of the current graphs or evidence of contradictory outcomes under identical conditions. The provenance record names the saved legacy/current tables and exact source hashes. Corrected query-block inference supersedes row-independent significance interpretations while preserving its effect estimand and 35-comparison family. It does not convert reused benchmark observations into new independent evidence.

Historical selector/headroom work is context rather than a fourth independent test of impossibility. A 419-query mixed-protocol summary over 122,203 dependent rows has preserve/repair/oracle means 0.481487/0.479852/0.483994 and query-level headroom 0.002507824, saved interval [0.002041054,0.003018928]. Benefit/harm/exact-neutral proportions are 28.2/27.2/44.6%. These meta-averages are not an untouched homogeneous efficacy study. Dependent-row feature p-values do not establish query-level predictability, and the mixed headroom cannot be compared with a canonical MDE as a practical-importance bound. Executed selector failures and mixed-protocol headroom do not establish predictive impossibility; other aggregation/stopping studies address distinct interventions.

S8.2 Data access and computational coverage

The public code and derived resources do not provide every experimental input. Processed observations support all 35 corrected query-block comparisons, including 342/290 query support and zero Holm rejections. Graph checks cover 1,025 inputs (646 acyclic, 379 cyclic), saved greedy feasibility and deletion cost. Other families’ historical p-values, exact statuses and timing remain preserved-output checks rather than new reconstructions of original scoring or certificates.

A private numerical extension supports the 213-query attribution/composition and six-graph envelope. It has been validated locally but is *not cleared for distribution*;

independent reviewer access and a genuinely non-public delivery route remain unconfirmed. The extension does not regenerate source comparisons or historical model serving. Raw text, source responses, original qrels and complete ranking/graph projections are not publicly redistributed. Existing confidential-review permissions do not establish unrestricted public redistribution rights. File manifests, checksums and execution instructions are maintained in separate access documentation rather than presented as additional scientific findings.

S8.3 Identified research resources and access limits

The public project repository is <https://github.com/SoroushVahidi/consistency-aware-llm-rankin>. Code access is distinct from access to excluded data and does not clear their redistribution. The source documentation records the scientific versions used in the analyses.

The public derived judgment resource is Consistency-Aware Judgments v1 (pinned). The recorded license is CC BY-SA 4.0; attribution is Soroush Vahidi. Its four record classes and exclusions are described in Section S1.4. The related public numerical resource is ranking-pool results (pinned). Neither resource supplies the missing canonical historical permutations or the recovered private small-instance optimization inputs. Structural resources at MWFAS heuristic metrics (pinned) and MWFAS benchmark extension (pinned) belong to related structural research and are not additional retrieval observations here.

Authoritative benchmark entry points are <https://github.com/allenai/scidocs>, the BEIR FiQA snapshot at <https://huggingface.co/datasets/BeIR/fiqa>, <https://hotpotqa.github.io/>, and <https://huggingface.co/datasets/xlangai/BRIGHT>. Hosted snapshots and license labels do not grant additional rights to underlying source text. The incompletely pinned historical HotpotQA retrieval conversion remains a reproducibility limitation.

External implementations are PRP-Graph at PRP-Graph (pinned source), SpringRank at SpringRank (pinned source), and the BGE model card at BGE (pinned model card). Third-party code is obtained from its source; the SpringRank implementation is not redistributed because of the recorded license inconsistency. These addresses identify resources, not a claim of unrestricted end-to-end reconstruction.

The BEIR corpus snapshots in Table S3 have separate qrel repositories: <https://huggingface.co/datasets/BeIR/fiqa-qrels> and <https://huggingface.co/datasets/BeIR/scidocs-qrels>. Their current public revisions do not establish the versions used historically; frozen qrel hashes remain the comparison anchors. All four original canonical processed qrel files and all 12 BM25/TF-IDF/MiniLM score files remain available privately and match the canonical protocol manifests. Those manifests contain 1,026 query-regime records; the exact study’s unchanged 1,025-record roster omits its one historical empty graph. This byte-level recovery does not establish the exact historical corpus conversion or MiniLM weight/tokenizer revision, and no scorer inference is rerun. FiQA’s original challenge at <https://sites.google.com/view/fiqa/home> specifies noncommercial source-data use. The mirror’s license label does not replace that condition, and the paper does not redistribute its text or qrels.

Official checkpoint sources for Table S4 are Qwen2.5-7B-Instruct (pinned), SmolLM3-3B (pinned), and Flan-T5-large (pinned), alongside the BGE card above. The historical MiniLM source is MiniLM model card; its exact historical weight revision has not been recovered. A current model card is not evidence of an immutable historical hosted-provider deployment.

The exact-solver implementations are SCIP, <https://github.com/scipopt/scip>, and PySCIPOpt, <https://github.com/scipopt/PySCIPOpt> [1, 8]. Recorded canonical versions are SCIP 10.0.2 and PySCIPOpt 6.2.1; current upstream heads are not substituted for those runtime versions. The public PRP-Graph code is Apache 2.0. At the pinned SpringRank revision, the LICENSE file states MIT while pyproject.toml states GPL-3.0-or-later; neither that implementation nor its local runtime adapters are bundled as cleared third-party software.

The public ranking-pool resource contains 93,504 measurements, 66 configurations and 1,056 summary rows across 342 query identities. Its standard-library validator checks checksums, schemas and means; it does not reconstruct original scoring or the corrected 35-test analysis. Neither that resource nor the public judgment release supplies the private attribution and optimum-envelope inputs.

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] 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>
- [3] 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
- [4] 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>
- [5] De Bacco C, Larremore DB, Moore C (2018) A physical model for efficient ranking in networks. *Science Advances* 4(7):eaar8260. <https://doi.org/10.1126/sciadv.aar8260>
- [6] 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>

- [7] 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>
- [8] 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
- [9] 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>
- [10] 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>
- [11] 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
- [12] 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>