RADAR: RADIAL ATTRIBUTE DISTILLATION FOR AUGMENTED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense embeddings primarily encode similarity through the vector direction, while vector magnitude is often normalized away or lacks an explicit semantic role. We introduce RADAR (Radial Attribute Distillation for Augmented Representations), a general framework for augmenting pretrained embeddings with latent attributes without disrupting their underlying semantic geometry. RADAR learns from multiple heterogeneous and noisy proxy signals to identify a latent attribute that is both supported by the proxies and predictable from the original representation. It then encodes the predicted attribute through bounded radial scaling: embedding direction—and hence the original semantic representation—is preserved, while vector norm carries complementary information that can be exploited directly by inner-product retrieval. The framework is domain-agnostic and does not require retraining the underlying embedding model or observing proxy signals at inference time. We evaluate RADAR using scientific literature retrieval as a challenging testbed, where citations and venue-related signals provide imperfect proxies for scholarly impact. Across four retrieval benchmarks, augmenting Qwen3-Embedding-8B with RADAR improves average nDCG@10 from 45.16 to 46.02 and Recall@50 from 50.01 to 51.31, while retaining single-stage vector retrieval. The distilled attribute also exhibits stronger associations with held-out peer-review judgments and prospective citation outcomes than direct proxy prediction and conventional factor-based alternatives on most measures. Our results suggest that radial degrees of freedom provide a simple mechanism for incorporating complementary latent attributes into existing embedding spaces while preserving their semantic structure and computational advantages.

1 INTRODUCTION

Modern dense retrieval systems typically ℓ_2-normalize representations and rank via cosine similarity, restricting retrieval entirely to angular proximity. However, effective ranking relies on a document's inherent impact as much as its semantic relevance. For instance, a query for a specific methodology should ideally surface the seminal manuscript rather than topically identical but marginal follow-ups—a distinction imperceptible to a purely angular retriever. While classical information retrieval integrated query-independent authority (Craswell et al., 2005; Bendersky et al., 2011) and recent work has re-emphasized attributes beyond topical relevance (Zhou et al., 2025), current embedding pipelines generally manage these signals outside the core representation. They either rely on costly encoder fine-tuning, which distorts the established semantic geometry, or introduce auxiliary reranking stages that inflate latency. Concurrently, a fundamental degree of freedom within the embedding space remains unexploited: its vector norm (Oyama et al., 2023).

The central challenge is that the target attribute is rarely observed directly. For example, scholarly impact of a paper, like many document qualities, lacks ground-truth labels and must be inferred from heterogeneous, noisy proxies such as citation counts and venue prestige (Ratner et al., 2016). Furthermore, because these proxies are typically unavailable at indexing time for new documents, the latent attribute must ultimately be predicted from the frozen semantic embedding. This dual requirement creates a fundamental tension in defining the attribute itself. Standard aggregation techniques, such as principal component analysis, isolate directions of maximum proxy variance that may be poorly predictable from textual content. Conversely, optimizing predictability without a fidelity constraint can emphasize proxy variation that is easy to predict from text but weakly

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.

aligned with the dominant shared proxy structure. Existing methods typically default to one extreme: canonical correlation analysis only identifies the most predictable direction, while reduced-rank regression only maximizes predicted variance without this trade-off.

To resolve this, we introduce **RADAR** (**R**adial **A**ttribute **D**istillation for **A**ugmented **R**epresentations). Its core component, *Proxy-Anchored Latent Impact Factorization* (PALIF), embeds inference-time predictability directly into attribute distillation. PALIF selects the proxy-space direction that minimizes out-of-fold prediction error from the frozen embedding, strictly constrained to retain at least a $1 - \varepsilon$ fraction of the leading proxy direction's variance. Our population analysis shows that the predictability criterion measures the fraction of proxy variance explainable from the frozen representation. A sharp spectral result further quantifies how ε, together with the proxy eigenspectrum, limits departure from dominant proxy variation.

To integrate the predicted attribute, RADAR employs a transparent mechanism: *radial encoding*. Each document embedding is rescaled by $\exp(\gamma \widehat{I}_d)$, where $\widehat{I}_d \in [0, 1]$ is a normalized predicted attribute. Because this operation modifies only vector magnitude, it preserves embedding directions and pairwise cosine similarities while remaining compatible with single-vector maximum-inner-product search (MIPS). The multiplicative form also bounds ranking perturbation: among positively scored documents, a reversal is possible only when the semantic score ratio is below e^{γ}, and the exact RADAR top-k can differ from the base top-k only within an explicit multiplicative band. Thus radial encoding makes the attribute influence both transparent and quantitatively controlled.

RADAR is fundamentally domain-agnostic, requiring only a frozen encoder and training-time proxy observations. While applicable to any domain with noisy utility signals (e.g., e-commerce or legal search), we instantiate it in scientific literature retrieval because this setting uniquely provides independent external validity signals—peer-review scores and future citations—to rigorously evaluate the distilled attribute. We assess RADAR across four diverse benchmarks, including tasks where relevance is explicitly decoupled from citation counts, using four distinct embedding backbones (Qwen3-Embedding-8B, NV-Embed-v2, E5-large-v2, and SPECTER2). RADAR consistently improves retrieval performance across all backbones and metrics. Notably, when augmenting Qwen3-Embedding-8B, it outperforms the Qwen3-Reranker-8B pipeline while operating $19\times$ to $290\times$ faster, as all attribute inference and scaling are computed offline. Furthermore, PALIF outperforms direct proxy prediction, multi-proxy aggregation, and classical factor models. The distilled attribute generally exhibits stronger alignment with held-out peer-review judgments and prospective citations than these alternatives, providing external-validity evidence that the distilled attribute captures information related to scholarly impact beyond the training proxies.

Our main contributions are:

- **Framework:** We augment frozen embeddings with latent attributes inferred from proxies, utilizing the vector norm to encode impact orthogonally to semantic direction.
- **Method & Guarantees:** PALIF targets explainable proxy variance with an explicit learned-predictor approximation bound and sharp spectral anchoring (Props. 1–2); radial encoding preserves semantic geometry while admitting tight semantic-margin and top-k stability guarantees (Prop. 4; Cor. 5).
- **Benchmarks:** We construct two novel scientific retrieval benchmarks: SuReIR (survey-to-bibliography) and GSRR (scholar rankings).
- **Evidence:** RADAR improves four backbones across all benchmarks, surpassing an 8B reranker at a fraction of the latency with consistent gains in all cases.

2 RELATED WORK

2.1 ATTRIBUTE-AWARE DENSE RETRIEVAL

Dense retrieval ranks documents by similarity in learned representation spaces (Karpukhin et al., 2020; Xiong et al., 2020), while retrieval can also exploit query-independent document signals such as quality and authority (Craswell et al., 2005; Bendersky et al., 2011). Quality-biased ranking directly incorporates document-quality features into the ranking function (Bendersky et al., 2011). Beyond score-level integration, EPIC learns a query-independent document-quality factor that multiplicatively scales vocabulary-space document representations during relevance training (MacA-

vaney et al., 2020). Embedding magnitude can itself encode information beyond vector direction: squared word-vector norms have been linked to information gain (Oyama et al., 2023), while item-vector magnitudes have been adjusted at inference time to control popularity effects in recommendation (Kim et al., 2023). In contrast, RADAR encodes a predicted latent attribute offline via bounded radial scaling, preserving the original semantic geometry and single-vector MIPS efficiency.

2.2 LATENT ATTRIBUTES FROM NOISY PROXIES

Latent attributes are commonly studied through multiple observable indicators that provide imperfect evidence about an underlying construct (Jöreskog & Goldberger, 1975). In scholarly impact prediction, a common simplification is to predict a single citation-based indicator directly from text (Li et al., 2019; Zhao et al., 2025). When multiple proxies are available and the learned attribute must later be recovered from a frozen representation, both proxy fidelity and predictability become relevant. Classical methods emphasize different criteria: PCA maximizes proxy variance (Jolliffe & Cadima, 2016), CCA maximizes cross-view correlation (Hotelling, 1992), and RRR fits a low-rank predictive map (Izenman, 1975); recent learning-based methods instead couple noisy sources with model training (Rühling Cachay et al., 2021) or weight quality scores using predicted validation loss (Zhuang et al., 2025). None of these criteria lets one prescribe how much proxy variance must be retained while predictability is optimized. RADAR makes this trade-off explicit, minimizing cross-fitted prediction error from the frozen embedding subject to retaining a $(1-\varepsilon)$ fraction of maximal proxy variance; ε interpolates between PCA ($\varepsilon = 0$) and a predictability-maximizing direction that reduces to CCA under an in-sample linear predictor, with an exact bound on departure from the dominant proxy axis (Proposition 2).

3 THE RADAR FRAMEWORK

RADAR is a framework (Figure 1) that augments frozen semantic embeddings with a query-independent latent attribute, such as scholarly impact. RADAR addresses two core conceptual challenges:

- **Proxy-Anchored Latent Impact Factorization (Section 3.1):** Because the target attribute is unobserved and must be inferred from the document and noisy proxies, we face a trade-off between retaining variance in the proxy space and ensuring predictability from the document embedding. PALIF formulates this as a structured optimization problem. It isolates a latent direction that minimizes out-of-fold prediction error from the semantic embeddings, strictly constrained to retain a dominant fraction of the leading proxy variance.
- **Geometric Preservation via Radial Scaling (Section 3.2):** Once the latent attribute is predicted, it is injected into the representation without distorting the established semantic space. RADAR achieves this by mapping the predicted attribute to a scalar and applying it as a positive radial transformation. This structural separation preserves pairwise cosine geometries while seamlessly modulating standard maximum-inner-product search (MIPS).

