CHORUS OR ECHO? AN ASSOCIATION-AWARE RANKING LABEL MODEL FOR UNSUPERVISED RETRIEVAL FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregating multiple retrieval sources is a natural strategy for scientific literature retrieval, where relevance judgments are expensive to obtain at scale. When several such sources agree, however, it is unclear whether their agreement should count as stronger evidence: sources may jointly promote relevant candidates or repeat the same mistakes. Recent evidence of correlated errors among large language models shows that stronger sources do not eliminate this ambiguity. We quantify the consequences of this ambiguity in a controlled experiment: increasing agreement on irrelevant candidates reduces equal-weight fusion's NDCG@10 from 0.39 to 0.17, despite unchanged individual-source ranking quality. We propose the **Deliberative Noisy Label Model** (DNLM), a label model for fusing retrieval sources without relevance judgments. DNLM combines Plackett–Luce-based learning of source reliability with pairwise associations estimated from score deviations relative to pair-excluded anchors to determine fusion weights. Across four scientific and biomedical benchmarks, DNLM achieves the highest NDCG@10 in four of five settings and the highest MRR in all five against fixed-rule fusion, individual sources, and weak-supervision baselines. Controlled analyses show that DNLM mitigates the loss from shared irrelevant agreement, and that disagreement matters as well: discarding the associations that record where sources deviate in opposite directions lowers NDCG@10 by up to 3.4 points.

1 INTRODUCTION

Ranking is central to scientific literature retrieval, but relevance judgments are expensive to obtain at scale (Ajith et al., 2024; Wu et al., 2024). A common alternative is to aggregate programmatic retrieval sources, which score candidates without annotation, such as lexical retrievers (Robertson & Zaragoza, 2009), dense encoders (Karpukhin et al., 2020; Wang et al., 2020; 2022; Li et al., 2023), and citation-based scoring functions (Kessler, 1963; Small, 1973; Page et al., 1999). These sources capture different aspects of relevance, but may also share limitations. This raises a basic question: *When several imperfect sources agree, should their agreement count as stronger evidence?*

Agreement alone does not answer this question. Several sources may correctly promote the same relevant candidates, but they may also reinforce the same incorrect recommendations. For example, graph-based sources built on a common citation structure may repeatedly favor well-connected but non-relevant papers. Counting their recommendations as separate evidence can amplify these mistakes. Yet suppressing agreement indiscriminately can also discard useful support. The challenge is therefore to limit the amplification of mistakes without discarding useful cross-source evidence.

We quantify this effect in a controlled experiment that changes which candidates sources rank highly together, while keeping each source's individual ranking quality unchanged. With agreement on relevant candidates held low, increasing agreement on irrelevant candidates reduces equal-weight fusion's NDCG@10 from 0.3897 to 0.1663. This substantial loss occurs without any deterioration in the individual sources, showing that their standalone quality alone cannot explain the effectiveness of their combination (Table 5 in Appendix D).

Prior work highlights two roles of agreement: supporting evidence and potential redundancy. RRF (Cormack et al., 2009) and CombMNZ (Fox & Shaw, 1994) accumulate support for candidates

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

across sources, while overlap-based weighting estimates source quality from how often a source's results appear in other sources (Wu & Crestani, 2002). Recent methods instead assess source quality through retriever-specific signals or LLM judgments of top results (Kalra et al., 2025; Hsu & Tzeng, 2025). Yet MoR (Kalra et al., 2025) observes that the best retriever pair need not comprise the individually strongest retrievers, highlighting the importance of relationships beyond standalone quality. From the redundancy perspective, correlation-based weighting discounts highly correlated sources (Wu & McClean, 2005), while weak supervision models source reliability together with dependencies among labeling functions (Ratner et al., 2016; Bach et al., 2017; Su et al., 2023). UWS (Shin et al., 2022) extends weak supervision to rankings, but its estimation procedure requires a specified dependency graph and suitable conditionally independent source triplets. Identifying such triplets can be difficult when retrievers share representations, seed candidates, or graph structures. Stronger sources do not remove this concern: recent LLM studies show that higher individual accuracy need not imply less correlated errors (Kim et al., 2025). We therefore combine source reliability learned from rankings with inter-source associations estimated from continuous retrieval scores, using both to determine fusion weights without target-task relevance judgments.

We propose the **Deliberative Noisy Label Model (DNLM)**, a label model that combines source association modeling with ranking-based reliability learning. An association component measures how source scores deviate together from anchors constructed from the other sources, summarizing these relationships in a matrix $\boldsymbol{A}$. A Plackett–Luce (PL) component (Plackett, 1975) learns source reliability indices using a consensus informed by these associations. An association-aware fusion objective combines $\boldsymbol{A}$ with the reliability indices to allocate source weights, which in turn shape the consensus used for reliability learning. The resulting weights aggregate retrieval scores into continuous relevance pseudo-labels.

Our main contributions are:

1. **An association-aware ranking label model.** We introduce DNLM, integrating source reliability and source associations for ranking aggregation without relevance judgments.
2. **An empirical characterization of agreement.** We show that individual-source quality alone cannot explain fusion effectiveness, and characterize when accounting for source associations helps aggregation.
3. **Evidence for effective unsupervised fusion.** We demonstrate DNLM's retrieval effectiveness across four benchmarks and, in the seven-source pool, isolate the contribution of negative associations through a controlled intervention on the estimated matrix.

2 Problem Formulation and Preliminaries

Problem formulation For a query q, let $\mathcal{D}_q = \{d_{q,i}\}_{i=1}^{N_q}$ be its candidate document set and let $\boldsymbol{S}_q \in \mathbb{R}^{N_q \times m}$ contain the scores of m retrieval sources. Sources may include lexical retrievers, neural retrievers, and graph-derived scoring functions. These sources may be noisy and differ in which candidate documents they favor. The sources provide score vectors $\{\boldsymbol{s}_{q,j}\}_{j=1}^{m}$, where $\boldsymbol{s}_{q,j} = \boldsymbol{S}_{q,:,j}$, and their induced rankings $\{\sigma_{qj}\}_{j=1}^{m}$. We aggregate these source scores without ground-truth relevance judgments to produce pseudo-labels in the form of continuous relevance scores for query–document pairs in scientific literature retrieval. These pseudo-labels can support downstream tasks or be used directly for ranking.

Preliminaries In our framework, the Plackett–Luce (PL) model provides a ranking likelihood for learning **source reliability indices** from the rankings supplied by multiple retrieval sources. The PL model (Plackett, 1975) assigns **positive strengths** to documents and uses them to describe how likely each observed ordering is.

For query q and source j, let $\boldsymbol{a}_{qj} \in \mathbb{R}_{>0}^{N_q}$ denote the document strengths used to model the observed ranking σ_{qj}. Each entry $a_{q,j,i}$ represents the model's relative preference for document $d_{q,i}$: a larger value makes the document more likely to be ranked ahead of the remaining candidates.

PL models a ranking from top to bottom: at each position, the documents not yet ranked compete for that position according to their strengths. Writing $\sigma_{qj}(t)$ for the index of the document at position

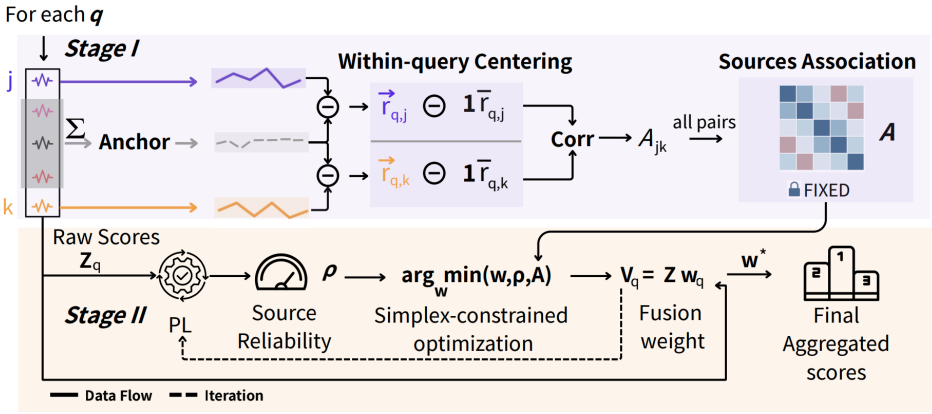


Figure 1: Overview of DNLM.

t, the conditional probability of its selection is

$$P(\sigma_{qj}(t) \mid \sigma_{qj}(1), \ldots, \sigma_{qj}(t-1); \boldsymbol{a}_{qj}) = \frac{a_{q,j,\sigma_{qj}(t)}}{\sum_{h=t}^{N_q} a_{q,j,\sigma_{qj}(h)}}. \tag{1}$$

The numerator is the strength of the document placed at position t, and the denominator is the total strength of all documents not yet ranked. The sequential factorization also yields the likelihood of a top-K prefix without enumerating the possible orderings of the remaining documents. Choosing K smaller than N_q focuses the fitting objective on the leading positions and reduces the number of log-likelihood terms evaluated per ranking. The resulting prefix likelihood is

$$P(\sigma_{qj}(1{:}K); \boldsymbol{a}_{qj}) = \prod_{t=1}^{K} \frac{a_{q,j,\sigma_{qj}(t)}}{\sum_{h=t}^{N_q} a_{q,j,\sigma_{qj}(h)}}, \qquad K \le N_q. \tag{2}$$

Let $\boldsymbol{\theta}$ denote the model parameters, with $\boldsymbol{a}_{qj}(\boldsymbol{\theta})$ giving the document strengths for query q and source j. These parameters can be estimated by minimizing the negative log-likelihood summed over the observed rankings from all sources and averaged over training queries:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \left[-\frac{1}{|\mathcal{Q}_{\mathrm{tr}}|} \sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \sum_{j=1}^{m} \sum_{t=1}^{K} \log \frac{a_{q,j,\sigma_{qj}(t)}(\boldsymbol{\theta})}{\sum_{h=t}^{N_q} a_{q,j,\sigma_{qj}(h)}(\boldsymbol{\theta})} \right]. \tag{3}$$

Here, $\mathcal{Q}_{\mathrm{tr}}$ denotes the unlabeled training query set, restricted to queries with at least K candidates.

3 DNLM

Overview The PL likelihood introduced above provides a criterion for fitting each source's ranking. However, source-wise fitting alone does not determine how overlapping evidence should be combined: sources may agree because they capture common relevance information or share systematic errors. Motivated by this ambiguity, we propose the Deliberative Noisy Label Model (DNLM), which jointly considers source reliability and source associations when aggregating retrieval sources.

