Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16208
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Semantic Embedding and Profile-Based Ranking for Automated Reviewer Recommendation

Azisya Luthfi Bintang[1)*], Ida Nurhaida[2)]
[1,2)]Department of Informatics, Universitas Pembangunan Jaya, Tangerang Selatan, Indonesia,
[2)]Center of Urban Studies, Universitas Pembangunan Jaya, Tangerang Selatan, Indonesia
[1)] azisya.luthfibintang@student.upj.ac.id, [2)] ida.nurhaida@upj.ac.id

Submitted :May 13, 2026 | **Accepted** : May 17, 2026 | **Published** : July 5, 2026

Abstract: Manual reviewer assignment in peer review is difficult to scale because submission volumes grow faster than editors can inspect reviewer expertise, and reviewer profiles shift across topics and time. Existing automated approaches often rely on keyword or lexical matching, which cannot capture semantic similarity, and few combine dense retrieval with interpretable reviewer evidence. This study develops and evaluates an explainable reviewer recommendation system using a BERT-first Reciprocal Rank Fusion semantic-profile backend. The system retrieves candidate evidence using BERT and SPECTER2 semantic representations, extracts candidate reviewers from retrieved paper authors, and ranks them using fused retrieval evidence supported by frequency, h-index, and recency signals. The expertise-scoring component was evaluated using the Stelmakh/OpenReview benchmark, while end-to-end recommendation was evaluated on an OpenAlex citation-based proxy dataset using a validation split for configuration selection and a held-out test split for final reporting. SPECTER2 max pooling achieved a weighted Kendall tau loss of 0.22 on the Stelmakh/OpenReview benchmark, consistent with the public SPECTER2 baseline. On the held-out test split, the selected BERT-first RRF semantic-profile backend achieved the highest NDCG@10 of 0.2621, significantly outperforming BERT, SPECTER2-only, BM25, TF-IDF, and the previous profile-heavy backend. These findings indicate that rank-level fusion of complementary dense retrieval signals can improve reviewer candidate ranking while retaining interpretable profile evidence for editorial workflows. The local evaluation uses citation-based proxy relevance rather than true editorial assignments, so further validation using human-annotated reviewer data is needed.

Keywords: Automated Reviewer Recommendation, BERT, Citation-Based Proxy Relevance, Peer Review, Profile-Based Ranking, Reciprocal Rank Fusion, Semantic Similarity, SPECTER2

INTRODUCTION

Scientific peer review is a central mechanism for evaluating the quality, validity, and relevance of scholarly manuscripts. In this process, editors or program chairs must assign each submission to reviewers who have suitable expertise, sufficient familiarity with the topic, and no apparent conflict of interest. This task is increasingly difficult because scholarly publication volume continues to grow, reviewer expertise changes over time, and editorial systems often lack complete metadata about potential reviewers. Prior studies describe the reviewer assignment problem as a long-standing challenge that requires accurate reviewer-paper relevance estimation before final assignment decisions can be made (Aksoy et al., 2023; Zhao & Zhang, 2022).

Manual reviewer selection becomes less reliable at scale. Editors may depend on keyword search, personal networks, prior reviewer lists, or visible publication records. These strategies are useful but limited. Keyword-based matching can miss reviewers whose publications use different terminology while still addressing the same research topic. Lexical models also tend to represent documents through surface word overlap rather than deeper semantic similarity. Recent reviewer assignment and scientific retrieval studies therefore explore semantic modeling and dense document representations to improve matching between submissions and reviewer profiles (Singh et al., 2023; Tan et al., 2021).

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Automated reviewer recommendation is important because poor reviewer assignment can affect review quality, editorial workload, and publication fairness. A useful system should not only retrieve candidates, but also provide interpretable evidence for why a candidate is relevant. This interpretability is important because reviewer recommendation systems are intended to support editorial judgment rather than replace it. Existing work has examined reviewer assignment methods, semantic topic modeling, and benchmark construction for reviewer-paper matching (Aksoy et al., 2023; Stelmakh et al., 2025). However, several gaps remain. Many systems still rely on lexical or keyword-based similarity. Several studies evaluate only a component of the reviewer assignment process, while fewer combine external human-rated expertise validation with local end-to-end recommendation evaluation. Moreover, dense retrieval models may identify relevant candidates but often provide limited editorial evidence about the factors behind the ranking. In addition, author-level signals such as publication impact, publication recency, and authorship evidence are rarely integrated with dense embedding retrieval in a single explainable ranking pipeline.