Standard dense retrieval systems score documents based exclusively on angular proximity. We consider a retrieval setting over a corpus $\mathcal{D}$ using a frozen general-purpose encoder that maps a query q and a document d to ℓ_2-normalized semantic embeddings $\mathbf{u}_q, \mathbf{u}_d \in \mathbb{S}^{p-1}$, where $\mathbb{S}^{p-1}$ denotes the unit sphere in $\mathbb{R}^p$. Documents are conventionally ranked in descending order of their cosine similarity, $\mathbf{u}_q^\top \mathbf{u}_d$, a formulation that inherently discards vector magnitude. Because our target attribute lacks ground-truth labels, it must be learned from proxies during training. For N training documents, we define the training set: $\mathcal{T} = \left\{ \left(\mathbf{u}_i, \{(x_{im}, \omega_{im})\}_{m=1}^{M}, g_i \right) \right\}_{i=1}^{N}$, where $\mathbf{u}_i \in \mathbb{S}^{p-1}$ denotes the frozen semantic embedding of document i. For each proxy $m = 1, \ldots, M$, the indicator $\omega_{im} \in \{0, 1\}$ denotes whether the proxy is observed, and x_{im} represents its value when $\omega_{im} = 1$. The variable g_i indexes a stratum (e.g., publication year or field) within which proxy values are comparable. These proxies provide noisy, heterogeneous evidence of latent impact.

Our goal is to construct a document representation $\mathbf{z}_d$ such that ranking by the inner product $\mathbf{u}_q^\top \mathbf{z}_d$ jointly reflects query-dependent semantic relevance and latent impact. This formulation is subject to two strict requirements: it must preserve the original semantic direction produced by the frozen encoder ($\mathbf{z}_d / \|\mathbf{z}_d\| = \mathbf{u}_d$), and it must require no proxy observations for retrieval-corpus documents

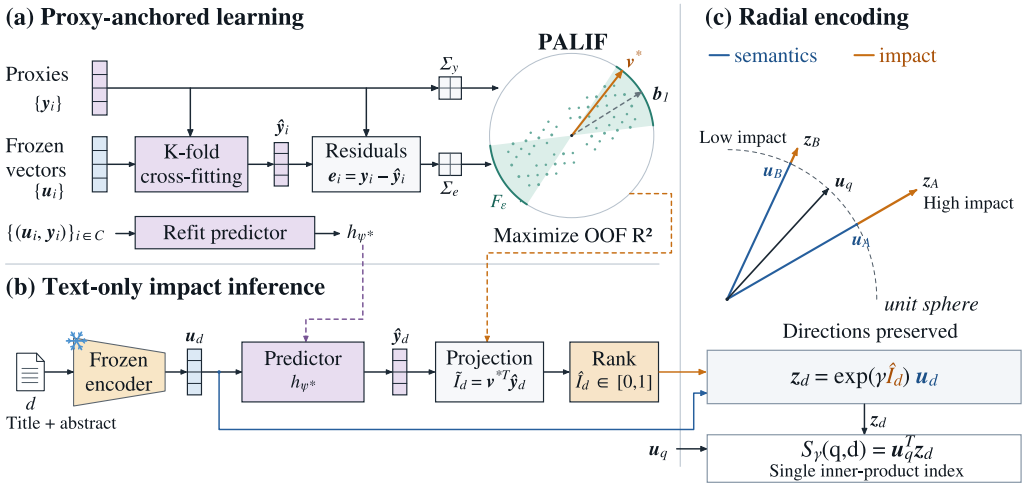


Figure 1: RADAR overview. **(a)** PALIF selects a proxy direction minimizing normalized out-of-fold prediction error under a variance-retention constraint. **(b)** A refitted predictor uses only frozen document embeddings. The predicted proxies are projected onto the selected direction and rank-normalized within the corpus. **(c)** Exponential radial scaling using normalized impact scores preserves embedding directions and pairwise cosine similarities while retaining single-vector MIPS.

at index time. In our implementation, the raw attribute score is predicted solely from $\mathbf{u}_d$ and is then mapped through an offline corpus-level monotone rank normalization before indexing.

3.1 Proxy-Anchored Latent Impact Factorization (PALIF)

Proxy Harmonization. We apply a rank-based inverse normal transformation within each stratum to place heterogeneous proxies on a comparable scale without reversing their within-stratum ordering. For each (i, m) with $\omega_{im} = 1$, let r_{im} denote the average rank of x_{im} among observations of proxy m in stratum g_i, with larger proxy values assigned larger ranks, and let $n_m(g_i)$ denote the number of such observations. Using the Rankit plotting position with offset $1/2$ (Bliss et al., 1956; Cooke & Haniffa, 2025), we define $\widetilde{x}_{im} = \Phi^{-1}\left[\mathrm{clip}\left(\frac{r_{im}-\frac{1}{2}}{n_m(g_i)}, \delta, 1-\delta\right)\right]$, where $\mathrm{clip}(a, l, u) = \min\{\max\{a, l\}, u\}$. Φ^{-1} denotes the standard normal quantile function, and $\delta \in (0, 1/2)$ controls the clipping of extreme quantiles. Let $\mathcal{C} = \{i \in \{1, \ldots, N\} : \omega_{im} = 1 \text{ for all } m = 1, \ldots, M\}$ denote the subset of training documents for which all proxies are observed, and let $N_{\mathcal{C}} = |\mathcal{C}|$. To form the complete M-dimensional proxy vectors required by PALIF without introducing an imputation model, we restrict factor estimation to $\mathcal{C}$. The rank transformation is computed within each stratum using all observed values, whereas PALIF operates on the subset $\mathcal{C}$. We therefore center and rescale each transformed proxy over $\mathcal{C}$. Let μ_m and σ_m denote the empirical mean and standard deviation of $\widetilde{x}_{im}$ over $i \in \mathcal{C}$, assuming $\sigma_m > 0$, and define $y_{im} = (\widetilde{x}_{im} - \mu_m)/\sigma_m$. For $i \in \mathcal{C}$, the harmonized proxy vector is $\mathbf{y}_i = (y_{i1}, \ldots, y_{iM})^\top$.

Out-of-Fold Proxy Prediction. To estimate proxy predictability from frozen embeddings without relying on potentially optimistic in-sample residuals, we construct out-of-fold predictions using K-fold cross-fitting (Bach et al., 2024). This construction prevents each training observation from being evaluated by the proxy predictor fitted on that observation. The resulting PALIF criterion is exactly equivalent to maximizing cross-fitted empirical coefficient of determination (R^2_{OOF}) over admissible directions. Cross-fitting addresses in-sample optimism in proxy prediction. Topic-related nuisance variation is addressed separately by the stratified proxy harmonization, while the variance-retention constraint controls departure from dominant proxy structure. Let $h_\psi : \mathbb{S}^{p-1} \to \mathbb{R}^M$ denote a multilayer perceptron with parameters ψ that predicts $\mathbf{y}_i$ from $\mathbf{u}_i$. Let $\{\mathcal{F}_k\}_{k=1}^K$ partition the indices

in $\mathcal{C}$. For each fold k, the predictor is fitted on the complementary folds:

$$\psi_k \in \arg\min_{\psi} \frac{1}{N_{\mathcal{C}} - |\mathcal{F}_k|} \sum_{\substack{i \in \mathcal{C} \\ i \notin \mathcal{F}_k}} \|h_\psi(\mathbf{u}_i) - \mathbf{y}_i\|_2^2 . \tag{1}$$

For $i \in \mathcal{F}_k$, $\widehat{\mathbf{y}}_i = h_{\psi_k}(\mathbf{u}_i)$ denotes the out-of-fold prediction and $\mathbf{e}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i$ denotes the residual, which is used to characterize proxy prediction error in the subsequent factor-selection step.

Proxy-Anchored Factor Selection. We define the empirical second-moment matrices

$$\mathbf{\Sigma}_y = \frac{1}{N_{\mathcal{C}}} \sum_{i \in \mathcal{C}} \mathbf{y}_i \mathbf{y}_i^\top, \qquad \mathbf{\Sigma}_e = \frac{1}{N_{\mathcal{C}}} \sum_{i \in \mathcal{C}} \mathbf{e}_i \mathbf{e}_i^\top .$$

Let $\lambda_{\max}(\mathbf{\Sigma}_y)$ denote the largest eigenvalue of $\mathbf{\Sigma}_y$, and assume $\lambda_{\max}(\mathbf{\Sigma}_y) > 0$. For $\varepsilon \in [0, 1)$, Proxy-Anchored Latent Impact Factorization (PALIF) defines the factor direction as

$$\mathbf{v}^\star \in \arg\min_{\mathbf{v} \in \mathbb{R}^M,\, \|\mathbf{v}\|_2 = 1} \frac{\mathbf{v}^\top \mathbf{\Sigma}_e \mathbf{v}}{\mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v}} \quad \text{s.t.} \quad \mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v} \geq (1 - \varepsilon) \lambda_{\max}(\mathbf{\Sigma}_y). \tag{2}$$

For a unit-norm direction $\mathbf{v}$, the numerator is the mean squared out-of-fold prediction error of $\mathbf{v}^\top \mathbf{y}_i$ using $\mathbf{v}^\top \widehat{\mathbf{y}}_i$, while the denominator is the empirical variance of $\mathbf{v}^\top \mathbf{y}_i$ over $\mathcal{C}$. The variance-retention constraint requires this variance to be at least a fraction $1 - \varepsilon$ of the maximum attainable variance over unit-norm directions, preventing the selection of low-variance directions with small normalized prediction error but limited proxy variation.

The quotient in equation 2 is a generalized Rayleigh quotient and is structurally related to minimum-noise-fraction criteria (Green et al., 1988; Gupta & Bajaj, 2019). Unlike these criteria, $\mathbf{\Sigma}_e$ is induced by out-of-fold prediction from frozen semantic embeddings rather than by a prescribed noise process. The variance-retention constraint defines the admissible directions, while out-of-fold predictability distinguishes among them. With the variance-retention constraint, equation 2 does not in general reduce to a single generalized eigendecomposition, and the feasible set on the unit sphere is non-convex. We therefore use multi-start Sequential Least Squares Programming (SLSQP) (Kraft, 1988), initialized around the leading eigenvector of $\mathbf{\Sigma}_y$. This is a numerical solver rather than a general global-optimality certificate; however, Appendix A.4 shows that if the unconstrained generalized-eigenvector solution is feasible it is globally optimal.