The framework consists of two stages. Stage I uses auxiliary PL fitting to support the estimation of a source association matrix $\boldsymbol{A}$. Stage II reruns PL fitting with $\boldsymbol{A}$ held fixed to learn global fusion weights. These weights produce an aggregated relevance score for each query–document pair. Figure 1 illustrates the full pipeline and Algorithm 1 in Appendix C gives the full procedure.

3.1 Stage I: Source Association Estimation

Stage I uses an auxiliary PL model to learn reliability indices for combining source scores into pair-excluded anchors. Each anchor combines the remaining sources' scores, against which we measure the target pair's co-varying deviations to construct the source association matrix $\boldsymbol{A}$.

Learning reliability indices for anchor construction. We first place source scores on a common scale by normalizing each source within each query:

$$Z_{q,i,j} = \frac{S_{q,i,j} - \min_h S_{q,h,j}}{\max_h S_{q,h,j} - \min_h S_{q,h,j}}. \tag{4}$$

The resulting matrix is denoted by $\boldsymbol{Z}_q \in \mathbb{R}^{N_q \times m}$.

We fit an auxiliary PL model over the full source pool to learn the reliability indices. For each source j, a parameter b_j[1]$\in \mathbb{R}$ defines its reliability index ρ_j through

$$\varepsilon_j = \mathrm{sigmoid}(b_j), \qquad \rho_j = 1 - \varepsilon_j. \tag{5}$$

These parameters are shared across queries. At iteration r, we normalize the reliability indices, denoted as $\boldsymbol{u}$, and construct the consensus $\boldsymbol{v}_q$[2]:

$$u_j^{(r)} = \frac{\rho_j^{(r)}}{\sum_{\ell=1}^m \rho_\ell^{(r)}}, \qquad \boldsymbol{v}_q^{(r)} = \max\left\{\boldsymbol{Z}_q \boldsymbol{u}^{(r)}, \delta_v \mathbf{1}\right\}, \tag{6}$$

where $\delta_v > 0$ is a numerical floor, $\mathbf{1} \in \mathbb{R}^{N_q}$ is the all-ones vector, and the maximum is element-wise.

Holding $\boldsymbol{v}_q^{(r)}$ fixed, we construct the PL strengths for source j as

$$a_{q,j,i}(\boldsymbol{b}; \boldsymbol{v}_q^{(r)}) = \max\left\{\rho_j \frac{v_{q,i}^{(r)}}{\mathbf{1}^\top \boldsymbol{v}_q^{(r)}} + \frac{1-\rho_j}{N_q}, \delta_a\right\}. \tag{7}$$

where $\boldsymbol{b} = (b_1, \ldots, b_m)^\top$ and $\delta_a > 0$ is a numerical floor. The index ρ_j controls the balance between consensus-based and uniform strengths. We take a fixed number of gradient steps on the PL objective in Eq. (3) with respect to $\boldsymbol{b}$, then recompute the normalized reliability indices $\boldsymbol{u}$ and consensus $\boldsymbol{v}_q$ for the next iteration.

Pair-excluded anchors. After fitting, we recompute and retain $\boldsymbol{u}$ from the final reliability indices. For each source pair (j,k), we exclude its direct contribution to the anchor $\boldsymbol{v}_q^{(-jk)}$ by aggregating only the remaining sources:

$$\boldsymbol{v}_q^{(-jk)} = \max\left\{\frac{\sum_{h \notin \{j,k\}} u_h \boldsymbol{z}_{q,h}}{\sum_{h \notin \{j,k\}} u_h}, \delta_v \mathbf{1}\right\}, \tag{8}$$

where $\boldsymbol{z}_{q,h} = \boldsymbol{Z}_{q,:,h}$ is source h's normalized score vector. This construction requires at least three sources; if $\sum_{h \notin \{j,k\}} u_h$ is numerically negligible, we use their unweighted mean before applying the numerical floor. Pair exclusion removes direct score contributions, but does not imply independence, since $\boldsymbol{u}$ is learned from the full source pool.

Estimating source associations. For a fixed pair (j,k), we compare both sources against the same anchor $\boldsymbol{v}_q^{(-jk)}$. For $\ell \in \{j,k\}$, let $\boldsymbol{e}_{q,\ell}$ denote its score deviation from this anchor and $\bar{e}_{q,\ell}$ its mean deviation within query q:

$$\boldsymbol{e}_{q,\ell} = \boldsymbol{z}_{q,\ell} - \boldsymbol{v}_q^{(-jk)}, \qquad \bar{e}_{q,\ell} = \frac{\mathbf{1}^\top \boldsymbol{e}_{q,\ell}}{N_q}, \qquad \boldsymbol{x}_{q,\ell} = \boldsymbol{e}_{q,\ell} - \bar{e}_{q,\ell}\mathbf{1}. \tag{9}$$

[1]Initialize $b_j = -\frac{3}{|\mathcal{Q}_{\mathrm{tr}}|}\sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \max\{0, \mathrm{Spearman}(\boldsymbol{Z}_{q,:,j}, \boldsymbol{Z}_q \mathbf{1}/m)\}$

[2]The consensus $\boldsymbol{v}_q^{(r)}$ aggregates scores from the full source pool for PL fitting. The pair-excluded anchor $\boldsymbol{v}_q^{(-jk)}$, introduced later, aggregates only sources other than j and k to estimate their association.

Here, $\boldsymbol{x}_{q,\ell}$ is the centered deviation vector; its dependence on the pair (j, k) is omitted for readability. Removing the query-specific mean $\bar{e}_{q,\ell}$ allows us to measure whether the two sources' deviations vary together across documents, rather than reflecting only their mean offsets from the anchor.

We estimate this association by correlating the centered deviations pooled across training queries:

$$A_{j,k} = \frac{\sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \boldsymbol{x}_{q,j}^{\top} \boldsymbol{x}_{q,k}}{\sqrt{\left(\sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \|\boldsymbol{x}_{q,j}\|_2^2 \right) \left(\sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \|\boldsymbol{x}_{q,k}\|_2^2 \right)}}. \tag{10}$$

Positive and negative values indicate aligned and opposing centered deviations, respectively. We set $A_{k,j} = A_{j,k}$ and $A_{j,j} = 1$, and assign zero when the denominator vanishes. These associations concern deviations from constructed anchors, rather than ground-truth errors.

The resulting source association matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ is held fixed throughout Stage II.

3.2 Stage II: Association-Aware Fusion

Stage II reruns PL fitting with the source association matrix $\boldsymbol{A}$ estimated in Stage I held fixed. We reinitialize $\boldsymbol{b}$ and learn the corresponding reliability indices $\boldsymbol{\rho}$ afresh, reusing the notation from Stage I. The PL parameterization and fitting objective remain the same, but the consensus is now constructed using fusion weights computed from both $\boldsymbol{\rho}$ and $\boldsymbol{A}$. The final weights aggregate source scores for each query–document pair.

Computing fusion weights. At each iteration, the current reliability indices $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)^{\top}$ and the fixed association matrix $\boldsymbol{A}$ determine the fusion weights:

$$\boldsymbol{w}^*(\boldsymbol{\rho}; \boldsymbol{A}) \in \arg\min_{\boldsymbol{w} \in \Delta_{m-1}} \left\{ \frac{1}{2} \boldsymbol{w}^{\top} \boldsymbol{A} \boldsymbol{w} - \boldsymbol{\rho}^{\top} \boldsymbol{w} \right\}, \qquad \Delta_{m-1} = \left\{ \boldsymbol{w} \in \mathbb{R}^m_{\geq 0} : \mathbf{1}^{\top} \boldsymbol{w} = 1 \right\}. \tag{11}$$

The term $-\boldsymbol{\rho}^{\top} \boldsymbol{w}$ favors assigning weight to sources with higher reliability indices. The quadratic term incorporates source associations: for each pair $j < k$, its contribution is $A_{j,k} w_j w_k$. A positive $A_{j,k}$ adds a cost when both sources receive weight, whereas a negative $A_{j,k}$ reduces that cost. The association entries $A_{j,k}$ may therefore have either sign, while the fusion weights satisfy $w_j \geq 0$ and $\sum_{j=1}^m w_j = 1$.

Since $\boldsymbol{A}$ need not be positive semidefinite, the weight objective can be nonconvex. We solve this subproblem by enumerating possible sets of retrieval sources receiving nonzero weight. For each candidate subset of retrieval sources, we set the weights of all excluded sources to zero and solve for the weights of the included sources under the nonnegativity and sum-to-one constraints. Appendix C provides details of the solution procedure and its theoretical scope.

PL fitting with association-aware weights. Each iteration r consists of consensus construction followed by a PL parameter update. Starting from $\boldsymbol{b}^{(r)}$, we compute $\boldsymbol{\rho}^{(r)}$ using Eq. (5) and obtain $\boldsymbol{w}^{(r)} = \boldsymbol{w}^*(\boldsymbol{\rho}^{(r)}; \boldsymbol{A})$. These weights construct the consensus for this iteration:

$$\boldsymbol{v}_q^{(r)} = \max \left\{ \boldsymbol{Z}_q \boldsymbol{w}^{(r)}, \delta_v \mathbf{1} \right\}. \tag{12}$$

We substitute $\boldsymbol{v}_q^{(r)}$ into Eq. (7) and treat it as a constant during the PL update within iteration r. Starting from $\boldsymbol{b}^{(r)}$, a fixed number of gradient steps on Eq. (3) produces $\boldsymbol{b}^{(r+1)}$. At the next iteration, we recompute the reliability indices, fusion weights, and consensus from these updated parameters.

Producing aggregated scores. After fitting, we solve Eq. (11) using the final reliability indices $\boldsymbol{\rho}^{\mathrm{final}}$ to obtain $\boldsymbol{w}^{\mathrm{final}}$. For a new query q, we normalize its source scores using Eq. (4) and compute the aggregated relevance score

$$f(q, d_{q,i}) = \sum_{j=1}^{m} w_j^{\mathrm{final}} Z_{q,i,j}. \tag{13}$$

The weights are shared across queries, and candidates are ranked in descending order of $f(q, d_{q,i})$.

Table 1: G5 results on shared candidate pools. N@10 denotes NDCG@10. LitSearch and TREC-COVID use pooled five-fold predictions; STaRK-MAG and STaRK-Prime use official splits. Prime synthetic/human evaluations share one fitted model. UWS reports column-wise maxima over three variants.