This study proposes an automated reviewer recommendation system based on BERT-first semantic retrieval, Reciprocal Rank Fusion, and lightweight profile-based ranking evidence. The system represents a manuscript using dense semantic embeddings, retrieves supporting papers using BERT and SPECTER2 representations, extracts candidate reviewers from retrieved paper authors, and ranks candidates using fused retrieval evidence supported by frequency, h-index, and publication recency. Authorship position was evaluated in the earlier profile-heavy configuration but was not retained in the revised ranking formula because it reduced local citation-proxy ranking performance. The resulting score is used as ranking evidence for editorial decision support, not as a deterministic measure of reviewer suitability.

The contributions of this study are sixfold. First, it develops an end-to-end reviewer recommendation system based on dense semantic retrieval and profile-based ranking evidence. Second, it validates SPECTER2-style semantic expertise scoring using the Stelmakh/OpenReview human-rated benchmark. Third, it introduces a BERT-first Reciprocal Rank Fusion backend that combines BERT and SPECTER2 retrieval evidence for local reviewer candidate ranking. Fourth, it integrates lightweight author-level evidence, including repeated retrieval frequency, h-index, and publication recency, to support interpretability. Fifth, it evaluates the full recommendation pipeline using an OpenAlex citation-based proxy dataset with validation-based configuration selection and held-out test reporting. Sixth, it applies paired Wilcoxon signed-rank testing on manuscript-level NDCG@10 scores. These contributions position the system as an explainable reviewer discovery tool that provides ranking evidence for editorial decision support. The citation-based proxy is treated as a local relevance proxy, not as true editorial ground truth.

LITERATURE REVIEW

This section discusses previous studies and theoretical foundations related to reviewer assignment systems, semantic embedding techniques, and evaluation frameworks used in scientific recommendation systems.

Reviewer Assignment and Recommendation Systems

Reviewer assignment is commonly understood as a constrained matching and ranking problem that connects manuscripts with appropriate reviewers by considering expertise, workload, conflicts of interest, and fairness requirements (Aksoy et al., 2023; Zhao & Zhang, 2022). Ribeiro et al. (2023) explains that this research area generally involves two related stages: identifying suitable expert reviewers and assigning them to submitted manuscripts. Earlier approaches relied on reviewer profiles, keywords, subject categories, topic models, and lexical similarity between manuscript text and reviewer publications. To improve scalability, systems such as PaRe projected papers and reviewers into a shared topic space, while later studies incorporated rule-based matching, knowledge graphs, latent Dirichlet allocation, classification, and link prediction to capture topical and relational evidence (Anjum et al., 2019; Nugroho et al., 2023; Yong et al., 2021). Although these methods reduced dependence on manual assignment, they were still limited when manuscript metadata were sparse or when reviewer expertise was expressed through semantically related but lexically different terms. Recent studies have therefore moved toward learning-based and hybrid approaches. Tan et al. (2021) showed that combining word-level and semantic features can improve reviewer assignment beyond surface keyword overlap. Zhao & Zhang (2022) further showed that peer review automation includes text mining, machine learning, optimization, and constraint-aware assignment methods. However, recent surveys indicate that reviewer assignment research remains fragmented across matching models, profile evidence, and evaluation design (Aksoy et al., 2023; Zhao & Zhang, 2022). This gap suggests the need for an explainable pipeline that integrates dense semantic retrieval, author-level profile ranking, external human-rated validation, and local end-to-end evaluation within a single reviewer recommendation framework.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16208
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Semantic Embedding for Scientific Text Matching