Latent Impact Score and Latent Impact Inference. The PALIF objective and variance-retention constraint are invariant under $\mathbf{v} \mapsto -\mathbf{v}$, so the sign of the resulting direction $\mathbf{v}^\star$ is not uniquely determined. Using the citation-increment orientation reference r_i defined in Appendix D.1, we choose the sign so that the sample covariance $\widehat{\mathrm{Cov}}_{i \in \mathcal{C}}(\mathbf{v}^{\star\top} \mathbf{y}_i, r_i) \geq 0$ (with either sign allowed in the zero-covariance tie case). For each $i \in \mathcal{C}$, we then define $I_i = \mathbf{v}^{\star\top} \mathbf{y}_i$. Cross-fitting is used only to construct the out-of-fold residuals for factor selection. For inference, we fit a single proxy predictor $h_{\psi^\star}$ on the full set $\mathcal{C}$, using the same model architecture and training protocol as the fold-specific predictors and the squared-error loss in equation 1, with the summation taken over all $i \in \mathcal{C}$. For each $d \in \mathcal{D}$, we define the predicted latent impact score as $\widetilde{I}_d = \mathbf{v}^{\star\top} h_{\psi^\star}(\mathbf{u}_d)$. The prediction depends only on the frozen semantic representation $\mathbf{u}_d$ and requires no proxy observations for d.

3.2 Radial Impact Encoding and Retrieval

Corpus-Normalized Impact Score. For an indexed corpus $\mathcal{D}$ with $|\mathcal{D}| \geq 2$, let r_d denote the average rank of the predicted latent impact score $\widetilde{I}_d$ among $\{\widetilde{I}_j : j \in \mathcal{D}\}$, with larger scores assigned larger ranks. We define the corpus-normalized impact score used for retrieval as

$$\widehat{I}_d = \frac{r_d - 1}{|\mathcal{D}| - 1} \in [0, 1]. \tag{3}$$

This transformation preserves the ordering by $\widetilde{I}_d$, is invariant to strictly increasing score transformations, and is insensitive to extreme scores that can dominate min–max scaling. Most importantly, the guaranteed range $[0, 1]$ makes γ a fixed upper budget on radial rescaling: $\exp(\gamma \widehat{I}_d) \in [1, e^\gamma]$.

Radial Impact Encoding. Given $\widehat{I}_d$, we define the impact-aware document representation as

$$\mathbf{z}_d = \exp\left(\gamma \widehat{I}_d\right) \mathbf{u}_d, \qquad \mathbf{z}_q = \mathbf{u}_q, \qquad \gamma \geq 0. \tag{4}$$

The hyper-parameter γ controls the strength of document-side radial impact scaling. Setting $\gamma = 0$ recovers the base semantic retrieval model. Retrieval is performed by the inner product

$$S_\gamma(q, d) = \mathbf{z}_q^\top \mathbf{z}_d = \exp\left(\gamma \widehat{I}_d\right) \mathbf{u}_q^\top \mathbf{u}_d. \tag{5}$$

$\exp\left(\gamma \widehat{I}_d\right)$ preserves the sign of the base semantic score and increases its magnitude proportionally.

For any $d, d' \in \mathcal{D}$, the radial transformation preserves document directions and cosine similarities:

$$\frac{\mathbf{z}_d}{\|\mathbf{z}_d\|_2} = \mathbf{u}_d, \qquad \cos(\mathbf{z}_d, \mathbf{z}_{d'}) = \mathbf{u}_d^\top \mathbf{u}_{d'}.$$

Thus, latent impact modifies only the radial component of the document embedding while preserving the semantic angular geometry. Since equation 5 remains a single inner product between query and document vectors, RADAR is compatible with standard single-vector maximum-inner-product search (MIPS) (Shrivastava & Li, 2014; Guo et al., 2020).

4 THEORETICAL ANALYSIS

We establish three properties of RADAR's design choices: predictability, variance retention, and bounded radial perturbation. Lowercase subscripts ($\mathbf{\Sigma}_y$, $\mathbf{\Sigma}_e$) denote the empirical matrices of Section 3.1; uppercase subscripts denote population counterparts. Proofs are in Appendix A.

4.1 PALIF Targets Explainable Proxy Variance

Proposition 1 (Explainable variance and learned-predictor gap)**.** *Let $(\mathbf{U}, \mathbf{Y})$ be a population pair of frozen representations and centered proxy vectors, and define $\mathbf{m}(\mathbf{U}) = \mathbb{E}[\mathbf{Y} \mid \mathbf{U}]$, $\mathbf{\Sigma}_Y = \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$, $\mathbf{\Sigma}_M = \mathbb{E}[\mathbf{m}\mathbf{m}^\top]$, and $\mathbf{\Sigma}_E = \mathbb{E}[(\mathbf{Y} - \mathbf{m})(\mathbf{Y} - \mathbf{m})^\top]$. For every $\mathbf{v}$ with $\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v} > 0$, the fraction of the variance of $\mathbf{v}^\top \mathbf{Y}$ explainable from $\mathbf{U}$ is*

$$R^2_{\mathrm{Bayes}}(\mathbf{v}) = 1 - \frac{\mathbf{v}^\top \mathbf{\Sigma}_E \mathbf{v}}{\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v}} = \frac{\mathbf{v}^\top \mathbf{\Sigma}_M \mathbf{v}}{\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v}}. \tag{6}$$

For any square-integrable predictor $\mathbf{h}(\mathbf{U})$, let $\mathbf{\Sigma}_h = \mathbb{E}[(\mathbf{Y} - \mathbf{h})(\mathbf{Y} - \mathbf{h})^\top]$, $\boldsymbol{A}_h = \mathbb{E}[(\mathbf{m} - \mathbf{h})(\mathbf{m} - \mathbf{h})^\top]$, and $R^2_h(\mathbf{v}) = 1 - \mathbf{v}^\top \mathbf{\Sigma}_h \mathbf{v} / \mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v}$. Then

$$\mathbf{\Sigma}_h = \mathbf{\Sigma}_E + \boldsymbol{A}_h, \qquad R^2_h(\mathbf{v}) = R^2_{\mathrm{Bayes}}(\mathbf{v}) - \frac{\mathbf{v}^\top \boldsymbol{A}_h \mathbf{v}}{\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v}}, \tag{7}$$

and for all unit $\mathbf{v}$ with $\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v} \geq \tau > 0$,

$$0 \leq R^2_{\mathrm{Bayes}}(\mathbf{v}) - R^2_h(\mathbf{v}) \leq \mathbb{E}\|\mathbf{h}(\mathbf{U}) - \mathbf{m}(\mathbf{U})\|_2^2 / \tau. \tag{8}$$

The ratio in equation 2 equals $1 - \widehat{R}^2_{\mathrm{OOF}}(\mathbf{v})$, the cross-fitted analogue of $1 - R^2_h(\mathbf{v})$ (Appendix A.2), so PALIF selects the variance-retaining direction most predictable from the frozen representation. The constraint also bounds the learned-predictor gap: on the population feasible set, $\tau = (1 - \varepsilon)\lambda_1(\mathbf{\Sigma}_Y) \geq 1 - \varepsilon$ because standardized proxies give $\mathbf{\Sigma}_Y$ a unit diagonal, so the gap in equation 8 is at most $\mathbb{E}\|\mathbf{h} - \mathbf{m}\|_2^2 / (1 - \varepsilon)$. Without the constraint, τ can fall to $\lambda_M(\mathbf{\Sigma}_Y)$, which is small for strongly correlated proxies, letting predictor error along low-variance directions dominate selection. Variance retention thus both anchors the attribute (Section 4.2) and stabilizes the selection.

4.2 Variance Retention: Spectral Anchoring and Interpolation

Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_M \geq 0$ be the eigenvalues of $\mathbf{\Sigma}_y$ (so $\lambda_1 = \lambda_{\max}(\mathbf{\Sigma}_y)$) with orthonormal eigenvectors $\mathbf{b}_1, \ldots, \mathbf{b}_M$, and define the feasible set $\mathcal{A}_\varepsilon = \{\mathbf{v} : \|\mathbf{v}\|_2 = 1,\ \mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v} \geq (1 - \varepsilon)\lambda_1\}$.

Proposition 2 (Sharp spectral anchoring). *For every $\varepsilon \in [0, 1)$,*

$$\max_{\mathbf{v} \in \mathcal{A}_\varepsilon} \sin^2\theta(\mathbf{v}, \mathbf{b}_1) = \min\left\{1, \frac{\varepsilon\lambda_1}{\lambda_1 - \lambda_2}\right\}, \qquad \theta(\mathbf{v}, \mathbf{b}_1) = \arccos|\mathbf{v}^\top \mathbf{b}_1|, \tag{9}$$

and the maximum is attained.

Hence ε controls proxy fidelity jointly with the eigengap; e.g., for two standardized proxies with correlation $\rho = 0.5$ and $\varepsilon = 0.03$, every admissible direction lies within about 12° of $\mathbf{b}_1$. Appendix Proposition 6 extends the guarantee to dominant r-dimensional eigenspaces. The endpoints of the ε range recover classical methods.

Corollary 3 (PALIF interpolates PCA and CCA). *Assume $\lambda_1 > \lambda_2$ and $\mathbf{\Sigma}_y \succ 0$. (i) At $\varepsilon = 0$, the PALIF solution is $\pm\mathbf{b}_1$, the leading principal direction. (ii) If the out-of-fold predictions are replaced by in-sample affine least-squares predictions of $\mathbf{y}_i$ from $\mathbf{u}_i$, the PALIF ratio equals $1 - \rho^2(\mathbf{v})$, where $\rho(\mathbf{v})$ is the sample multiple correlation between $\mathbf{v}^\top\mathbf{y}_i$ and $\mathbf{u}_i$; for all ε sufficiently close to 1, the PALIF solution is a leading canonical direction of $(\mathbf{y}, \mathbf{u})$.*

PALIF's cross-fitted nonlinear predictor extends this spectrum to out-of-sample predictability.