	LitSearch		STaRK-MAG		STaRK-Prime				TREC-COVID	
					Synthetic		Human			
Method	N@10	MRR	N@10	MRR	N@10	MRR	N@10	MRR	N@10	MRR
Fixed-rule fusion										
Equal	.3518	.2905	.2729	.2774	.1662	.1516	.1843	.1688	<u>.6029</u>	<u>.8600</u>
RRF	.3411	.2768	.2831	.2924	**.1951**	<u>.1912</u>	.2067	.1995	.5539	.8529
CombMNZ	.3190	.2510	.2591	.2547	.1797	.1654	.1880	.1688	.5937	.8417
Majority Vote	.3356	.2785	.2758	.2851	<u>.1924</u>	.1836	.2118	.2059	.5599	.8533
Individual sources										
BM25	<u>.3933</u>	<u>.3613</u>	<u>.3105</u>	<u>.3451</u>	.1596	.1708	<u>.2394</u>	<u>.2573</u>	.5927	.8279
MiniLM	.3540	.3211	.2710	.3082	.1300	.1404	.1562	.1718	.4721	.7275
Graph 1†	.1933	.1733	.1208	.1312	.1198	.1158	.0821	.0767	.4054	.7577
Graph 2†	.1510	.1345	.1329	.1576	.1164	.1086	.1174	.1131	.3864	.7326
Local PPR†	.2681	.1889	.2494	.2544	.1419	.1316	.1506	.1237	.5622	.7515
Weak-supervision models										
UWS (best)	.3300	.2636	.2532	.2497	.1572	.1418	.1599	.1498	.5964	.8367
DNLM (ours)	**.4671**	**.4381**	**.3666**	**.4032**	.1858	**.1950**	**.2732**	**.2776**	**.6684**	**.8717**
External reference										
GTE-large-v1.5	.5093	.4765	.3684	.4086	.1854	.1913	.2273	.2415	.7716	.9300

Bold/underline mark the best/second-best point estimates in the first three blocks, not statistical significance. GTE is outside G5. UWS variants: OPT-Medoid, OPT-Score, and TauB-Score. †Graph 1/2: bibliographic coupling/co-citation; on Prime, typed one-/two-hop propagation. Local PPR uses the corresponding graph.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on four benchmarks with text and graphs: LitSearch (Ajith et al., 2024) and STaRK-MAG (Wu et al., 2024) for scientific literature retrieval, STaRK-Prime (Wu et al., 2024) for biomedical entity retrieval, and TREC-COVID (Voorhees et al., 2020) for COVID-19 literature retrieval. Dataset statistics, graph coverage, and preprocessing details are provided in Appendix A.

Retrieval sources. Text sources are BM25, all-MiniLM-L6-v2, all-mpnet-base-v2 (Song et al., 2020), and E5-large-v2. LitSearch, STaRK-MAG, and TREC-COVID additionally use bibliographic coupling, co-citation, and local personalized PageRank on citation graphs; STaRK-Prime uses relation-type-balanced one-/two-hop propagation and local PageRank on its knowledge graph. All graph sources use the deduplicated union of BM25's and MiniLM's top-five results as query-specific seeds, without relevance judgments. Construction details appear in Appendix A.

Source pools. Our main configuration, G5, combines BM25, MiniLM, and the three dataset-specific graph sources. H7 adds MPNet and E5 to examine the effect of expanding the source pool. Scores are min–max normalized within each query.

Baselines. We compare individual sources, Equal fusion, RRF ($k = 60$), CombMNZ, top-100 majority voting, and three UWS adaptations (Shin et al., 2022); the main table reports UWS's per-column maximum. GTE-large-v1.5 is an external reference outside G5/H7. Definitions, tie handling, and adaptation details appear in Appendix B.

Implementation and evaluation. DNLM is fitted without relevance labels. Each stage uses 20 alternating iterations with five PL updates per iteration, a top-K likelihood with $K = 20$, and Adam with learning rate 0.01. We use the same configuration across datasets and source pools,

with seed 42. We report NDCG@10 and MRR, using five-fold cross-validation on LitSearch and TREC-COVID and official splits on STaRK-MAG and STaRK-Prime. Prime's synthetic and human-written queries are evaluated separately using the same fitted model. Detailed evaluation protocols and additional metrics are provided in Appendix A.

4.2 Main Results

Main results. Table 1 reports G5 performance on shared candidate pools. DNLM achieves the highest NDCG@10 in four of five settings and the highest MRR in all five. Its NDCG@10 gains over the strongest baseline are 7.38 percentage points on LitSearch, 5.61 on STaRK-MAG, 3.38 on STaRK-Prime human, and 6.55 on TREC-COVID. On STaRK-Prime synthetic, RRF leads in NDCG@10 (0.1951 versus 0.1858), while DNLM leads in MRR (0.1950 versus 0.1912).

DNLM also outperforms every individual G5 source on both metrics. Equal fusion and RRF, however, fall below BM25 on LitSearch, STaRK-MAG, and STaRK-Prime human, showing that combining sources alone does not ensure better retrieval. Against the external GTE-large-v1.5 reference, DNLM obtains close point estimates on STaRK-MAG and STaRK-Prime synthetic and exceeds GTE on Prime human, while GTE leads on LitSearch and TREC-COVID.

Effect of expanding the source pool. Adding MPNet and E5-large to form H7 improves DNLM's NDCG@10 by 1.90–5.27 percentage points across all five settings (Table 10 in Appendix F). DNLM remains the strongest fusion method in four of five settings, with RRF slightly ahead on STaRK-Prime synthetic. Thus, DNLM benefits from the additional neural sources, although its gains are not consistently larger than those of other fusion methods.

4.3 Ablation and Analysis

Component analysis. Table 2 starts with PL-only, which fuses scores using normalized PL-based reliability estimates. The second row adds a fixed association matrix A, estimated using a reference constructed from all sources, and the association-aware second-stage fusion. The final row changes only the reference construction by excluding the evaluated source pair, yielding the full DNLM. Both additions improve mean NDCG@10 on the two datasets, with pair exclusion providing further gains beyond the full-source reference.

Table 2: G5 ablation: mean NDCG@10 $\pm$ sample SD over five folds; STaRK-MAG uses its training split.

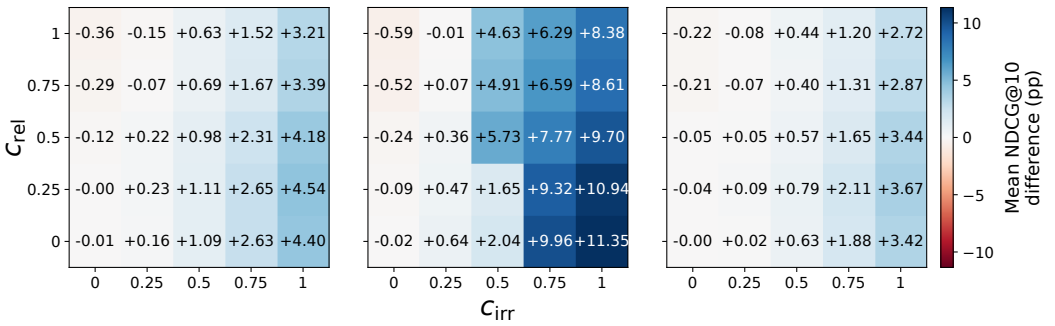
Component	LitSearch	STaRK-MAG
PL-only	$.358 \pm .029$	$.285 \pm .003$
$+ \boldsymbol{A}$	$.430 \pm .030$	$.349 \pm .003$
+ Pair exclusion	$\mathbf{.467} \pm .026$	$\mathbf{.368} \pm .008$

Alignment with individual-source quality. We first examine whether sources with higher standalone retrieval quality receive greater fusion weight. For each dataset and source pool, we measure single-source quality using NDCG@10 and compute its Spearman correlation with DNLM's final fusion weights w_j and PL-only's reliability indices ρ_j^{PL}.

Table 3: Spearman correlation of held-out single-source NDCG@10 with DNLM weights w_j or PL-only indices ρ_j^{PL}. Δ is the former minus the latter.

	G5			H7		
Dataset	w_j	ρ_j^{PL}	Δ	w_j	ρ_j^{PL}	Δ
LitSearch	.821	.700	+.121	.721	.429	+.292
STaRK-MAG	.894	.600	+.294	.679	.571	+.107
Prime (synthetic)	.671	.800	-.129	.739	.214	+.524
Prime (human)	.894	1.000	-.106	.883	.321	+.561
TREC-COVID	.671	.700	-.029	.556	.357	+.199

Table 3 shows positive alignment between w_j and individual-source quality in every setting. This alignment exceeds that of the ρ_j^{PL} in all H7 settings, while the comparison is mixed in G5: on STaRK-Prime, even the strongest G5 source, BM25, achieves only 0.1596 and 0.2394 NDCG@10 on the synthetic and human splits, respectively. The lower alignment may reflect limited useful cross-source evidence in this weak-source setting. Beyond standalone source quality, fusion also depends on which candidates different sources recommend together: sources may agree in ranking relevant candidates highly, or repeat the same mistakes by ranking irrelevant candidates highly.



(a) DNLM vs. PL-only. (b) DNLM vs. $\boldsymbol{A} = \boldsymbol{I}$. (c) DNLM vs. Equal.

Figure 2: Mean NDCG@10 differences (percentage points) between DNLM and each control over ten data seeds. The agreement controls c_{rel} and c_{irr} vary independently while each source's ranked relevance-grade sequence is held fixed. All panels share the color scale shown on the right.

Agreement on relevant versus irrelevant candidates. We vary agreement within three of five synthetic ranking sources. The controls $c_{\text{rel}}, c_{\text{irr}} \in \{0, 0.25, 0.5, 0.75, 1\}$ align rankings of relevant and irrelevant candidates, respectively, from independent to fully shared assignments. Each source's ranked relevance-grade sequence and standalone retrieval metrics (e.g., NDCG@10) remain fixed across all 25 conditions.

Figure 2 shows that DNLM's mean gains over all three controls increase with c_{irr} at every fixed c_{rel}, without requiring reduced relevant agreement. When agreement is high only on irrelevant candidates ($c_{\text{rel}} = 0$ and $c_{\text{irr}} = 1$), DNLM improves over PL-only, the identity-association control, and Equal by 4.40, 11.35, and 3.42 percentage points, respectively. When agreement is high only on relevant candidates ($c_{\text{rel}} = 1,\ c_{\text{irr}} = 0$), DNLM incurs modest losses of 0.36, 0.59, and 0.22 points, respectively. Together, these results show that association-aware weighting can mitigate shared mistakes, with a modest cost at the opposite extreme of exclusively relevant agreement. Appendix D provides the construction and weight diagnostics.