Lexical and semantic retrieval offer complementary approaches to scholarly text matching because lexical methods preserve exact term overlap, while dense representations capture broader semantic relationships between queries and documents (Lin & Lin, 2023). This distinction is important in reviewer recommendation, where a reviewer may be relevant even when their publications use terminology that differs from the submitted manuscript. Neural language models help address this problem by learning contextual text representations. BERT introduced deep bidirectional transformer representations for general language understanding, while SciBERT adapted transformer pretraining to scientific corpora and improved performance on scientific text tasks (Beltagy et al., 2019; Devlin et al., 2019). SPECTER further advanced scientific document representation by using citation-informed training, making it more suitable for paper-paper similarity than generic encoders (Cohan et al., 2020). More recent work has refined scientific embeddings through task adaptation and more specific representation learning. SPECTER2 extended this direction through the SciRepEval benchmark and adapter-based training across multiple scientific tasks, supporting paper similarity and expertise scoring (Singh et al., 2023). Ostendorff, Rethmeier, et al. (2022) strengthened scientific document embeddings using neighborhood contrastive learning with citation embeddings, while Ostendorff, Blume, et al. (2022) showed that aspect-specific embeddings can make paper similarity more explainable by separating task, method, and dataset dimensions. These studies support the use of dense scientific embeddings for identifying semantically related scholarly texts, but embedding similarity alone is insufficient because it does not capture author-level evidence such as research impact, recent activity, or authorship role. This limitation creates a need for reviewer recommendation pipelines that combine semantic retrieval with interpretable profile-based ranking signals. The need for combination also reflects a broader retrieval challenge: no single retrieval model captures all forms of relevance equally well. Hybrid retrieval addresses this issue by combining lexical and dense signals, and Bruch et al. (2023) show that fusion methods can improve retrieval quality while noting that Reciprocal Rank Fusion is not always parameter-free or universally superior; score-level convex combination may be stronger when normalization and tuning data are available, whereas RRF remains useful when rank evidence from different models must be combined without directly calibrating raw similarity scores.

Evaluation Frameworks and Profile-Based Signals

Evaluation remains a major challenge in reviewer recommendation because actual editorial assignments are difficult to access and may reflect institutional constraints, reviewer availability, or conflict-of-interest considerations that are not visible in public datasets (Aksoy et al., 2023). Human-rated benchmarks can provide direct evidence for reviewer-paper expertise scoring, but they do not necessarily evaluate the full discovery process from manuscript input to a ranked reviewer list. For example, the Stelmakh/OpenReview benchmark is useful for assessing semantic expertise scoring through human ratings, but it does not fully validate reviewer discovery or editorial suitability (Stelmakh et al., 2025). Citation-based evaluation can therefore serve as a scalable silver-standard alternative by treating citing authors as topically related candidates, although citation links should not be interpreted as actual editorial reviewer assignments. This aligns with Otero et al. (2023), who emphasized that information retrieval evaluation depends on relevance judgments and requires careful attention to assessment cost, reliability, and reuse, supporting the need for transparent proxy relevance design. Beyond evaluation, profile-based signals can complement semantic similarity by adding author-level evidence to reviewer ranking. Bibliometric indicators such as publication count, citation count, and h-index are often used to represent scholarly impact, but Momeni et al. (2023) showed that these indicators should be interpreted as profile evidence rather than direct measures of reviewer suitability. Recency is also important because recent publications may better reflect current expertise, and Jiang et al. (2023) incorporated time-aware publication behavior in scientific recommendation. Authorship position may provide additional evidence about contribution patterns, but Baum et al. (2023) showed that byline order is influenced by disciplinary and social conventions, so it should be used cautiously. Overall, these studies suggest that reviewer recommendation can be strengthened by integrating profile signals such as h-index, recency, and authorship position with dense scientific embedding retrieval and dual evaluation.