4.3 Radial Encoding Is a Bounded Reranking Rule

For a fixed query q, write $c_d = \mathbf{u}_q^\top \mathbf{u}_d$ and $S_\gamma(d) = S_\gamma(q, d) = e^{\gamma\widehat{I}_d} c_d$. We call any $\{\widehat{I}_d\}_{d\in\mathcal{D}} \subset [0, 1]$ admissible; the results below hold for all admissible assignments, regardless of predictor quality or corpus. For $c_d > 0$, ranking by S_γ equals ranking by $\log c_d + \gamma\widehat{I}_d$, combining log-relevance with a query-independent document prior as in classical IR (Craswell et al., 2005), yet as a single inner product.

Proposition 4 (Semantic-margin bound). *For every document, $\operatorname{sign}(S_\gamma(d)) = \operatorname{sign}(c_d)$. If $c_d > c_{d'} > 0$, their order reverses if and only if*

$$0 < \log\frac{c_d}{c_{d'}} < \gamma(\widehat{I}_{d'} - \widehat{I}_d). \tag{10}$$

Therefore no positive pair with $c_d/c_{d'} \geq e^\gamma$ can reverse, and the threshold e^γ is tight.

Corollary 5 (Top-k stability). *Let $c_{(1)} \geq \cdots \geq c_{(|\mathcal{D}|)}$ be the ordered base scores and suppose $c_{(k)} > 0$. A document outside the base top-k can enter RADAR's top-k only if*

$$c_d \geq e^{-\gamma} c_{(k)}. \tag{11}$$

The entire base top-k set is invariant for every admissible impact assignment if $c_{(k+1)} \leq 0$, or if

$$c_{(k)}/c_{(k+1)} > e^\gamma. \tag{12}$$

The band is tight: if $c_d > e^{-\gamma}c_{(k)}$, setting $\widehat{I}_d = 1$ and all other impacts to zero places d in the top-k. Thus γ is an explicit semantic-perturbation budget. At $\gamma = 0.05$, no positive pair whose base scores differ by more than about 5.1% can reverse, and only documents within about 4.9% of the k-th base score can newly enter the top-k; sign preservation further prevents impact from promoting negatively aligned candidates. Radial encoding is therefore a controlled adjustment, not an unrestricted reranker.

5 Experiments

5.1 Experimental Setup

Datasets We use separate datasets for retrieval evaluation, latent-impact learning, and impact validation. For retrieval evaluation, we use four scientific literature benchmarks. MasterSet (Ratul et al., 2026) defines must-cite positives as papers corresponding to experimental baselines, methods, or datasets on which the query paper directly builds. RMC (Long et al., 2024) uses papers explicitly identified by reviewers as missing citations. We additionally construct SuReIR and GSRR. SuReIR uses survey and review papers as queries and papers cited in their bibliographies as positives, while GSRR uses scientific papers as queries and Google Scholar exact-title-search rankings as retrieval judgments. Further details on SuReIR and GSRR are provided in Appendix C. Latent-impact models are trained on OpenAlex using citation increments and historical venue assessments as scholarly proxies. We use the SciRepEval Peer Review Score task (Singh et al., 2023) for external validation.

Table 1: **Retrieval performance on four real-world scientific retrieval benchmarks.** We report nDCG@10 and Recall@50, both in %. The Avg. columns report the unweighted macro-average over the four benchmarks. Best results are in **bold**; second-best results are underlined.

	SuReIR		GSRR		RMC		MasterSet		Avg.	
Baseline	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**
Sparse / Dense Embedding										
BM25	50.82	29.30	72.72	68.03	12.89	31.85	13.13	24.82	37.39	38.50
SPECTER	49.51	27.74	60.19	59.90	10.29	30.15	10.62	22.76	32.65	35.14
SciNCL	52.87	30.57	69.00	68.17	12.39	34.51	11.58	25.10	36.46	39.59
SPECTER2	54.15	30.36	69.51	69.93	13.10	35.49	12.29	25.69	37.26	40.37
E5-large-v2	51.67	30.19	82.82	80.50	12.37	32.53	12.73	25.11	39.90	42.08
BGE-M3	51.85	27.13	75.50	74.17	11.91	29.41	12.07	23.19	37.83	38.48
NV-Embed-v2	56.92	33.59	81.60	77.97	12.30	39.44	15.15	31.39	41.49	45.60
Qwen3-Embedding-8B	58.41	34.41	84.92	83.20	18.38	47.04	18.93	35.37	45.16	50.01
Reranking										
BGE-reranker-v2-m3	52.52	30.33	81.87	82.53	13.22	36.76	12.30	32.97	39.98	45.65
Qwen3-Reranker-8B	58.47	34.11	84.61	82.67	19.35	48.10	19.21	35.45	45.41	50.08
Impact-Aware Methods (Qwen3-Embedding-8B)										
+CitationPred	57.98	34.31	85.03	82.93	19.38	47.90	19.33	36.38	45.43	50.38
+VenuePred	58.36	34.37	85.07	83.17	19.22	48.14	18.98	35.47	45.41	50.29
+MultiProxyPred	58.11	34.31	84.98	82.93	19.28	48.06	19.29	36.35	45.42	50.41
+FactorDistill	57.80	34.05	84.63	82.73	19.24	**48.29**	19.39	36.43	45.27	50.38
RADAR (Ours)	**59.99**	**36.54**	**85.11**	**83.37**	**19.54**	48.27	**19.45**	**37.04**	**46.02**	**51.31**

Evaluation Metrics. For retrieval evaluation, we report nDCG@10 and Recall@50, measuring top-ranked retrieval quality and relevant-paper coverage, respectively. For external validation, we report tie-aware Kendall's τ to measure rank agreement between predicted impact scores and held-out peer-review judgments.

Baselines. We evaluate 14 baselines in five categories. **Sparse retrieval:** BM25 (Robertson & Zaragoza, 2009), a non-neural lexical baseline. **Scientific-domain dense embeddings:** SPECTER (Cohan et al., 2020), SciNCL (Ostendorff et al., 2022), and SPECTER2 (Singh et al., 2023), which use scientific-domain supervision to learn document representations. **General-purpose dense embeddings:** E5-large-v2 (Wang et al., 2022), BGE-M3 (Chen et al., 2024), NV-Embed-v2 (Lee et al., 2025), and Qwen3-Embedding-8B (Zhang et al., 2025), spanning encoder-based and LLM-based embedding families. **Rerankers:** BGE-reranker-v2-m3 (Li et al., 2023) and Qwen3-Reranker-8B (Zhang et al., 2025), which rescore candidates from a first-stage retriever. **Impact-aware methods:** CitationPred and VenuePred use citation- and venue-based proxy targets, respectively; MultiProxyPred fits the two predictors independently, standardizes their outputs using training-set statistics, and averages the resulting scores with equal weights; FactorDistill predicts a latent-factor score obtained from a classical linear factor model. All four impact-aware baselines use the same frozen Qwen3-Embedding-8B backbone and retrieval scoring rule as RADAR.

5.2 RESULT OVERVIEW

Table 1 summarizes retrieval performance across the four scientific retrieval benchmarks. RADAR achieves the highest macro-average on both metrics, with 46.02 nDCG@10 and 51.31 Recall@50. It ranks first in seven of the eight dataset–metric pairs and second only on RMC Recall@50, where FactorDistill leads by 0.02 points. Relative to the underlying Qwen3-Embedding-8B retriever, RADAR improves average nDCG@10 by 0.86 points and Recall@50 by 1.30 points, with gains on both metrics across all four benchmarks. These results indicate that incorporating predicted latent impact provides additional ranking information beyond the frozen semantic representation.

The impact-aware baselines provide a controlled comparison because they share the same frozen Qwen3-Embedding-8B backbone while differing in how the document-level impact score is learned. Among these alternatives, CitationPred obtains the highest average nDCG@10 (45.43), while MultiProxyPred obtains the highest average Recall@50 (50.41). RADAR improves over these best baseline scores by 0.59 and 0.90 points, respectively. None of the single-proxy, multi-proxy averaging, or classical factor-distillation alternatives matches RADAR's average retrieval performance, favoring the proposed latent impact formulation over these simpler ways of constructing an impact score.

Table 2: Ablation for RADAR's components: proxy families, PALIF factor selection, and radial scaling. Removing each component degrades performance, with direct impact scaling yielding the largest drop.

	nDCG@10	Recall@50
RADAR	**46.02%**	**51.31%**
- Citation Proxies	45.40% (↓ 0.62)	50.36% (↓ 0.95)
- Venue Proxies	45.32% (↓ 0.70)	50.58% (↓ 0.73)
- Text Predictability	45.13% (↓ 0.89)	50.35% (↓ 0.96)
- Variance Retention	44.90% (↓ 1.12)	50.12% (↓ 1.19)
- Exponential Scaling	30.72% (↓ 15.30)	29.18% (↓ 22.13)

Table 3: Rank association between predicted impact and peer-review scores on SciRepEval.

Method	**Kendall's τ ↑**
CitationPred	0.0834 ± 0.0084
VenuePred	0.1425 ± 0.0015
MultiProxyPred	0.1379 ± 0.0052
FactorDistill	0.1371
RADAR	$\mathbf{0.1504 \pm 0.0016}$

RADAR also exceeds the stronger reranking baseline, Qwen3-Reranker-8B, by 0.61 nDCG@10 and 1.23 Recall@50 points on average. Whereas the reranker rescores the top-100 candidates returned by Qwen3-Embedding-8B, RADAR retrieves directly from the full corpus through a single-vector inner product. Thus, RADAR outperforms the computationally expensive cross-encoder reranker without reranking, while avoiding the recall ceiling imposed by the top-100 candidate pool.

5.3 Ablation Study

We perform an ablation study to evaluate each component's contribution. Following the protocol in Section 5.1, we (i) remove the citation proxy family, (ii) remove the venue proxy family, (iii) replace PALIF's text-predictability criterion with the maximum-variance proxy direction, (iv) remove its variance-retention constraint, and (v) replace exponential radial scaling with direct impact scaling. Table 2 shows that every ablation reduces average retrieval performance. Removing either proxy family or either PALIF component lowers both nDCG@10 and Recall@50, while direct impact scaling yields the largest degradation.