Contribution of negative source associations. We now examine disagreement as reflected in negative entries of $\boldsymbol{A}$, testing whether these associations contribute to fusion.

We zero all negative off-diagonal entries of $\boldsymbol{A}$, retain the remaining entries, and rerun Stage II with the same initialization and training configuration. This reduces NDCG@10 by 0.84–3.40 percentage points across all five H7 settings. In G5, retaining negative associations improves NDCG@10 by 0.66 points on LitSearch, with changes elsewhere within 0.05 points. These smaller effects coincide with limited weight changes: in several G5 settings, BM25 and MiniLM receive most of the fusion weight, and their respective weights change little under the intervention. H7 exhibits larger reallocations. These results support using negative alongside positive associations to guide fusion. Appendix E reports full results, weight diagnostics, and the intervention protocol.

Computational cost. Table 11 in Appendix F separates source preparation, DNLM fitting, and inference on LitSearch G5. Stage I, including anchor learning and source-association estimation, takes 8.19 s, while Stage II takes 8.02 s. The association matrix is fixed during inference. Given the shared candidate pool, DNLM takes 17.63 ms per query, including query encoding, source scoring, fusion, and sorting, compared with 27.99 ms for GTE-large-v1.5.

Sensitivity to training budgets and PL truncation. On LitSearch, varying the shared iteration budget from 1 to 40 yields NDCG@10 between 0.4637 and 0.4689, with limited gains beyond the default 20 iterations per stage. Performance also varies little across $K \in \{5, 10, 20, 50, 100, 200\}$, ranging from 0.4652 to 0.4672 (Appendix F , Figure 3).

5 Related Work

Weak supervision and data programming. Data programming (Ratner et al., 2016) and Snorkel (Ratner et al., 2017) aggregate noisy labeling functions into training labels. Extensions learn dependencies among labeling functions (Bach et al., 2017), including LLM-prompted ones (Su et al., 2023), and support continuous-valued outputs (Chatterjee et al., 2020). For ranking, Xu et al. (2019) apply data programming to passage ranking, while UWS (Shin et al., 2022) extends weak supervision to structured outputs, including permutations through a heterogeneous Mallows model. UWS estimates source accuracy using conditionally independent triplets selected according to a specified dependency graph. DNLM instead estimates source associations from retrieval scores without a supplied dependency graph.

Probabilistic ranking models. The Plackett–Luce model (Plackett, 1975) describes rankings through sequential choices based on item strengths, whereas Mallows models (Mallows, 1957) describe deviations from a central ranking. PL mixtures (Nguyen & Zhang, 2023) represent heterogeneous preference patterns. Source expertise has also been modeled explicitly: Klementiev et al. (2008) learn ranker-specific parameters in an extended Mallows model for unsupervised aggregation. DNLM uses the PL likelihood to learn source reliability indices that inform fusion weights.

Signal fusion and inter-source structure. CombMNZ (Fox & Shaw, 1994) and RRF (Cormack et al., 2009) accumulate support for candidates across sources. Early work hypothesizes that fusion benefits from greater overlap among relevant than non-relevant results (Lee, 1997). Weighted methods estimate source quality from cross-source overlap (Wu & Crestani, 2002), discount correlated sources (Wu & McClean, 2005), tune convex weights on labeled queries (Bruch et al., 2023), or assign query-specific weights using retriever-specific signals or LLM judgments of top results (Kalra et al., 2025; Hsu & Tzeng, 2025). Vote-in-Context (Eltahir et al., 2025) encodes cross-list consensus in a VLM prompt without explicitly estimating source-dependence parameters. Correlated errors have also been documented among LLMs (Kim et al., 2025). Outside retrieval, aggregation methods model shared annotator confusions (Chu et al., 2021), instance-dependent labeling-function accuracy (Zhang et al., 2023), or correlations among LLM answers (Ai et al., 2026). DNLM combines pairwise score-deviation associations estimated using pair-excluded anchors with PL-based reliability indices to allocate fusion weights.

Scientific document retrieval. SPECTER (Cohan et al., 2020) and SciNCL (Ostendorff et al., 2022) use citation information to learn scientific document representations. Weakly supervised neural ranking can use an unsupervised retriever such as BM25 to generate training signals (Dehghani et al., 2017). DNLM complements these approaches by aggregating lexical, neural, and graph-derived scores without target-task relevance judgments or retraining the underlying retrievers.

6 Conclusion and Discussion

Whether agreement among retrieval sources should count as stronger evidence depends on what the sources agree on. DNLM addresses this problem by estimating source associations against pair-excluded anchors without relevance judgments and combining them with PL-based reliability learning in an association-aware weighting objective. In the main G5 experiments, DNLM achieves the highest NDCG@10 among the evaluated in-pool baselines in four of five evaluation settings and the highest MRR in all five.

The controlled analyses clarify why source relationships matter. Individual-source quality alone is insufficient to explain fusion effectiveness: increasing shared irrelevant recommendations can substantially reduce performance even when every source's quality remains unchanged. Negative associations also contribute to fusion: removing them lowers NDCG@10 by up to 3.4 points in the seven-source pool.

Two limitations remain. First, the association matrix measures co-deviation relative to anchors rather than correctness, so association-aware weighting can also suppress useful agreement. Second, DNLM learns a single weight vector per source pool, and its current solver enumerates candidate supports, whose number grows exponentially with the number of sources. Query- and candidate-dependent weighting, together with a more scalable solver, is a natural extension.

AI USE STATEMENT

In this work, we used generative AI tools for early-stage brainstorming on how to construct the controlled experiments (Section 4.3 and Appendices D–E) and for assistance in writing and optimizing parts of the experimental pipeline and evaluation code. We did not use generative AI tools for any other task in the required-disclosure list of the ICLR 2027 AI Policy. Additionally, we used generative AI tools to polish and edit the writing and to suggest document structure and formatting.

We have reviewed all AI-assisted work: AI-assisted code was tested by the authors and all reported results were produced with the authors' own runs under the fixed configuration of Section 4.1; and all drafted or edited text was reviewed by the authors for consistency with the reported results. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify DNLM's two-stage fitting procedure and the association-aware fusion objective, and Section 4.1 lists the fixed configuration (iteration budgets, PL prefix length, optimizer, seed) used in all experiments. Appendix A documents the datasets, retrieval sources, candidate-pool construction, training splits, and evaluation protocols; Appendix B specifies the baseline implementations and tie handling; Appendix C describes the fusion solver and its verification checks; Appendices D and E give the construction of the controlled agreement grid and the negative-association intervention; Appendix F reports the sensitivity analyses. Main experiments use the fixed configuration in Section 4.1 without per-dataset tuning. Sensitivity analyses vary only the stated parameters; controlled experiments use ten data seeds with the algorithm seed fixed at 42.

REFERENCES

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: LLM aggregation by leveraging higher-order information. *ICML*, 2026.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. *EMNLP*, 2024.

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. *ICML*, 2017.

Sebastian Bruch, Siyu Gai, and Amir Ingber. An analysis of fusion functions for hybrid retrieval. *ACM Trans. Inf. Syst.*, 2023.

Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Robust data programming with precision-guided labeling functions. *AAAI*, 2020.

Zhendong Chu, Jing Ma, and Hongning Wang. Learning from crowds by modeling common confusions. *AAAI*, 2021.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. *ACL*, 2020.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *SIGIR*, 2009.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. Neural ranking models with weak supervision. *SIGIR*, 2017.

Mohamed Eltahir, Ali Habibullah, Lama Ayash, Tanveer Hussain, and Naeemullah Khan. Vote-in-context: Turning VLMs into zero-shot rank fusers. *arXiv*, 2025.

Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. *TREC-2*, 1994.

Hsin-Ling Hsu and Jengnan Tzeng. DAT: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation. *arXiv*, 2025.

Jushaan Singh Kalra, Xinran Zhao, To Eun Kim, Fengyu Cai, Fernando Diaz, and Tongshuang Wu. MoR: Better handling diverse queries with a mixture of sparse, dense, and human retrievers. *EMNLP*, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *EMNLP*, 2020.

M. M. Kessler. Bibliographic coupling between scientific papers. *Am. Doc.*, 1963.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *ICML*, 2025.

Alexandre Klementiev, Dan Roth, and Kevin Small. Unsupervised rank aggregation with distance-based models. *ICML*, 2008.

Joon Ho Lee. Analyses of multiple evidence combination. *SIGIR*, 1997.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv*, 2023.

C. L. Mallows. Non-null ranking models. I. *Biometrika*, 1957.

Duc Nguyen and Anderson Y. Zhang. Efficient and accurate learning of mixtures of Plackett–Luce models. *AAAI*, 2023.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. *EMNLP*, 2022.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Robin L. Plackett. The analysis of permutations. *Applied Statistics*, 1975.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *NeurIPS*, 2016.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *VLDB Endow*, 2017.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 2009.

Changho Shin, Winfred Li, Harit Vishwakarma, Nicholas Carl Roberts, and Frederic Sala. Universalizing weak supervision. *ICLR*, 2022.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.*, 1973.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: masked and permuted pre-training for language understanding. *NeurIPS*, 2020.

Jinyan Su, Peilin Yu, Jieyu Zhang, and Stephen H. Bach. Leveraging large language models for structure learning in prompted weak supervision. *IEEE BigData*, 2023.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 2020.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv*, 2022.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 2020.

Shengli Wu and Fabio Crestani. Data fusion with estimated weights. *CIKM*, 2002.

Shengli Wu and Sally McClean. Data fusion with correlation weights. *ECIR*, 2005.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N. Ioannidis, Karthik Subbian, James Zou, and Jure Leskovec. STaRK: Benchmarking LLM retrieval on textual and relational knowledge bases. *NeurIPS*, 2024.

Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Passage ranking with weak supervision. *ICLR Workshop on Learning from Limited Labeled Data*, 2019.

Jieyu Zhang, Linxin Song, and Alexander Ratner. Leveraging instance features for label aggregation in programmatic weak supervision. *AISTATS*, 2023.