Table 1. Comparison of Prior Reviewer Recommendation Studies

Study	Retrieval/Matching Approach	Profile Signals Used	Evaluation Method	Key Contribution
Anjum et al. (2019)	Common topic space for paper-reviewer matching	Reviewer publication profiles	Paper-reviewer matching evaluation	Reduced vocabulary mismatch through shared topic representation

Zhao & Zhang (2022)	Survey of rule-based, topic-based, and learning-based reviewer assignment algorithms	Not profile-based	Systematic review of prior methods	Comprehensive survey covering classical and modern reviewer assignment approaches
Nugroho et al. (2023)	LDA, classification, and link prediction	Topic and conflict-related relational evidence	MAP comparison	Reviewer assignment method considering conflict-of-interest constraints
Cohan et al. (2020)	Citation-informed scientific document embeddings	Not profile-based	Scientific document representation benchmarks	SPECTER document-level embedding for scientific papers
Singh et al. (2023)	Multi-format scientific document representation	Not profile-based	SciRepEval benchmark	SPECTER2 model family for multiple scientific representation formats
Stelmakh et al. (2025)	Similarity algorithms for reviewer-paper expertise scoring	Reviewer-paper expertise ratings	Human-rated self-reported benchmark	Gold-standard dataset for expertise-score evaluation
This Study	BERT-first Reciprocal Rank Fusion combining BERT and SPECTER2 retrieval evidence	Retrieval frequency, h-index, and recency as lightweight profile evidence	Stelmakh/OpenReview expertise benchmark and OpenAlex citation-proxy validation/test evaluation	Explainable end-to-end reviewer recommendation with rank-level semantic fusion, profile evidence, and held-out local evaluation

Overall, prior work addresses reviewer assignment, scientific text representation, evaluation design, and author profile modeling as related but mostly separate problems. Classical reviewer assignment systems provide scalable matching, but they often depend on lexical or topic-level representations. Dense scientific embeddings improve semantic matching, but they usually do not incorporate interpretable reviewer profile signals. Evaluation studies provide either human-rated expertise judgments or scalable proxy labels, but they rarely combine both in one pipeline. This study fills the combined gap by integrating dense semantic retrieval, BERT-first Reciprocal Rank Fusion, lightweight profile-based evidence, external human-rated expertise validation, and local citation-proxy end-to-end evaluation in a single explainable reviewer recommendation system.

METHOD

The proposed method consists of a deterministic reviewer recommendation pipeline that combines dense semantic retrieval, rank-level fusion, author-level candidate extraction, and lightweight profile-based ranking evidence. The system receives a manuscript title and abstract as input and returns a ranked list of reviewer candidates. The revised local recommendation backend uses BERT-first Reciprocal Rank Fusion to combine BERT and SPECTER2 retrieval evidence, then incorporates frequency, h-index, and recency as supporting profile signals. The methodology comprises dataset preparation, semantic paper retrieval, candidate extraction, RRF-based semantic-profile ranking, conflict filtering, external expertise-scoring evaluation, local validation and held-out test evaluation, statistical testing, and implementation configuration. Figure 1 illustrates the overall workflow of the system.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data Preparation

Two datasets were prepared for this study: a backend scholarly corpus and a local evaluation dataset, both sourced from OpenAlex (Priem et al., 2022). OpenAlex was selected because it provides open bibliographic metadata, author identifiers, citation relations, and concept-based filtering suitable for reproducible scholarly evaluation. The backend corpus covers Computer Science papers published between 2020 and 2025, with paper metadata and authorship records stored in PostgreSQL and SPECTER2 embeddings indexed in Qdrant using cosine distance. The evaluation dataset was constructed by selecting 500 manuscripts published between 2020 and 2024 with citation counts between 10 and 500, using random seed 42. For each manuscript, authors of citing works were retrieved from OpenAlex and treated as proxy relevant reviewers after excluding the original manuscript authors and filtering to authors available in the local database. This dataset represents citation-based proxy relevance, not true editorial reviewer assignments. For the revised local evaluation, the 500 manuscripts were divided using a fixed 300/100/100 split. The validation split was used to select among predefined BERT-first and RRF scoring configurations, while the held-out test split was used once for final performance reporting. Table 2 summarizes both datasets.