5.4 Latent Impact Validation

To assess the measurement validity of the learned latent impact, we evaluate its rank association with peer-review scores on the SciRepEval Peer Review Score task (Singh et al., 2023). We report Kendall's τ between predicted impact scores and peer-review judgments.

While the absolute correlation magnitude is inherently capped by the well-documented high variance and subjectivity of human peer-review scores (Cortes & Lawrence, 2021; Beygelzimer et al., 2023), Table 3 shows that RADAR yields the highest relative rank association among the tested methods.

We additionally assess prospective validity on a chronologically held-out OpenAlex cohort using subsequent three-year citation outcomes. RADAR leads on three of four prospective metrics, with full evaluation details and results reported in Appendix E.2.

6 Conclusion and Discussion

We introduced RADAR, a general framework for augmenting pretrained embeddings with latent document quality learned from heterogeneous noisy proxies. Two principles guide its design: predictability from the frozen representation informs attribute selection, while variance retention keeps the attribute supported by substantial proxy variation. RADAR encodes the predicted attribute solely through radial scaling, preserving semantic directions and compatibility with single-vector maximum-inner-product retrieval. It requires neither retraining the embedding nor proxy observations for corpus documents. In scientific retrieval, citation and venue signals provide a testbed for latent impact distillation. Across four benchmarks, RADAR improves on the underlying semantic representation and competitive impact-aware alternatives, while its distilled attribute agrees more strongly with held-out peer-review judgments and prospective citation outcomes on most measures. These results show that radial scaling can incorporate critical latent targets without sacrificing semantic structure or retrieval efficiency. The proposed framework can be generalized to image embeddings and other modalities, and evaluating it in other modalities remains future work.

AI USE STATEMENT

In this work, we used generative AI tools for three purposes. First, to aid and polish writing: editing text for readability and formatting references. Second, for retrieval and discovery: searching for and summarizing related literature. Third, to assist with proving mathematical claims: suggesting proof strategies, supplying candidate intermediate steps, and checking algebraic manipulations in the propositions and their appendix proofs.

We did not use generative AI tools to generate synthetic datasets or to interpret the reported results. The research questions, the theoretical framework, the method, the experimental design, and all empirical conclusions are the authors' own.

We have reviewed all AI-assisted work. Every proposition and proof was independently verified by the authors. Every reference identified by an AI tool was checked against its original published source for existence, authorship, venue, and the accuracy of the claim attributed to it. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The theoretical assumptions are stated in Section 4, with supporting proofs and additional analysis provided in Appendix A. The construction of the two benchmarks we introduce, including query selection, corpus sampling, and relevance judgments, is documented in Appendix C. The training corpus, proxy definitions, stratification, and main training and retrieval settings are described in Appendix D. Code and both benchmarks will be made publicly available upon acceptance.

REFERENCES

Philipp Bach, Malte S. Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56, 2024. doi: 10.18637/jss.v108.i03.

Michael Bendersky, W Bruce Croft, and Yanlei Diao. Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 95–104, 2011.

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? The NeurIPS 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.

C. I. Bliss, Mary L. Greenwood, and Edna Sakamoto White. A rankit analysis of paired comparisons for measuring the effect of sprays on flavor. *Biometrics*, 12(4):381–403, 1956. doi: 10.2307/3001679.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5), 2024.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2270–2282, 2020.

Terence Cooke and Roszaini Haniffa. The development of inverse normal transformations: A review. *IMA Journal of Management Mathematics*, 36(4):991–1010, 2025. doi: 10.1093/imaman/dpaf024.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Nick Craswell, Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 416–423, 2005.

Andrew A. Green, Mark Berman, Paul Switzer, and Maurice D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74, 1988. doi: 10.1109/36.3001.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3887–3896. PMLR, 2020.

Soumyajit Gupta and Chandrajit Bajaj. A streaming model for generalized Rayleigh with extension to minimum noise fraction. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 74–83, 2019. doi: 10.1109/BigData47090.2019.9006512.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202, 2016.

Karl G Jöreskog and Arthur S Goldberger. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American statistical Association*, 70 (351a):631–639, 1975.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Dain Kim, Jinhyeok Park, and Dongwoo Kim. Test-time embedding normalization for popularity bias mitigation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 4023–4027, 2023.

Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*, 3, 2023.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. A neural citation count prediction model based on peer review text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4914–4924, 2019.

Kehan Long, Shasha Li, Pancheng Wang, Chenlong Bao, Jintao Tang, and Ting Wang. Recommending missed citations identified by reviewers: A new task, dataset and baselines. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13699–13711, 2024.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Expansion via prediction of importance with contextualization. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 1573–1576, 2020.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 11670–11688, 2022.

Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2108–2130. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.131. URL https://aclanthology.org/2023.emnlp-main.131/.

Alexander J Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016.

Md Toyaha Rahman Ratul, Zhiqian Chen, Kaiqun Fu, Taoran Ji, and Lei Zhang. Masterset: A large-scale benchmark for must-cite citation recommendation in the ai/ml literature. *arXiv preprint arXiv:2604.17680*, 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-end weak supervision. *Advances in Neural Information Processing Systems*, 34:1845–1857, 2021.

Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *Advances in Neural Information Processing Systems*, volume 27, pp. 2321–2329, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/c98e7c4b8f20d384e3ad857d0ee226cc-Abstract.html.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 5548–5566, 2023.

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 preprint arXiv:2212.03533*, 2022.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. From words to worth: Newborn article impact prediction with llm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1183–1191, 2025.

Jianqun Zhou, Yuanlei Zheng, Wei Chen, Qianqian Zheng, Shang Zeyuan, Wei Zhang, Rui Meng, and Xiaoyu Shen. Beyond content relevance: Evaluating instruction following in retrieval models. In *International Conference on Learning Representations*, volume 2025, pp. 84965–84996, 2025.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10856–10896, 2025.

A Theoretical Proofs for Section 4

A.1 Proof of Proposition 1

Proof. Because $\mathbb{E}[\mathbf{Y}] = \mathbf{0}$, the tower property gives $\mathbb{E}[\mathbf{m}(\mathbf{U})] = \mathbf{0}$. Conditional expectation is the L_2 projection onto functions of $\mathbf{U}$, so for $T_{\mathbf{v}} = \mathbf{v}^\top \mathbf{Y}$,

$$\mathbb{E}[T_{\mathbf{v}} \mid \mathbf{U}] = \mathbf{v}^\top \mathbf{m}(\mathbf{U}),$$

and the minimum mean-squared prediction error is

$$\inf_g \mathbb{E}[(T_{\mathbf{v}} - g(\mathbf{U}))^2] = \mathbf{v}^\top \mathbf{\Sigma}_E \mathbf{v}.$$

The cross term in $\mathbf{Y} = \mathbf{m}(\mathbf{U}) + (\mathbf{Y} - \mathbf{m}(\mathbf{U}))$ vanishes because

$$\mathbb{E}[\mathbf{Y} - \mathbf{m}(\mathbf{U}) \mid \mathbf{U}] = \mathbf{0}.$$

Hence $\mathbf{\Sigma}_Y = \mathbf{\Sigma}_M + \mathbf{\Sigma}_E$, which yields equation 6.

For an arbitrary predictor $\mathbf{h}(\mathbf{U})$, write

$$\mathbf{Y} - \mathbf{h} = (\mathbf{Y} - \mathbf{m}) + (\mathbf{m} - \mathbf{h}).$$

The two terms are orthogonal in L_2 because the second is measurable with respect to $\mathbf{U}$ and the first has conditional mean zero. Therefore

$$\mathbf{\Sigma}_h = \mathbf{\Sigma}_E + \boldsymbol{A}_h,$$

which gives equation 7. On a set of unit vectors with $\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v} \geq \tau$,

$$0 \leq \frac{\mathbf{v}^\top \boldsymbol{A}_h \mathbf{v}}{\mathbf{v}^\top \mathbf{\Sigma}_Y \mathbf{v}} \leq \frac{\|\boldsymbol{A}_h\|_{\mathrm{op}}}{\tau} \leq \frac{\mathrm{tr}(\boldsymbol{A}_h)}{\tau} = \frac{\mathbb{E}\|\mathbf{h}(\mathbf{U}) - \mathbf{m}(\mathbf{U})\|_2^2}{\tau},$$

proving equation 8.

If $\mathbf{\Sigma}_Y \succ 0$ and the variance-retention constraint is removed, equation 6 is a generalized Rayleigh quotient, so its maximizers are leading generalized eigenvectors of $(\mathbf{\Sigma}_M, \mathbf{\Sigma}_Y)$, equivalently minimizers of the quotient associated with $(\mathbf{\Sigma}_E, \mathbf{\Sigma}_Y)$. □

A.2 Finite-Sample Cross-Fitted Identity and Consistency

For $i \in \mathcal{C}$, let $t_i(\mathbf{v}) = \mathbf{v}^\top \mathbf{y}_i$, $\widehat{t}_i(\mathbf{v}) = \mathbf{v}^\top \widehat{\mathbf{y}}_i$, and $\mathbf{e}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i$. Then

$$\frac{\mathbf{v}^\top \mathbf{\Sigma}_e \mathbf{v}}{\mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v}} = \frac{\sum_{i \in \mathcal{C}} (t_i(\mathbf{v}) - \widehat{t}_i(\mathbf{v}))^2}{\sum_{i \in \mathcal{C}} t_i(\mathbf{v})^2} = 1 - R^2_{\mathrm{OOF}}(\mathbf{v}),$$

where the last equality uses empirical centering of the harmonized proxies. Thus equation 2 exactly maximizes cross-fitted empirical R^2 over the admissible directions. Conditional on each fold-trained predictor, its held-out residuals evaluate that predictor on observations excluded from its fit; the resulting maximum is a model-selection criterion, not an independent post-selection estimate of generalization R^2.