APPENDIX

A Datasets, Retrieval Sources, and Evaluation **13**

A.1 Datasets and partitions 13

A.2 Text retrieval sources 13

A.3 Graph retrieval sources 13

A.4 Candidate construction and normalization 14

A.5 Training and reporting 14

B Baseline Definitions and Implementation **15**

B.1 Shared inputs 15

B.2 Fixed-rule fusion 15

B.3 UWS retrieval adaptations 15

C Solving the Association-Aware Fusion Subproblem and DNLM's Algorithm **17**

C.1 Solving the Association-Aware Fusion Subproblem 17

C.2 DNLM's Algorithm 19

D Controlled Agreement at Fixed Per-Source Ranking Quality **19**

D.1 Experimental construction 19

D.2 Training and evaluation 20

D.3 Full-grid results 21

D.4 Joint variation of the two agreement controls 21

D.5 Realized agreement and fusion weights 22

E Contribution of Negative Source Associations **22**

F Additional Experiments and Analyses **24**

A Datasets, Retrieval Sources, and Evaluation

A.1 Datasets and partitions

Table 4 summarizes the collections and evaluation partitions used in our main experiments. LitSearch (Ajith et al., 2024), STaRK-MAG and STaRK-Prime (Wu et al., 2024), and TREC-COVID (Voorhees et al., 2020) cover scientific literature and biomedical retrieval.

For LitSearch and TREC-COVID, we group queries by their case-folded, whitespace-normalized text and construct five folds with seed 42. Each query is evaluated once using a model fitted on the other folds. For STaRK-MAG and STaRK-Prime, models are fitted on 7,993 and 6,162 official training queries, respectively, without using their relevance labels. Training queries whose normalized text overlaps with held-out partitions are excluded. The same fitted Prime model is evaluated separately on the synthetic and human-written query sets.

The STaRK-MAG component ablation uses a separate protocol: five-fold cross-validation within its 7,993 official training queries. The official test set is not included in this cross-validation. Its scores therefore need not match the official-test results in the main table.

Table 4: Dataset sizes and main evaluation protocols. STaRK-Prime lists synthetic and human-written evaluation queries separately. TREC-COVID excludes one empty-text document from retrieval while retaining its original relevance judgments.

Dataset	Searchable items	Evaluation queries	Main protocol
LitSearch	57,986	597	Five-fold CV
STaRK-MAG	700,244	2,665	Official split
STaRK-Prime	129,375	2,801 / 98	Official split
TREC-COVID	171,331	50	Five-fold CV

A.2 Text retrieval sources

We concatenate each item's available title and textual content for retrieval. BM25 uses the Lucene-style implementation in `bm25s`, with $k_1 = 1.5$ and $b = 0.75$, lowercasing, English stopword removal, and no stemming.

The dense sources use the following pretrained checkpoints:

- `sentence-transformers/all-MiniLM-L6-v2`;
- `sentence-transformers/all-mpnet-base-v2`;
- `intfloat/e5-large-v2`.

We use attention-mask-aware mean pooling, followed by ℓ_2 normalization, and compute cosine similarity. Maximum input lengths are 256, 384, and 512 tokens, respectively. E5 inputs use the prefixes `query:` and `passage:`; the other encoders use no prefix. The encoders are not fine-tuned on benchmark relevance judgments.

A.3 Graph retrieval sources

Query-dependent seeds. For each query, let $\mathcal{S}_q$ be the deduplicated union of the top-five results from BM25 and MiniLM. These retrieved items serve as graph seeds, not as known relevant items. The graph sources consequently depend on the text retrievers through seed selection.

Citation-based sources. For LitSearch, STaRK-MAG, and TREC-COVID, let $\mathcal{N}^+(d)$ and $\mathcal{N}^-(d)$ denote the outgoing and incoming citation neighbors of document d. We compute bibliographic-

coupling and co-citation scores as

$$s_{\mathrm{bib}}(q,d) = \left| \mathcal{N}^{+}(d) \cap \bigcup_{u \in \mathcal{S}_q} \mathcal{N}^{+}(u) \right|, \tag{14}$$

$$s_{\mathrm{cocite}}(q,d) = \left| \mathcal{N}^{-}(d) \cap \bigcup_{u \in \mathcal{S}_q} \mathcal{N}^{-}(u) \right|. \tag{15}$$

These definitions count distinct shared neighbors of the seed set, rather than summing pairwise similarities to individual seeds.

For local personalized PageRank, we expand from the seeds using incoming and outgoing edges to construct a local subgraph with at most 5,000 nodes. The random walk follows outgoing citation edges within this subgraph, with damping factor 0.85 and uniform restart probability over the seeds. Dangling-node mass is redistributed to the seed distribution. Iteration stops when the ℓ_1 change is below 10^{-8}, with a maximum of 200 iterations. Candidates outside the local subgraph receive zero PageRank scores.

Typed graph sources on STaRK-Prime. We construct an undirected adjacency matrix for each relation type, remove self-loops, and normalize its nonempty rows. The transition matrix P averages these normalized transitions over the relation types available at each node. This balances relation types locally rather than weighting them by their raw edge counts.

Let $\mathbf{t}_q$ be the uniform row distribution over $\mathcal{S}_q$. The one-hop and two-hop scores are $(\mathbf{t}_q P)_d$ and $(\mathbf{t}_q P^2)_d$. For local PageRank, we retain all candidates and add high-scoring one-hop neighbors up to a total of 5,000 nodes. We restrict and renormalize P on this subgraph and use the same damping and convergence settings as above. These sources are propagation scores on a typed knowledge graph, not bibliographic-coupling or co-citation scores.

Graph coverage. TREC-COVID citations are reconstructed from the historical CORD-19 document parses dated July 16, 2020, using explicit identifiers to link references to corpus documents. The resulting corpus-induced graph contains 190,634 distinct directed edges over 171,332 corpus records. Parse files are available for 76,483 records, and 28,371 records have at least one resolved outgoing internal citation.

A.4 Candidate construction and normalization

For each query, we construct $\mathcal{C}_q$ by taking the union of the top-1,000 results from each of the four text sources, giving at most 4,000 candidates. Every text source scores every item in this union, including items outside its own retrieved top-1,000. Graph sources score the same candidates without adding new candidates to the evaluated pool.

All methods and source-pool configurations share $\mathcal{C}_q$. In particular, G5 uses only BM25, MiniLM, and the three graph scores for fusion, but its candidate pool is still constructed from all four text retrievers. This controls candidate recall when comparing G5 and H7. Relevance judgments are not used to construct candidates or source scores.

Source scores are min–max normalized separately within each query. If a source's score range is at most 10^{-8}, its normalized column is set to zero. Finite zero scores remain part of the input rather than being treated as missing observations. Nonfinite scoring outputs are rejected. For retrieval and final evaluation, exact score ties are resolved by ascending string-valued item identifier.

A.5 Training and reporting

Both DNLM stages are fitted using only the training queries' source scores and rankings. Each stage uses 20 outer iterations, five PL gradient updates per iteration, truncation depth $K = 20$, and Adam with learning rate 0.01. The association matrix is estimated after Stage I and held fixed throughout Stage II. The same configuration is used across datasets and source pools.

We evaluate using `pytrec_eval`, with NDCG@10 and MRR as the main metrics. NDCG uses the complete relevance judgments, rather than restricting its ideal ranking to retrieved candidates. Queries with no relevant candidates are retained.

Main cross-validation results pool the held-out per-query metrics across folds and average over queries. Official-split results average over their evaluation queries. In the component-ablation table, we instead report the arithmetic mean of the five fold scores and their sample standard deviation. This fold standard deviation is descriptive and is not a confidence interval or a significance test.

B Baseline Definitions and Implementation

B.1 Shared inputs

All baselines use the same candidate sets, source pools, and evaluation partitions as DNLM. Let $z_{qj}(d)$ denote the min–max-normalized score of source j for candidate d, and let $r_{qj}(d)$ be its one-based rank. Complete source rankings are obtained by sorting raw scores in descending order, with ties resolved by ascending string-valued item identifier.

Individual-source results rank candidates by that source's score on the shared pool. The best-single-source result is selected retrospectively by its aggregate NDCG@10 within the evaluated source pool.

B.2 Fixed-rule fusion

Equal fusion. Equal fusion averages normalized scores:

$$f_{\mathrm{Equal}}(q,d) = \frac{1}{M}\sum_{j=1}^{M} z_{qj}(d). \tag{16}$$

A constant source contributes a zero normalized column.

Reciprocal rank fusion. RRF uses $k = 60$:

$$f_{\mathrm{RRF}}(q,d) = \sum_{j\in\mathcal{J}_q} \frac{1}{60 + r_{qj}(d)}, \tag{17}$$

where $\mathcal{J}_q$ contains sources whose raw score range exceeds 10^{-8} for query q. Excluding constant sources prevents arbitrary tie completion from contributing artificial rank differences. Ranks are computed over the complete shared candidate pool, not only each source's original top-1,000 list.

CombMNZ. CombMNZ multiplies the sum of normalized scores by the number of supporting sources:

$$f_{\mathrm{CombMNZ}}(q,d) = \left(\sum_{j=1}^{M} \mathbb{I}[z_{qj}(d) > 10^{-8}]\right)\left(\sum_{j=1}^{M} z_{qj}(d)\right). \tag{18}$$

Top-100 majority voting. Each nonconstant source gives one vote to each of its top-100 candidates, or to all candidates when fewer than 100 are available. The implementation uses

$$f_{\mathrm{MV}}(q,d) = \sum_{j\in\mathcal{J}_q} \mathbb{I}[r_{qj}(d) \le \min(100, |\mathcal{C}_q|)] + 10^{-3} f_{\mathrm{Equal}}(q,d). \tag{19}$$

The small Equal-Fusion term resolves ties in vote count without overturning a one-vote difference. Any remaining ties use the common identifier rule.

B.3 UWS retrieval adaptations

Applying UWS (Shin et al., 2022) to our retrieval setting requires explicit choices for variable-length candidate sets, tied source scores, and ranking generation. These choices affect both parameter

estimation and the resulting rankings. We therefore evaluate three configurations. The main table reports their per-column maximum, while their definitions are given below.

UWS-OPT-Medoid retains the median-triplet Mallows-mapping structure and restricted-medoid decoding of the selected public implementation branches, with numerical and variable-length adaptations described below. UWS-OPT-Score uses the same estimated weights with score-based fusion, isolating the decoding choice. UWS-TauB-Score provides an alternative that accounts for score ties in its distance statistic, together with a distinct parameter mapping.

All configurations use the same candidate pools, source scores, and evaluation partitions as DNLM, and fit source weights without relevance labels. We distinguish preserved upstream components from retrieval-specific adaptations, rather than claiming an exact reproduction of the original end-to-end pipeline. The variant names below identify our experimental configurations, not official names from the UWS paper.