Table 2. Dataset summary

Item	**Backend Corpus**	**Evaluation Dataset**
Source	OpenAlex	OpenAlex
Domain / Strategy	Computer Science	Citation-based proxy
Year range	2020–2025	2020–2024
Papers / Manuscripts	50,190	500
Authors	220,832	N/A
Authorship records	310,738	N/A
Minimum citations	N/A	10
Maximum citations	N/A	500
Min. proxy reviewers per manuscript	N/A	5
Average proxy reviewers	N/A	118.11
Unique proxy reviewers	N/A	40,370
Random seed	N/A	42
Validation manuscripts	N/A	100
Held-out test manuscripts	N/A	100

System Architecture

Figure 1 summarizes the revised reviewer recommendation workflow. The system receives a manuscript title and abstract and encodes the text using two dense semantic representations. BERT is used as the primary local retrieval signal, while SPECTER2 provides complementary scientific-document retrieval evidence. Retrieved papers from both semantic rankings are linked to authors through PostgreSQL authorship records. Candidate reviewers are then represented using rank-based semantic evidence and lightweight profile evidence. The final local ranking is computed using BERT-first Reciprocal Rank Fusion supported by retrieval frequency, h-index, and recency. Affiliation-based conflict filtering is applied when metadata are available, and the final output is a ranked list of reviewer candidates.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16208
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Fig. 1 Reviewer recommendation pipeline using BERT-first Reciprocal Rank Fusion.

Semantic Paper Retrieval and BERT-First RRF Reviewer Ranking

The manuscript title and abstract were concatenated into a single query text. For SPECTER2, the input follows the title–abstract format used in scientific document representation. The same manuscript content was encoded using two dense encoders: BERT as the primary local retrieval signal and SPECTER2 as a complementary scientific-document representation. BERT was prioritized because it produced the strongest dense baseline in the earlier citation-proxy evaluation, while SPECTER2 was retained because it provides domain-specific scientific retrieval evidence and supports the external expertise-scoring evaluation.

$$x_m = title_m\ [SEP]\ abstract_m \qquad (1)$$

where x_m denotes the textual representation of manuscript m, The manuscript representation was then encoded into BERT and SPECTER2 query embeddings as follows:

$$v_m^B = f_B(x_m), \qquad v_m^S = f_S(x_m) \qquad (2)$$

where v_m^B and v_m^S are the BERT and SPECTER2 query embeddings, and f_B and f_S denote the corresponding encoders. Each query embedding is compared with corpus-paper embeddings from the same representation space using cosine similarity:

$$sim_j(m,p) = \frac{v_m^j \cdot v_p^j}{\left\|v_m^j\right\| \left\|v_p^j\right\|}, \qquad j \in \{B,S\} \qquad (3)$$

$sim_j(m,p)$ denotes the similarity between manuscript m and corpus paper p, j denotes the retrieval model, B denotes BERT, S denotes SPECTER2. v_m^j is the manuscript embedding, and v_p^j is the embedding of corpus paper p. These similarity scores produce two ranked lists of supporting papers for the same manuscript: a BERT-ranked list and a SPECTER2-ranked list.

Reviewer candidates were extracted from the authors of retrieved papers by querying PostgreSQL authorship records and grouping evidence by OpenAlex author ID. For each candidate, the system records supporting papers from both retrieval rankings, retrieval frequency, publication years, and author profile metadata. Authorship position was excluded from the revised final ranking formula because earlier ablation showed that it reduced NDCG@10 on the citation-proxy evaluation.