Consistency remark. If the proxy harmonization and covariance estimators are consistent, $\widehat{\mathbf{\Sigma}}_y \to \mathbf{\Sigma}_Y$ in operator norm, and the fold-wise predictors converge to $\mathbf{m}$ in L_2, then Proposition 1 gives uniform convergence of the learned-predictor criterion on variance-retaining sets. Operator- norm convergence also implies convergence of the leading eigenvalue and, away from a binding spectral degeneracy, convergence of the feasible sets induced by the estimated threshold $(1 - \varepsilon)\widehat{\lambda}_1$. If the constrained population optimizer is unique up to sign and well separated, standard argmin continuity then yields consistency of the selected direction. Explicit finite-sample rates would additionally depend on the complexity of the chosen predictor class and preprocessing estimators and are outside the scope of this paper.

A.3 Proof of Proposition 2 and Subspace Extension

Proposition 6 (Dominant-subspace extension). *Let $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$ be the eigenvalues of $\mathbf{\Sigma}_y$ with orthonormal eigenvectors $\{\mathbf{b}_j\}_{j=1}^M$. For $\mathcal{B}_r = \operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_r\}$, assume $\lambda_r > \lambda_{r+1}$. Then*

$$\max_{\mathbf{v} \in \mathcal{A}_\varepsilon} \|P_{\mathcal{B}_r^\perp} \mathbf{v}\|_2^2 = \min\left\{1, \frac{\varepsilon \lambda_1}{\lambda_1 - \lambda_{r+1}}\right\}. \tag{13}$$

The numerical bound itself requires only $\lambda_1 > \lambda_{r+1}$; the eigengap $\lambda_r > \lambda_{r+1}$ makes $\mathcal{B}_r$ uniquely defined.

Proof of Proposition 6. Write $\mathbf{v} = \sum_{j=1}^M a_j \mathbf{b}_j$ with $\sum_j a_j^2 = 1$ and let $s = \sum_{j=r+1}^M a_j^2 = \|P_{\mathcal{B}_r^\perp} \mathbf{v}\|_2^2$. Then

$$\mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v} \leq \lambda_1(1-s) + \lambda_{r+1} s = \lambda_1 - (\lambda_1 - \lambda_{r+1})s.$$

For $\mathbf{v} \in \mathcal{A}_\varepsilon$, this implies

$$s \leq \frac{\varepsilon \lambda_1}{\lambda_1 - \lambda_{r+1}},$$

together with the trivial bound $s \leq 1$.

If $\varepsilon \lambda_1 \leq \lambda_1 - \lambda_{r+1}$, set

$$s_\varepsilon = \frac{\varepsilon \lambda_1}{\lambda_1 - \lambda_{r+1}}, \qquad \widetilde{\mathbf{v}} = \sqrt{1 - s_\varepsilon}\, \mathbf{b}_1 + \sqrt{s_\varepsilon}\, \mathbf{b}_{r+1}.$$

Then $\widetilde{\mathbf{v}}^\top \mathbf{\Sigma}_y \widetilde{\mathbf{v}} = (1-\varepsilon)\lambda_1$ and the bound is attained. If instead $\varepsilon \lambda_1 \geq \lambda_1 - \lambda_{r+1}$, then $\mathbf{b}_{r+1} \in \mathcal{A}_\varepsilon$ and attains $s = 1$. Thus equation 13 is sharp. □

Proposition 2 is the case $r = 1$, since $\|P_{\mathcal{B}_1^\perp} \mathbf{v}\|_2^2 = 1 - (\mathbf{v}^\top \mathbf{b}_1)^2 = \sin^2 \theta(\mathbf{v}, \mathbf{b}_1)$.

Additional feasible-set properties. If $\lambda_1 > \lambda_2$, then $\mathcal{A}_0 = \{\mathbf{b}_1, -\mathbf{b}_1\}$ and $\mathcal{A}_{\varepsilon_1} \subseteq \mathcal{A}_{\varepsilon_2}$ for $0 \leq \varepsilon_1 \leq \varepsilon_2 < 1$. Each feasible set is nonempty and compact, the PALIF denominator is bounded below by $(1-\varepsilon)\lambda_1 > 0$, and therefore a constrained optimum exists. Feasible-set nesting implies that the optimal PALIF objective cannot increase as ε increases; this does not imply monotonic downstream retrieval performance.

A.4 A Solver Characterization

Proposition 7 (Interior solutions of the constrained quotient). *Assume $\mathbf{\Sigma}_y \succ 0$ and let $q(\mathbf{v}) = \mathbf{v}^\top \mathbf{\Sigma}_e \mathbf{v} / (\mathbf{v}^\top \mathbf{\Sigma}_y \mathbf{v})$. If a generalized eigenvector associated with the smallest generalized eigenvalue of $(\mathbf{\Sigma}_e, \mathbf{\Sigma}_y)$ is feasible for PALIF, it is a global constrained minimizer. If no such generalized eigenvector is feasible, every global constrained minimizer satisfies the variance constraint with equality.*

Proof. The first statement is immediate because a global minimizer of q over all nonzero vectors remains optimal over any subset that contains it. For the second, suppose a constrained global minimizer $\mathbf{v}^\star$ has strict variance slack. It is then a local minimizer of q on the unit sphere. Under $\mathbf{\Sigma}_y \succ 0$, the map

$$\mathbf{x} = \frac{\mathbf{\Sigma}_y^{1/2} \mathbf{v}}{\|\mathbf{\Sigma}_y^{1/2} \mathbf{v}\|_2}$$

is a continuous bijection of the unit sphere with a continuous inverse, and $q(\mathbf{v}) = \mathbf{x}^\top (\mathbf{\Sigma}_y^{-1/2} \mathbf{\Sigma}_e \mathbf{\Sigma}_y^{-1/2}) \mathbf{x}$. Thus a local minimum of q corresponds to a local minimum of an ordinary Rayleigh quotient. Such a quotient has no nonglobal local minima: its local minima are exactly the eigenvectors associated with its smallest eigenvalue. Hence $\mathbf{v}^\star$ must be a smallest generalized eigenvector, contrary to the assumption that none is feasible. Therefore any constrained global minimizer must lie on the active variance boundary. □

A.5 Proof of Proposition 4

Proof. Since $S_\gamma(d) = e^{\gamma \widehat{I}_d} c_d$ and the exponential factor is positive, sign is preserved. For $c_d, c_{d'} > 0$,

$$\log \frac{S_\gamma(d)}{S_\gamma(d')} = \log \frac{c_d}{c_{d'}} + \gamma(\widehat{I}_d - \widehat{I}_{d'}).$$

If $c_d > c_{d'}$, strict reversal is therefore equivalent to equation 10. Since $\widehat{I}_{d'} - \widehat{I}_d \leq 1$, reversal requires $c_d/c_{d'} < e^\gamma$. Tightness follows by taking any $\kappa \in (1, e^\gamma)$, setting $c_d/c_{d'} = \kappa$, $\widehat{I}_d = 0$, and $\widehat{I}_{d'} = 1$. □

A.6 Proof of Corollary 5

Proof. For every positive base score,

$$c_d \leq S_\gamma(d) \leq e^\gamma c_d.$$

All k base top-k documents therefore have RADAR score at least $c_{(k)}$. If an outside document satisfies $c_d < e^{-\gamma} c_{(k)}$, then $S_\gamma(d) < c_{(k)}$ and it cannot enter, proving equation 11.

If $c_{(k+1)} \leq 0$, sign preservation separates every positive base top-k score from every nonpositive outside score. Otherwise, when $c_{(k+1)} > 0$ and $c_{(k)}/c_{(k+1)} > e^\gamma$, every original top-$k$ document has RADAR score at least $c_{(k)}$, while every outside document has score at most $e^\gamma c_{(k+1)} < c_{(k)}$. Hence the entire top-$k$ set is invariant. □

A stronger document-level sufficient condition also follows from the same envelope: a document d with $c_d > 0$ is guaranteed to remain in the RADAR top-k whenever fewer than k other documents satisfy $c_j \geq e^{-\gamma} c_d$.

B RADAR Algorithm

Algorithm 1 summarizes the complete RADAR pipeline. Proxy observations are used only during latent-attribute learning. Corpus-level attribute prediction, rank normalization, and radial scaling are performed offline before indexing using only frozen document embeddings. Once the augmented document representations are constructed, online retrieval requires only standard single-vector maximum-inner-product search.

C Dataset Construction

We construct two evaluation-only benchmarks, SuReIR and GSRR. Table 4 summarizes their main statistics.

Table 4: Summary statistics for SuReIR and GSRR.

Statistic	**SuReIR**	**GSRR**
Queries	103	200
Validation / Test	20 / 83	50 / 150
Corpus families	29,421	20,000
Positive pairs	10,007	4,000
Query representation	Title + abstract	Title
Relevance	Graded (1/2)	Binary

C.1 SuReIR

SuReIR uses 103 manually verified computer-science surveys and reviews published between 2020 and 2025 as queries, spanning 10 subfields. Each query is represented by its title and abstract. The query set contains 20 validation and 83 test queries.

Algorithm 1 RADAR: Radial Attribute Distillation for Augmented Representations

Require: Frozen training embeddings $\{\mathbf{u}_i\}_{i=1}^N$; proxy observations $\{(x_{im}, \omega_{im}) : i = 1, \ldots, N,\ m = 1, \ldots, M\}$; strata $\{g_i\}_{i=1}^N$; an application-specific orientation reference; frozen corpus embeddings $\{\mathbf{u}_d : d \in \mathcal{D}\}$, with $|\mathcal{D}| \geq 2$; number of folds K; Rankit clipping parameter δ; variance-retention parameter ε; radial scale γ.

Ensure: Augmented corpus representations $\{\mathbf{z}_d : d \in \mathcal{D}\}$.