UWS-OPT: distance estimation. For each query, source scores induce complete rankings with exact ties resolved by ascending item identifier. Let $K(\sigma_{qi}, \sigma_{qj})$ denote their Kendall inversion count, and let N_q be the candidate-set size. We estimate

$$D_{ij} = \frac{1}{|\mathcal{Q}_{\mathrm{tr}}|} \sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \frac{K(\sigma_{qi}, \sigma_{qj})}{\binom{N_q}{2}}. \tag{20}$$

The median-triplet moment for source i is

$$\mu_i = \operatorname{median}_{\substack{j<k \\ j,k \neq i}} \frac{D_{ij} + D_{ik} - D_{jk}}{2}. \tag{21}$$

All source pairs use the same training queries, and all eligible source triples are included.

UWS-OPT: parameter mapping. We map these moments to Kendall–Mallows concentration parameters using the expected inversion count

$$g_N(\theta) = \sum_{k=1}^{N} \left[\frac{1}{e^{\theta} - 1} - \frac{k}{e^{k\theta} - 1} \right], \qquad g_N(0) = \frac{N(N-1)}{4}. \tag{22}$$

For variable candidate-set sizes, we solve

$$\frac{1}{|\mathcal{Q}_{\mathrm{tr}}|} \sum_{q \in \mathcal{Q}_{\mathrm{tr}}} \frac{g_{N_q}(\theta_i)}{\binom{N_q}{2}} = \max(\mu_i, 10^{-8}). \tag{23}$$

This adaptation assumes a shared raw-Kendall concentration parameter per source across training queries; test-query lengths do not enter estimation.

We restrict θ_i to $[0, 64]$, set it to zero when the target moment is at least 0.5, and record numerical boundary activations. Weights are $w_i = \theta_i / \sum_j \theta_j$, with uniform weights used if all parameters are zero.

UWS-OPT-Medoid. Using these weights, the decoder selects an input ranking:

$$j_q^{\star} \in \arg\min_j \sum_k w_k \frac{K(\sigma_{qj}, \sigma_{qk})}{\binom{N_q}{2}}, \qquad \widehat{\sigma}_q = \sigma_{q j_q^{\star}}. \tag{24}$$

Objective ties follow the fixed source order. This preserves the restricted weighted-medoid behavior of the selected upstream decoder; it is not exact Kemeny optimization over all permutations.

UWS-OPT-Score. This variant uses the same fitted weights but replaces medoid selection with normalized-score fusion:

$$f(q, d) = \sum_j w_j z_{qj}(d). \tag{25}$$

It isolates the effect of decoding under the same parameter estimates.

UWS-TauB-Score. This alternative computes Kendall's tau-b directly on source scores, retaining ties in the distance statistic. For each source pair, it averages $(1 - \tau_b)/2$ over training queries where both score vectors are nonconstant and the statistic is finite. Each pair must have at least one valid training query.

We apply Eq. 21 to this distance matrix and use the mapping

$$a_i = \begin{cases} 1/\mu_i, & \mu_i > 0.01, \\ 0.01, & \text{otherwise}, \end{cases} \qquad w_i = \frac{\mathrm{clip}(a_i, 0.01, 100)}{\sum_j \mathrm{clip}(a_j, 0.01, 100)}. \tag{26}$$

The resulting weights are used for normalized-score fusion. This thresholded mapping is not the Mallows inversion used by UWS-OPT, and pairwise query filtering can give different source pairs different estimation supports.

Implementation scope. Our upstream reference is the public UWS repository at commit `503460b1472b`. The OPT implementation follows its median-triplet Mallows-mapping branch, with explicit adaptations for variable-length retrieval rankings and stable numerical inversion. Score fusion is an additional retrieval decoder.

We do not provide a dependence graph or select independent source triples in these configurations. Accordingly, these comparisons do not imply that the general UWS framework cannot model dependence. The three configurations remain separately named because their distance statistics, parameter mappings, and decoders are not interchangeable.

C Solving the Association-Aware Fusion Subproblem and DNLM's Algorithm

C.1 Solving the Association-Aware Fusion Subproblem

For fixed source associations $\boldsymbol{A}$ and reliability indices $\boldsymbol{\rho}$, Stage II solves

$$\min_{\boldsymbol{w} \in \Delta_{m-1}} F(\boldsymbol{w}) = \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{w} - \boldsymbol{\rho}^\top \boldsymbol{w}. \tag{27}$$

The matrix $\boldsymbol{A}$ is symmetric but need not be positive semidefinite. The objective is therefore potentially nonconvex. Nevertheless, a global minimum exists because F is continuous and the simplex is compact. We exploit the small number of sources by enumerating all nonempty supports rather than applying a local optimizer.

Support enumeration. Let $\mathcal{S} \subseteq \{1, \ldots, m\}$ be a nonempty candidate support. We set $w_j = 0$ for $j \notin \mathcal{S}$. A stationary point in the relative interior of this face satisfies

$$\underbrace{\begin{bmatrix} \boldsymbol{A}_{\mathcal{S},\mathcal{S}} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}}_{\boldsymbol{K}_{\mathcal{S}}} \begin{bmatrix} \boldsymbol{w}_{\mathcal{S}} \\ \lambda \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\rho}_{\mathcal{S}} \\ 1 \end{bmatrix}}_{\boldsymbol{c}_{\mathcal{S}}}. \tag{28}$$

Here, λ is the multiplier for the sum-to-one constraint. For a nonsingular $\boldsymbol{K}_{\mathcal{S}}$, we obtain its unique solution. We retain candidates satisfying nonnegativity, embed them into the full m-dimensional simplex, and compare their objective values. Singleton supports include all simplex vertices. The selected candidate has the lowest objective value among the retained candidates.

Why nonsingular faces and vertices suffice. The following argument concerns exact arithmetic. Among the global minimizers, choose one, $\boldsymbol{w}^\star$, with the smallest support $\mathcal{S}$. Its positive coordinates place it in the relative interior of that face, so it satisfies Eq 28.

Suppose $\boldsymbol{K}_{\mathcal{S}}$ were singular. There would exist a nonzero vector $(\boldsymbol{d}, \mu)$ such that

$$\boldsymbol{A}_{\mathcal{S},\mathcal{S}} \boldsymbol{d} + \mu \mathbf{1} = \mathbf{0}, \qquad \mathbf{1}^\top \boldsymbol{d} = 0. \tag{29}$$

Necessarily $\boldsymbol{d} \neq \mathbf{0}$, and

$$\boldsymbol{d}^\top \boldsymbol{A}_{\mathcal{S},\mathcal{S}} \boldsymbol{d} = 0. \tag{30}$$

Stationarity also gives

$$\boldsymbol{d}^\top \left(\boldsymbol{A}_{\mathcal{S},\mathcal{S}}\boldsymbol{w}^\star_{\mathcal{S}} - \boldsymbol{\rho}_{\mathcal{S}}\right) = 0. \tag{31}$$

Consequently, the quadratic objective remains constant along $\boldsymbol{w}^\star_{\mathcal{S}} + t\boldsymbol{d}$. Since $\boldsymbol{d}$ is nonzero and sums to zero, we can move along this direction until a positive coordinate reaches zero while preserving feasibility. This produces an equally good minimizer with a smaller support, contradicting the choice of $\boldsymbol{w}^\star$.

Thus, at least one global minimizer is represented by a nonsingular face system, including a singleton face when the minimizer is a vertex. Enumerating these candidates and comparing their objective values therefore suffices in exact arithmetic. This does not require $\boldsymbol{A}$ to be positive semidefinite.

Numerically singular systems. The implementation uses double-precision arithmetic. A face system is classified as numerically singular when its numerical rank, computed with an absolute singular-value threshold of 10^{-10}, is deficient. For such systems, we compute the Moore–Penrose pseudoinverse with a relative singular-value cutoff of 10^{-12}. For nonsingular systems, the inverse is cached. These matrices depend only on the fixed $\boldsymbol{A}$ and are reused as $\boldsymbol{\rho}$ changes.

A pseudoinverse solution is only an additional candidate: it is not assumed to be feasible, nor to represent every solution of a singular system. It must pass the same residual and feasibility checks as the other candidates. The exact-arithmetic coverage argument above does not depend on retaining a candidate from a singular face.

Candidate checks. For a computed solution $\boldsymbol{z}_{\mathcal{S}} = (\boldsymbol{w}_{\mathcal{S}}, \lambda)$, we require

$$\left\|\boldsymbol{K}_{\mathcal{S}}\boldsymbol{z}_{\mathcal{S}} - \boldsymbol{c}_{\mathcal{S}}\right\|_\infty \le 10^{-8}\left(1 + \|\boldsymbol{c}_{\mathcal{S}}\|_2\right). \tag{32}$$

Candidates with any weight below -10^{-10} are discarded. Remaining small negative weights are clipped to zero, and the weights are renormalized to sum to one. Objective values are evaluated after this adjustment. The incumbent candidate is replaced only when the objective decreases by more than 10^{-14}. If no candidate is retained, the solver raises an error.

Final stationarity and curvature checks. For the selected weights, define

$$\mathcal{S}_+ = \{j : w_j > 10^{-8}\}, \qquad \boldsymbol{g} = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{\rho}, \qquad \bar{g} = \frac{1}{|\mathcal{S}_+|}\sum_{j\in\mathcal{S}_+} g_j. \tag{33}$$

We check the first-order simplex conditions using

$$r_{\mathrm{KKT}} = \max\left\{\max_{j\in\mathcal{S}_+}|g_j - \bar{g}|,\ \max_{j\notin\mathcal{S}_+}[\bar{g} - g_j]_+\right\}, \tag{34}$$

where $[x]_+ = \max\{x, 0\}$ and the inactive-set term is zero when that set is empty. We require $r_{\mathrm{KKT}} \le 10^{-7}$.

When $|\mathcal{S}_+| > 1$, let $\boldsymbol{Q}$ have orthonormal columns spanning the tangent space $\{\boldsymbol{d} : \mathbf{1}^\top\boldsymbol{d} = 0\}$ on the active face. We additionally require

$$\lambda_{\min}\left(\boldsymbol{Q}^\top\boldsymbol{A}_{\mathcal{S}_+,\mathcal{S}_+}\boldsymbol{Q}\right) \ge -10^{-8}. \tag{35}$$

A failure of either check raises an error rather than silently returning the candidate.

Computational and theoretical scope. There are $2^m - 1$ nonempty supports: 31 for five sources and 127 for seven sources. The current implementation supports at most ten sources. It neither projects $\boldsymbol{A}$ onto the positive-semidefinite cone nor uses its inverse as a covariance-based weighting rule.