Stage I: Latent Attribute Distillation

Apply the stratum-wise rank-based inverse normal transformation to each proxy using all available observations within each stratum, obtaining $\widetilde{x}_{im}$.
Form the complete-case subset $\mathcal{C} = \{i : \omega_{im} = 1,\ \forall m = 1, \ldots, M\}$.
Center and standardize each transformed proxy over $\mathcal{C}$ to construct the harmonized proxy vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{iM})^\top$.
Partition $\mathcal{C}$ into K folds $\{\mathcal{F}_k\}_{k=1}^K$.
for $k = 1, \ldots, K$ **do**
    Fit the proxy predictor h_{ψ_k} on $\mathcal{C} \setminus \mathcal{F}_k$ using the squared-error objective in equation 1.
    For all $i \in \mathcal{F}_k$, obtain the out-of-fold predictions $\widehat{\mathbf{y}}_i = h_{\psi_k}(\mathbf{u}_i)$ and residuals $\mathbf{e}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i$.
end for
Compute $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_e$ from $\{\mathbf{y}_i, \mathbf{e}_i : i \in \mathcal{C}\}$ as defined in Section 3.1.
Solve equation 2 with multi-start SLSQP to obtain the factor direction $\mathbf{v}^\star$;
Choose the sign of $\mathbf{v}^\star$ so that its projected factor scores have nonnegative sample covariance with the designated orientation reference.
In the scholarly-impact instantiation, citation increments serve as the orientation reference.
Refit a single proxy predictor $h_{\psi^\star}$ from scratch on all $i \in \mathcal{C}$, using the same architecture, training protocol, and squared-error objective as in cross-fitting.

Stage II: Offline Corpus Augmentation

for $d \in \mathcal{D}$ **do**
    Compute the predicted latent impact score from the frozen embedding: $\widetilde{I}_d \leftarrow \mathbf{v}^{\star\top} h_{\psi^\star}(\mathbf{u}_d)$.
end for
Compute the average rank r_d of each $\widetilde{I}_d$ among $\{\widetilde{I}_j : j \in \mathcal{D}\}$, assigning larger ranks to larger scores.
for $d \in \mathcal{D}$ **do**
    Compute the corpus-normalized score $\widehat{I}_d \leftarrow (r_d - 1)/(|\mathcal{D}| - 1)$.
    Construct the RADAR document representation $\mathbf{z}_d \leftarrow \exp(\gamma \widehat{I}_d)\mathbf{u}_d$.
end for
return $\{\mathbf{z}_d : d \in \mathcal{D}\}$.

Documents are grouped into canonical paper families to merge different versions of the same work. The corpus contains 29,421 distinct families, combining 20,000 background families sampled from OpenAlex across computer-science subfields and publication periods with 9,449 families appearing in the query bibliographies; 28 families occur in both sets and are counted once.

A bibliography paper is retained as a positive if it can be mapped to a corpus family, was available before the query paper's cutoff date, and is not the query paper itself. All cited positives receive a base relevance grade of 1. A positive is upgraded to grade 2 if its citation context contains a predefined cue of foundational status or explicit emphasis (e.g., "seminal," "first proposed," or "widely adopted") that refers to the cited work or its contribution and can be unambiguously attributed to its canonical paper family, or if the family has at least three reliably resolved in-text citation occurrences. SuReIR is evaluated using graded nDCG@10 and Recall@50.

C.2 GSRR

GSRR contains 200 computer-science paper queries published between 2020 and 2024, balanced across 10 subfields and five publication years. We select four queries from each subfield–year cell using a deterministic, model-independent sampling procedure, assigning one to validation and three to test. Each query is represented by its exact full title.

For each query, we search Google Scholar using the full title under its default relevance ordering. The query paper itself and alternative versions of the same work are removed, and duplicate re-

sults are merged into canonical paper families. We retain the first 20 unique non-focal families as positives.

All queries share a 20,000-family corpus containing the 3,992 unique positive families and 16,008 background families sampled from OpenAlex. Candidate papers are represented by title and abstract when available; 613 of the 20,000 families use a title-only fallback because no reliable abstract is available.

GSRR uses binary relevance: membership in the captured Google Scholar Top-20 defines a positive judgment. The original Scholar ordering is retained as auxiliary information but is not used to define graded relevance. We evaluate GSRR using nDCG@10 and Recall@50.

D IMPLEMENTATION DETAILS

D.1 TRAINING CORPUS AND PROXY CONSTRUCTION

We construct the proxy-learning corpus from OpenAlex. We retain English computer science papers with valid abstracts, published either as conference papers or as articles associated with conference venues, and exclude retracted works. For latent-attribute learning, we use papers published between 2010 and 2017 and remove three papers overlapping with the SciRepEval test set, yielding 105,857 training papers. The latent-impact training set is disjoint from all four retrieval corpora; we verified that no paper family appears in both.

We use three proxies for scholarly impact. The first, $C_i^{(0,1]}$, counts citations from papers published one year after paper i, while $C_i^{(1,3]}$ counts citations from papers published two or three years later. Citations from the publication year itself are excluded from both windows. The third proxy, V_i, represents historical venue standing using the most recent CORE/ERA category available at the time paper i was published. We map $\{\text{Unranked}, \text{C}, \text{B}, \text{A}, \text{A}^*\}$ to $\{0, 1, 2, 3, 4\}$ and divide by four, yielding a venue proxy in $[0, 1]$.

For proxy harmonization, each paper is assigned to a stratum defined by the Cartesian product of its OpenAlex primary field, publication year, and a coarse semantic topic estimated from the training data only. We use 500 train-only Qwen semantic topics, resulting in 3,989 observed strata in the final training set. As described in Section 3.1, the rank transformation for each proxy uses all available observations within its corresponding stratum, whereas PALIF operates on the complete-case subset.

Finally, we resolve the sign ambiguity of the PALIF direction using a citation-increment orientation reference.

$$r_i = \log\left(1 + C_i^{(0,1]}\right) + \log\left(1 + C_i^{(1,3]}\right).$$

The sign of $\mathbf{v}^\star$ is chosen so that the sample covariance between $\mathbf{v}^{\star\top}\mathbf{y}_i$ and r_i is nonnegative, ensuring that larger latent-impact scores are oriented toward stronger citation evidence. All proxy transformations, semantic-topic assignments, cross-fitted predictions, covariance estimates, PALIF factor selection, and orientation statistics are estimated using the training data only.

D.2 TRAINING AND RETRIEVAL CONFIGURATION

We obtain out-of-fold proxy predictions using five-fold cross-fitting. In each fold, a one-hidden-layer MLP maps the frozen, ℓ_2-normalized 4096-dimensional Qwen3-Embedding-8B representation to the three harmonized proxies. The MLP has a hidden width of 512 with a GELU activation and is optimized using Adam with a learning rate of 10^{-3}, a batch size of 4096, and 2,800 optimizer updates. The final predictor is refitted on the full complete-case training set using the same architecture and optimization settings. Proxy harmonization uses a Rankit clipping parameter of $\delta = 10^{-4}$.

Dense retrievers rank the full corpus by inner product between ℓ_2-normalized query and document embeddings. The reranking baselines rescore the top-100 candidates retrieved by Qwen3-Embedding-8B; their recall is therefore bounded by this candidate pool.

For RADAR, PALIF is optimized using multi-start SLSQP initialized around the leading eigenvector of $\mathbf{\Sigma}_y$. The main experiments use a variance-retention tolerance of $\varepsilon = 0.03$. After factor selection, corpus-level latent-impact prediction, rank normalization, and radial scaling are performed offline

before indexing. The radial scale is fixed to $\gamma = 0.05$, selected on validation queries, and retrieval uses standard single-vector maximum-inner-product search.

E ADDITIONAL RESULTS AND ANALYSES

E.1 BACKBONE GENERALIZATION

To examine whether RADAR's retrieval gains depend on the Qwen3-Embedding-8B backbone used in the main experiments, we apply the same augmentation procedure to three additional embedding models: SPECTER2, E5-large-v2, and NV-Embed-v2. These backbones span both scientific-domain and general-purpose retrieval models, allowing us to test whether radial attribute augmentation transfers across different pretrained representation spaces.

Table 5: Backbone generalization of RADAR. We report nDCG@10 and Recall@50, both in %. The Avg. columns report the macro-average over the four benchmarks.

	SuReIR		GSRR		RMC		MasterSet		Avg.	
Baseline	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**	**nDCG @10**	**Recall @50**
SPECTER2	54.15	30.36	69.51	69.93	13.10	35.49	12.29	25.69	37.26	40.37
SPECTER2 + RADAR	**55.79**	**32.24**	**70.18**	**70.76**	**14.32**	**36.85**	**13.13**	**27.26**	**38.36**	**41.78**
E5-large-v2	51.67	30.19	82.82	80.50	12.37	32.53	12.73	25.11	39.90	42.08
E5-large-v2 + RADAR	**53.48**	**31.91**	**83.56**	**81.18**	**13.75**	**34.02**	**13.69**	**26.95**	**41.12**	**43.52**
NV-Embed-v2	56.92	33.59	81.60	77.97	12.30	39.44	15.15	31.39	41.49	45.60
NV-Embed-v2 + RADAR	**58.35**	**35.55**	**82.22**	**78.74**	**13.41**	**40.75**	**15.94**	**33.07**	**42.48**	**47.03**

As shown in Table 5, RADAR improves both nDCG@10 and Recall@50 for all three backbones on every evaluated benchmark. On average, RADAR increases nDCG@10 from 37.26 to 38.36 and Recall@50 from 40.37 to 41.78 for SPECTER2, from 39.90 to 41.12 and 42.08 to 43.52 for E5-large-v2, and from 41.49 to 42.48 and 45.60 to 47.03 for NV-Embed-v2. These results indicate that the gains from radial attribute augmentation are not specific to the Qwen3 embedding space and extend across both scientific-domain and general-purpose embedding backbones.

E.2 PROSPECTIVE IMPACT PREDICTION

To assess whether latent impact can be predicted prospectively from paper content, we evaluate prospective impact prediction on a chronologically held-out OpenAlex cohort of 26,651 papers published in 2022, using 2018 papers for validation. Predictions use frozen semantic representations without citation or venue observations from the test papers. Three-year citation outcomes count citing works published in 2023–2025, with a data cutoff of December 31, 2025. Spearman correlation uses citation counts z-score normalized within field–topic cohorts in the test set, where topics are the 500 semantic topics fitted on training data only; the C-index uses raw citation counts. Top-10% positives are defined by the 90th percentile of cohort-normalized outcomes over the full test set, using linear interpolation and including ties at the threshold.

Table 6: Prospective impact prediction on the chronologically held-out OpenAlex cohort. Best and second-best results are shown in **bold** and underlined, respectively.