The residual, stationarity, and curvature checks are numerical safeguards, not standalone certificates of global optimality for an arbitrary nonconvex problem. The exact-arithmetic result follows from exhaustive support enumeration; it does not provide a quantified floating-point optimality bound. Moreover, this result applies only to the weight subproblem with $\boldsymbol{\rho}$ fixed. It does not establish convergence or global optimality of the alternating PL and fusion procedure.

Algorithm 1 DNLM

Require: Training scores $\{\boldsymbol{S}_q\}_{q\in\mathcal{Q}_{\mathrm{tr}}}$; source count $m \geq 3$; PL prefix length K; iteration budgets T, H; initial parameters $\boldsymbol{b}_0$
Ensure: Fusion weights $\boldsymbol{w}^{\mathrm{final}}$
Normalize scores into $\{\boldsymbol{Z}_q\}$ using Eq. (4); obtain source rankings $\{\sigma_{qj}\}$.
Stage I: Source association estimation
$\boldsymbol{b} \leftarrow \boldsymbol{b}_0$; initialize Adam.
for $r = 1, \ldots, T$ **do**
  $\boldsymbol{\rho} \leftarrow \mathbf{1} - \mathrm{sigmoid}(\boldsymbol{b})$; $\boldsymbol{u} \leftarrow \boldsymbol{\rho}/(\mathbf{1}^\top \boldsymbol{\rho})$
  $\boldsymbol{v}_q \leftarrow \max\{\boldsymbol{Z}_q \boldsymbol{u}, \delta_v \mathbf{1}\}$ for all training queries
  Update $\boldsymbol{b}$ with H Adam steps on Eq. (3), using Eq. (7) and holding $\{\boldsymbol{v}_q\}$ fixed.
end for
Recompute $\boldsymbol{\rho}$ and $\boldsymbol{u}$ from the final $\boldsymbol{b}$.
$\boldsymbol{A} \leftarrow \boldsymbol{I}_m$
for each source pair (j, k) with $j < k$ **do**
  Construct anchors $\{\boldsymbol{v}_q^{(-jk)}\}$ using Eq. (8).
  Compute centered deviations using Eq. (9).
  Compute $A_{j,k} = A_{k,j}$ using Eq. (10).
end for
Stage II: Association-aware fusion
$\boldsymbol{b} \leftarrow \boldsymbol{b}_0$; reset Adam; keep $\boldsymbol{A}$ fixed.
for $r = 1, \ldots, T$ **do**
  $\boldsymbol{\rho} \leftarrow \mathbf{1} - \mathrm{sigmoid}(\boldsymbol{b})$
  $\boldsymbol{w} \leftarrow \boldsymbol{w}^*(\boldsymbol{\rho}; \boldsymbol{A})$ using Eq. (11)
  $\boldsymbol{v}_q \leftarrow \max\{\boldsymbol{Z}_q \boldsymbol{w}, \delta_v \mathbf{1}\}$ for all training queries
  Update $\boldsymbol{b}$ with H Adam steps on Eq. (3), using Eq. (7) and holding $\{\boldsymbol{v}_q\}$ fixed.
end for
$\boldsymbol{\rho}^{\mathrm{final}} \leftarrow \mathbf{1} - \mathrm{sigmoid}(\boldsymbol{b})$
$\boldsymbol{w}^{\mathrm{final}} \leftarrow \boldsymbol{w}^*(\boldsymbol{\rho}^{\mathrm{final}}; \boldsymbol{A})$ using Eq. (11)
return $\boldsymbol{w}^{\mathrm{final}}$

C.2 DNLM's Algorithm

Algorithm 1 gives the full procedure of DNLM.

D Controlled Agreement at Fixed Per-Source Ranking Quality

D.1 Experimental construction

We generate synthetic source rankings to examine how agreement on relevant and irrelevant candidates affects fusion while holding each source's individual ranking quality fixed. The experiment directly constructs score matrices; no query or document text is involved.

Fixed relevance and ranking templates. Each query contains 100 candidates and five sources. We randomly assign relevance grades: five candidates have grade 3, five have grade 2, ten have grade 1, and eighty have grade 0. Positive-grade candidates are considered relevant.

For each source, a rank-grade template specifies the relevance grade at every ranking position. The first 20 positions contain two candidates of each positive grade and fourteen grade-0 candidates, randomly ordered. The remaining positions contain the remaining candidates, also randomly ordered. Sources 1 and 2 receive independently sampled templates, while sources 3–5 share a template and form the target group. Relevance assignments and templates are reused across all conditions within each query and seed.

We first populate every source's template using independent permutations of candidate identities within each relevance grade. This defines the zero-control condition. Sources 1–3 remain unchanged throughout the grid; the agreement controls modify sources 4 and 5 relative to source 3.

Controlling agreement. We independently vary relevant and irrelevant agreement using

$$c_{\mathrm{rel}}, c_{\mathrm{irr}} \in \{0, 0.25, 0.5, 0.75, 1\}.$$

For each query and relevance grade, let n_g be the number of positions of that grade. We sample a random priority ordering of these positions and a value $U_g \sim \mathrm{Uniform}(0, 1)$. At control level c, we lock the first

$$L_g(c) = \lfloor cn_g + U_g \rfloor$$

positions in this ordering. The same ordering and U_g are reused across control levels, giving nested locked-position sets with expected size cn_g. Positive grades use $c = c_{\mathrm{rel}}$; grade 0 uses $c = c_{\mathrm{irr}}$.

At locked positions, sources 4 and 5 receive the same candidate identities as source 3. Their remaining positions are filled in the order of their original within-grade permutations, after removing the locked candidates. This preserves valid rankings without duplicate candidate identities within a source. The filtering step can also rearrange unlocked positions; the control specifies deliberately shared positions, not the total fraction of positions that change.

At $c = 0$, assignments remain independent; at $c = 1$, assignments are fully shared within the corresponding grades. Intermediate values progressively introduce shared assignments across the full ranking. The four endpoint conditions exactly reproduce the original low/low, low/high, high/low, and high/high construction. When both controls equal one, the target sources have identical rankings.

The controls are not observed overlap rates. Unlocked assignments may coincide by chance, and the realized top-20 overlap need not equal c.

Scores and matched quantities. A candidate at one-based rank r receives score

$$s(r) = \exp\left(-\frac{r-1}{20}\right), \qquad r = 1, \ldots, 100.$$

We apply the standard per-query, per-source min–max normalization before fusion. Every source has the same score distribution, and the generated source rankings have no ties.

Because candidate identities are reassigned only within relevance grades, each source's complete ranked grade sequence remains unchanged across all 25 conditions. Its individual relevance-based ranking metrics and score distribution are therefore exactly matched. This matching is across conditions for each source; sources with different templates need not have equal quality. Their combined coverage and information available to fusion can change.

D.2 Training and evaluation

For each of ten data seeds, we generate 200 training queries and 100 held-out test queries. Each method is fitted separately in each condition. Training receives only the training score matrices; relevance grades and target-group identities are used for generation and evaluation, not fitting. Test queries follow the same construction as their corresponding training condition.

We compare four methods: PL-only, which uses the Stage-I anchor coefficients for fusion; the identity-association control ($\boldsymbol{A} = \boldsymbol{I}$), which retains the Stage-II reliability and fusion updates; full DNLM; and Equal fusion. All methods receive the same five sources and candidate sets. The identity-association control and DNLM use the same Stage-II procedure and differ in the association matrix.

The implementation and hyperparameters remain unchanged across the grid. Each trained stage uses 20 alternating iterations, five PL gradient updates per iteration, a top-20 PL likelihood, and Adam with learning rate 0.01. The algorithm seed is fixed at 42. Stage-II parameters and optimizer state are initialized afresh, and no condition-specific tuning is performed.

For each condition and seed, we average NDCG@10 over the 100 test queries. We report means across the ten data seeds and sample standard deviations of seed-level paired method differences. These standard deviations describe dataset-generation variability and are not confidence intervals. Evaluation ties are resolved deterministically by candidate identity.

We verify valid candidate permutations, preserved rank-grade templates and score distributions, and exactly matched per-query single-source metrics. The four endpoint datasets and their fitted results reproduce the original experiment exactly. Fusion scores are checked against reconstruction from the learned weights, and evaluation metrics are checked against `pytrec_eval`.

Table 5: Mean NDCG@10 over ten data seeds as irrelevant agreement c_{irr} varies, with relevant agreement fixed at $c_{\text{rel}} = 0$. Bold denotes the highest displayed mean in each row.

c_{irr}	Equal	PL-only	DNLM
0.00	.3897	**.3898**	.3897
0.25	.3132	.3117	**.3133**
0.50	.2521	.2474	**.2583**
0.75	.2043	.1968	**.2231**
1.00	.1663	.1565	**.2005**

Table 6: DNLM minus PL-only NDCG@10, in percentage points: mean ± sample standard deviation of paired differences over ten data seeds. Rows and columns specify the relevant and irrelevant agreement controls.

	c_{irr}				
c_{rel}	0	0.25	0.50	0.75	1
0	-0.01 ± 0.09	$+0.16 \pm 0.26$	$+1.09 \pm 0.32$	$+2.63 \pm 0.42$	$+4.40 \pm 0.44$
0.25	-0.00 ± 0.13	$+0.23 \pm 0.28$	$+1.11 \pm 0.51$	$+2.65 \pm 0.49$	$+4.54 \pm 0.61$
0.50	-0.12 ± 0.16	$+0.22 \pm 0.30$	$+0.98 \pm 0.60$	$+2.31 \pm 0.62$	$+4.18 \pm 0.55$
0.75	-0.29 ± 0.20	-0.07 ± 0.35	$+0.69 \pm 0.37$	$+1.67 \pm 0.72$	$+3.39 \pm 0.43$
1	-0.36 ± 0.18	-0.15 ± 0.36	$+0.63 \pm 0.52$	$+1.52 \pm 0.71$	$+3.21 \pm 0.55$

D.3 Full-grid results

Table 6 reports the paired DNLM–PL-only differences across all 25 combinations of independently varied relevant and irrelevant agreement. At every fixed level of relevant agreement, DNLM's mean advantage over PL-only increases with irrelevant agreement. As c_{irr} increases from zero to one, the difference changes from -0.01 to $+4.40$ percentage points when $c_{\text{rel}} = 0$, and from -0.36 to $+3.21$ points when $c_{\text{rel}} = 1$. Thus, this pattern does not require a simultaneous reduction in relevant agreement.