Method	**Spearman ↑**	**C-index ↑**	**Top-10% AUROC ↑**	**AP ↑**
CitationPred	0.2654 ± 0.0156	0.6709 ± 0.0070	0.6168 ± 0.0117	0.1453 ± 0.0093
VenuePred	0.1694 ± 0.0020	0.6100 ± 0.0014	0.6124 ± 0.0018	0.1511 ± 0.0004
MultiProxyPred	0.2380 ± 0.0086	0.6566 ± 0.0043	0.6258 ± 0.0065	0.1565 ± 0.0050
FactorDistill	<u>0.3110</u>	**0.6930**	<u>0.6526</u>	<u>0.1837</u>
RADAR	$\mathbf{0.3409 \pm 0.0021}$	<u>0.6741 ± 0.0017</u>	$\mathbf{0.6999 \pm 0.0013}$	$\mathbf{0.2160 \pm 0.0004}$

As shown in Table 6, RADAR achieves the highest Spearman correlation, Top-10% AUROC, and average precision, while FactorDistill obtains the highest C-index. The results therefore differ across

the two forms of prospective evaluation: FactorDistill shows the highest pairwise concordance with raw future citation outcomes, whereas RADAR yields the highest association with normalized future-impact rankings and the highest performance in identifying the top-impact group.

E.3 LEAVE-ONE-PROXY-OUT VALIDATION

We further examine whether the latent attribute transfers to proxy families that are excluded during training. We conduct a leave-one-proxy-family-out evaluation in which either all citation-based proxies or the venue-based proxy is withheld when fitting the latent attribute model, and the resulting scores are evaluated against the excluded proxy family. This provides a stricter test of whether RADAR captures structure shared across heterogeneous proxy sources rather than only reproducing the signals used for fitting. For Held-out Citation, RADAR reduces to predicting the harmonized venue proxy, oriented by venue order rather than citations; VenuePred predicts the original venue rank.

Table 7: Leave-one-proxy-out construct validity. For held-out citation evaluation, all citation proxies are excluded from training and models are evaluated on unseen citation outcomes. For held-out venue evaluation, venue ranks are excluded from training and models are evaluated on unseen ordinal venue outcomes. Stochastic models are reported as mean ± sample standard deviation over five seeds; FactorDistill is reported as a single deterministic result.

Held-out Citation		
Baseline	**Spearman ↑**	**C-index ↑**
VenuePred	0.1186 ± 0.0027	0.5814 ± 0.0019
FactorDistill	$\underline{0.1517}$	$\underline{0.6179}$
RADAR	$\mathbf{0.1634 \pm 0.0022}$	$\mathbf{0.6372 \pm 0.0015}$
Held-out Venue		
Baseline	**Macro-AUROC ↑**	**QWK ↑**
CitationPred	0.6238 ± 0.0031	0.2147 ± 0.0046
FactorDistill	$\underline{0.6716}$	$\mathbf{0.2768}$
RADAR	$\mathbf{0.6849 \pm 0.0025}$	$\underline{0.2705} \pm 0.0042$

As shown in Table 7, RADAR transfers favorably to proxy families that are excluded during fitting. When citation-based proxies are held out, RADAR achieves the highest Spearman correlation (0.1634) and C-index (0.6372), improving over both VenuePred and FactorDistill. When venue information is held out, RADAR obtains the highest Macro-AUROC (0.6849) and a QWK of 0.2705, slightly below FactorDistill (0.2768) but above CitationPred (0.2147). Overall, RADAR achieves the best result on three of the four held-out metrics. These results provide additional evidence that the distilled attribute captures information that transfers across heterogeneous proxy families, rather than being tied exclusively to the proxy signals available during fitting.

E.4 EFFICIENCY AND LATENCY

We compare the online retrieval latency of RADAR with two reranking pipelines built on the same Qwen3-Embedding-8B first-stage retriever. RADAR performs latent-attribute inference and radial document scaling offline before indexing; its online execution therefore consists only of a single Qwen3 query encoding followed by one maximum-inner-product search. For each dataset, we measure 100 fixed queries after two warm-up queries and report the mean, sample standard deviation, and 95th-percentile (P95) latency. All measurements are in milliseconds per query on a single NVIDIA A100-PCIE-40GB GPU. For each reranker, end-to-end latency includes both the measured dense first-stage retrieval and Top-100 reranking, while model-loading time is excluded for all methods.

Table 8: Online retrieval latency on the four scientific literature benchmarks. Values report mean ± sample standard deviation, with P95 latency in parentheses, over 100 fixed measured queries per

dataset after two warm-up queries. All times are milliseconds per query on one NVIDIA A100-PCIE-40GB. Slowdown is measured relative to RADAR's single-stage retrieval.

Dataset	Corpus	RADAR Mean ± SD (P95)	BGE Reranker E2E Mean ± SD (P95)	Qwen3 Reranker E2E Mean ± SD (P95)	Slowdown BGE / Qwen3
SuReIR	29,421	62.61 ± 10.62 (82.33)	409.41 ± 26.33 (440.01)	4812.75 ± 203.22 (4928.32)	6.5× / 76.9×
GSRR	20,000	53.71 ± 1.14 (55.38)	427.78 ± 49.04 (501.72)	3418.16 ± 335.44 (3778.84)	8.0× / 63.6×
RMC	62,408	63.47 ± 1.32 (65.48)	1243.08 ± 161.07 (1475.53)	18408.35 ± 1560.32 (20245.11)	19.6× / 290.0×
MasterSet	67,761	57.79 ± 2.91 (63.12)	405.28 ± 33.94 (458.37)	1107.62 ± 29.95 (1149.63)	7.0× / 19.2×

As shown in Table 8, RADAR retains low online latency across all four benchmarks, with mean query times between 53.71 and 63.47 ms. In contrast, the BGE reranking pipeline requires between 405.28 and 1243.08 ms per query, corresponding to a 6.5×–19.6× slowdown relative to RADAR. The Qwen3 reranker is substantially more expensive, with mean end-to-end latency ranging from 1107.62 to 18408.35 ms and slowdowns of 19.2×–290.0×. These results highlight a key computational property of RADAR: latent-attribute prediction and radial scaling are performed offline, while online retrieval preserves the execution pattern of standard single-stage dense retrieval rather than introducing an additional reranking stage.

E.5 SENSITIVITY ANALYSIS

We examine sensitivity to the two RADAR-specific hyperparameters: the PALIF variance-retention tolerance ε and the radial scaling strength γ. Unless otherwise stated, all remaining settings are fixed to those used in the main experiments.

Variance-retention tolerance. Figure 2 examines the effect of ε, which controls the maximum proxy-variance loss allowed by the PALIF constraint. As ε increases, PALIF is allowed to sacrifice more proxy variance in exchange for improved predictability from the embedding. The frozen main setting, $\varepsilon = 0.03$, retains approximately 97% of the maximum proxy variance while reducing the out-of-fold prediction error relative to more restrictive settings. Across the five-seed validation analysis, performance generally improves up to this region, whereas the more permissive $\varepsilon = 0.05$ setting exhibits substantially greater variability. These results place the main setting near a favorable fidelity–predictability trade-off.

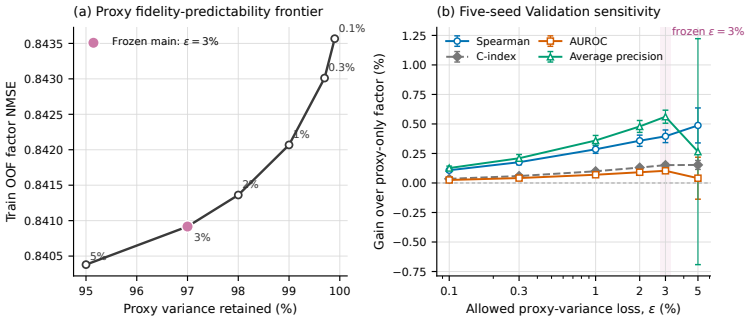


Figure 2: Sensitivity to the PALIF variance-retention tolerance ε. **(a)** Trade-off between retained proxy variance and train-only out-of-fold factor-prediction NMSE. **(b)** Five-seed validation gains over the proxy-only leading factor as ε varies. The reference uses the same proxy preprocessing and selects the leading eigenvector of $\mathbf{\Sigma}_y$. The highlighted setting corresponds to the frozen main value $\varepsilon = 0.03$.

Radial scaling strength. Figure 3 varies the radial scaling parameter γ. When $\gamma = 0$, RADAR reduces to the original cosine-based dense retriever; increasing γ gives progressively greater weight to the distilled attribute. The results show no single monotonic trend across datasets: larger values continue to benefit MasterSet, whereas aggressive scaling eventually degrades performance on the other benchmarks. The frozen main setting, $\gamma = 0.05$, lies before this stronger cross-dataset divergence and is used as a single global scaling value in the main experiments.

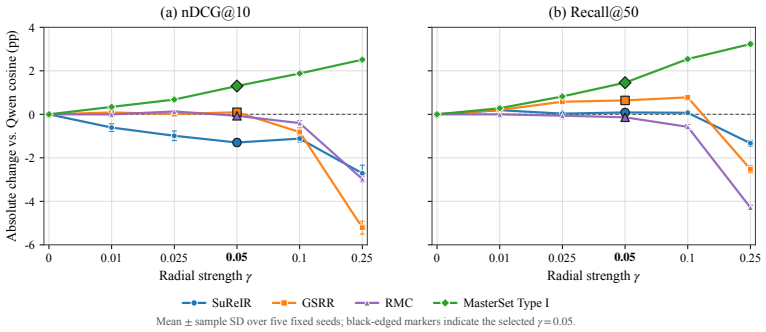


Figure 3: Sensitivity to the radial scaling strength γ on the validation splits. The panels report absolute changes in **(a)** nDCG@10 and **(b)** Recall@50 relative to the unscaled Qwen3-Embedding-8B cosine retriever. Points show means over five fixed seeds, with error bars indicating sample standard deviations. Black-edged markers denote the frozen main setting $\gamma = 0.05$.