Table 5 reports absolute scores with $c_{\text{rel}} = 0$. As c_{irr} increases from zero to one, Equal fusion's NDCG@10 decreases from 0.3897 to 0.1663, despite unchanged individual-source ranking quality. PL-only follows a similar decline, from 0.3898 to 0.1565. DNLM also declines, from 0.3897 to 0.2005, but retains higher effectiveness than both controls at each nonzero level of irrelevant agreement. Its benefit therefore reflects mitigation of the performance loss as shared irrelevant recommendations increase.

When $c_{\text{irr}} = 0$, DNLM's mean differences from PL-only are near zero or negative. The largest mean loss is 0.36 points at $(c_{\text{rel}}, c_{\text{irr}}) = (1, 0)$, while the largest mean gain is 4.54 points at $(0.25, 1)$. These results show that association-aware fusion can mitigate harmful repeated support, but can also suppress useful agreement. The observed patterns characterize this controlled construction rather than a general monotonicity guarantee.

D.4 Joint variation of the two agreement controls

To summarize the transition from shared relevant recommendations to shared irrelevant recommendations, Table 7 reports the anti-diagonal $c_{\text{rel}} = 1 - c_{\text{irr}}$. This is a slice through the two-dimensional grid, not a fixed budget of total agreement.

DNLM's mean gain over PL-only increases along this slice from -0.36 to $+4.40$ points. At $(0.5, 0.5)$, the gain is 0.98 points, with positive paired differences in all ten seeds. Its gain over Equal is 0.57 points, positive in nine seeds. At $(0.75, 0.25)$, however, the difference from PL-only is -0.07 points and positive in only four seeds; this condition does not establish a stable benefit or a precise transition threshold.

Table 7: Mean NDCG@10 over ten data seeds along $c_{\text{rel}} = 1 - c_{\text{irr}}$. Bold marks the highest displayed mean in each row.

c_{rel}	c_{irr}	Equal	PL-only	$\boldsymbol{A} = \boldsymbol{I}$	DNLM
1.00	0.00	.4687	.4701	.4723	.4665
0.75	0.25	.4036	.4037	.4023	.4029
0.50	0.50	.3235	.3194	.2718	.3292
0.25	0.75	.2425	.2371	.1705	**.2636**
0.00	1.00	.1663	.1565	.0871	**.2005**

Table 8: Realized top-20 overlap and target-group diagnostics along the anti-diagonal, averaged over ten data seeds. $\overline{A}_G$ is DNLM's within-group association; W_G is total target-group weight.

		Overlap			W_G		
c_{rel}	c_{irr}	Rel.	Irr.	$\overline{A}_G$	PL-only	$\boldsymbol{A} = \boldsymbol{I}$	DNLM
1.00	0.00	1.000	.175	.357	.608	.631	.587
0.75	0.25	.931	.383	.442	.613	.639	.574
0.50	0.50	.733	.598	.516	.616	.977	.561
0.25	0.75	.533	.800	.603	.619	1.000	.545
0.00	1.00	.333	1.000	.698	.620	1.000	.528

D.5 Realized agreement and fusion weights

For each target-source pair, we measure top-20 overlap separately among relevant and irrelevant candidates, dividing their intersection sizes by 6 and 14, respectively. We average over the three target-source pairs, test queries, and data seeds. We also report DNLM's mean within-group association

$$\overline{A}_G = \frac{A_{3,4} + A_{3,5} + A_{4,5}}{3}$$

and the target group's total fusion weight $W_G = w_3 + w_4 + w_5$. For PL-only, W_G uses its normalized anchor coefficients.

Table 8 shows that the controls and realized overlap are distinct. At $(0.5, 0.5)$, relevant and irrelevant overlap are 0.733 and 0.598, respectively. Even at zero control, incidental overlap remains.

Along the anti-diagonal, DNLM's mean within-group association rises from 0.357 to 0.698, while its target-group weight falls from 0.587 to 0.528. PL-only's group weight instead rises slightly, from 0.608 to 0.620. The identity-association control concentrates all weight on the target group at the final two points. This concentration accompanies its substantial performance deterioration and contextualizes DNLM's larger gains over this control than over PL-only.

The experiment shows increasing relative benefit from association-aware fusion as shared irrelevant recommendations become more prominent. This benefit reflects mitigation of performance loss: absolute retrieval effectiveness can still decline as irrelevant agreement increases. The construction also changes joint candidate coverage and aligns both identities and positions, with different numbers of relevant and irrelevant candidates.

E Contribution of Negative Source Associations

Intervention. We assess the contribution of negative associations by modifying the matrix estimated in Stage I:

$$A_{j,k}^{\text{nonneg}} = \begin{cases} 1, & j = k, \\ \max\{A_{j,k}, 0\}, & j \neq k. \end{cases} \tag{36}$$

This intervention removes negative off-diagonal entries while preserving positive associations and the unit diagonal. No positive-semidefinite projection is applied.

For each training split, we reuse the same fitted Stage-I matrix and compare Stage-II fitting with $\boldsymbol{A}$ against fitting with $\boldsymbol{A}^{\text{nonneg}}$. Both runs start from identical reliability parameters and use the same

Table 9: NDCG@10 after removing negative source associations. Nonnegative-A sets negative off-diagonal entries of $\boldsymbol{A}$ to zero before refitting Stage II. Δ denotes DNLM minus Nonnegative-A in percentage points.

Dataset	Pool	DNLM	Nonnegative-A	Δ
LitSearch	G5	.4671	.4606	+0.66
	H7	.5084	.4803	+2.81
STaRK-MAG	G5	.3666	.3666	+0.00
	H7	.3920	.3836	+0.84
Prime (synthetic)	G5	.1858	.1859	-0.00
	H7	.2086	.1979	+1.07
Prime (human)	G5	.2732	.2734	-0.02
	H7	.2922	.2582	+3.40
TREC-COVID	G5	.6684	.6679	+0.05
	H7	.7211	.7108	+1.03

Differences are computed before rounding. Signed zeros denote changes smaller than 0.005 percentage points in magnitude.

scores, query splits, optimizer, and training budget: 20 outer iterations with five PL updates per iteration and $K = 20$. The association matrix remains fixed within each run. Reliability indices and fusion weights are relearned, so the intervention measures the contribution of negative associations through the full Stage-II fitting procedure.

Evaluation. We evaluate both G5 and H7 using the main experimental protocol: pooled five-fold held-out predictions for LitSearch and TREC-COVID, and official splits for STaRK-MAG and STaRK-Prime. Prime's synthetic and human queries are evaluated separately using the same fitted model. Rerunning the original configuration reproduces the archived NDCG@10 and MRR results without metric differences. Table 9 reports held-out point estimates; no statistical significance is asserted.

Results. Retaining negative associations improves NDCG@10 in every H7 setting, with gains ranging from 0.84 points on STaRK-MAG to 3.40 points on Prime human queries. In G5, the gain is 0.66 points on LitSearch, while changes in the other settings are at most 0.05 points in magnitude.

The intervention also changes the allocation of weight across sources. On LitSearch G5, removing negative associations reduces the combined weight of BM25 and MiniLM from 96.6% to 88.2%, while increasing the bibliographic-coupling source's weight from 3.4% to 11.8%, averaged across folds. Thus, negative associations can affect the allocation of weight beyond the associated pair itself.

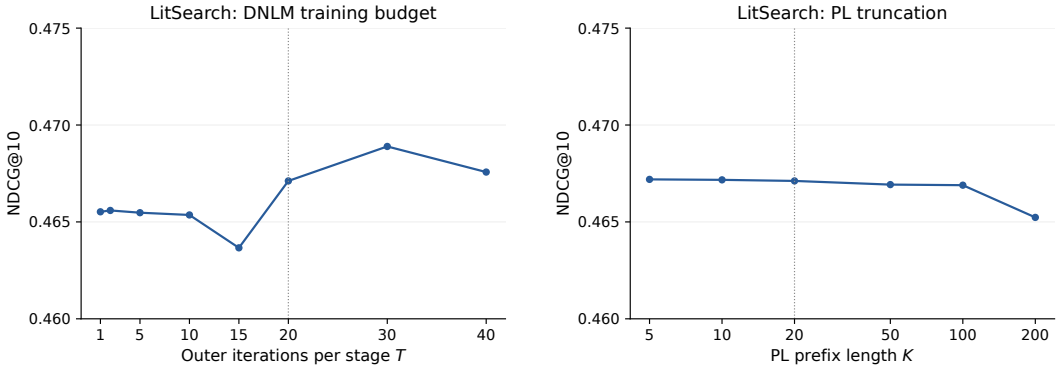
F Additional Experiments and Analyses

Table 10: H7 retrieval effectiveness and gains over G5 after adding MPNet and E5-large. H7 columns report NDCG@10; Δ denotes the change from G5 in percentage points. Bold marks the highest H7 NDCG@10 in each setting.

	LitSearch		STaRK-MAG		Prime (synthetic)		Prime (human)		TREC-COVID	
Method	H7	Δ	H7	Δ	H7	Δ	H7	Δ	H7	Δ
Equal	.4128	+6.10	.3173	+4.43	.1860	+1.99	.2208	+3.65	.6281	+2.52
RRF	.4206	+7.95	.3392	+5.61	**.2110**	+1.59	.2425	+3.59	.6510	+9.70
PL-only	.4263	+6.83	.3257	+4.59	.1875	+1.99	.2204	+3.65	.6372	+3.04
UWS-TauB-Score	.3971	+6.71	.2978	+4.46	.1757	+1.84	.2043	+4.44	.6243	+2.79
DNLM	**.5084**	+4.13	**.3920**	+2.54	.2086	+2.27	**.2922**	+1.90	**.7211**	+5.27

Table 11: Computational costs on LitSearch G5, with inference evaluated on shared candidate pools. Measurements use an Apple M4 Max: MPS for neural encoding and similarity scoring, and a single CPU thread for BM25, graph scoring, and DNLM.

Component	Offline (s)	Inference (ms/query)
DNLM (G5)		
MiniLM	58.83	11.85
BM25 + three graph sources	5.48	5.12
Source-association matrix (Stage I)	8.19	Cached
Reliability and fusion (Stage II)	8.02	0.62
Measured subtotal	80.52	17.63
External reference		
GTE-large-v1.5	1307.20	27.99



(a) Training budget.

(b) PL truncation length.

Figure 3: Sensitivity on LitSearch G5. Values are NDCG@10 pooled over five held-out folds. (a) Both stages use T outer iterations, with $K = 20$. (b) K varies with 20 iterations per stage; the horizontal axis is logarithmic. Each outer iteration contains five PL updates. Dotted lines mark the default settings.