After candidate extraction, the system computes rank-level semantic evidence using Reciprocal Rank Fusion. RRF is used because BERT and SPECTER2 produce separate ranked lists whose raw similarity scores need not be directly calibrated. For candidate reviewer a and retrieval model j, the candidate-level RRF evidence is computed as follows:

$$RRF_j(a) = \sum_{p \in P_a^j} \frac{1}{k + rank_j(p)}, \qquad j \in \{B,S\} \qquad (4)$$

where P_a^j denotes the supporting papers associated with candidate a in retrieval model j, $rank_j(p)$ is the rank of supporting paper p under that model, and k is the RRF constant, this study uses k = 60. The resulting BERT and SPECTER2 RRF values are normalized within the candidate pool before being combined with lightweight profile evidence. Retrieval frequency is normalized by the maximum candidate frequency, h-index is normalized by the maximum candidate h-index, and recency is computed as the average five-year linear recency score over the candidate's supporting papers. If all publication years are missing, the implementation uses a neutral recency value of 0.5.

The final candidate score combines normalized BERT RRF evidence, normalized SPECTER2 RRF evidence, normalized retrieval frequency, normalized h-index, and recency:

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16208
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$Score_a = 0.80\hat{R}_B(a) + 0.15\hat{R}_s(a) + 0.03\widehat{freq}_a + 0.01\hat{h}_a + 0.01R_a \quad (5)$$

where $Score_a$ denotes the final ranking score for candidate reviewer a, $\hat{R}_B(a)$ denotes normalized BERT RRF evidence, $\hat{R}_s(a)$ denotes normalized SPECTER2 RRF evidence, $\widehat{freq}_a$ denotes normalized retrieval frequency, $\hat{h}_a$ denotes normalized h-index, and R_a denotes the publication recency bonus. The weights were selected from predefined configurations using the validation split. The selected BERT-first RRF Semantic-Profile Backend was then evaluated once on the held-out test split. This design keeps semantic rank evidence dominant while preserving interpretable author-level evidence for editorial decision support. Table 3 presents the composite score components.

Table 3. Composite reviewer ranking score components.

Component	Symbol	Weight	Implementation Detail
BERT RRF evidence	$\hat{R}_B(a)$	0.80	Normalized reciprocal-rank evidence from BERT-ranked supporting papers
SPECTER2 RRF evidence	$\hat{R}_s(a)$	0.15	Normalized reciprocal-rank evidence from SPECTER2-ranked supporting papers
Retrieval frequency	$\widehat{freq}_a$	0.03	Normalized count of supporting retrieved papers
Normalized h-index	$\hat{h}_a$	0.01	Candidate h-index divided by the maximum h-index in the candidate pool
Recency bonus	R_a	0.01	Average five-year linear recency score of supporting papers

Conflict Filtering

The backend includes an affiliation-based conflict filtering component. Candidate affiliations and manuscript author affiliations are normalized by lowercasing, removing common institutional suffixes, and comparing normalized institution names. Candidates with overlapping institutions are excluded from the final ranking when affiliation metadata are available. This component was not evaluated in the validation or held-out test evaluation because manuscript author affiliation metadata were not reliably available in the citation-proxy dataset.

External Expertise-Scoring Evaluation

The semantic expertise-scoring component was externally evaluated using the Stelmakh/OpenReview reviewer-paper matching benchmark (Stelmakh et al., 2025). This benchmark provides human self-reported expertise judgments for reviewer-paper pairs. The evaluation tests the expertise-scoring component only. It does not test the complete reviewer discovery pipeline because reviewer candidates are already defined by the benchmark.

The main external scoring variant uses max pooling over the reviewer's profile papers. For a target manuscript and a reviewer profile, the system assigns the highest semantic similarity score among the reviewer's profile papers:

$$score_{max}(a, m) = max_{p \epsilon Profile(a)} sim(m, p) \quad (6)$$

where $score_{max}(a, m)$ denotes the max-pooling expertise score for reviewer a and manuscript m, $Profile(a)$ denotes the set of profile papers associated with reviewer a, p denotes a profile paper, and $sim(m, p)$ denotes cosine similarity between manuscript m and profile paper p. This score reflects the strongest available semantic evidence that a reviewer has written work related to the manuscript. A top-3 mean variant was also tested as an alternative pooling strategy; if fewer than three profile papers were available, the implementation averaged over the available papers. The official benchmark evaluator reports weighted Kendall tau loss, where lower values indicate better agreement with human-rated expertise orderings. Table 4 summarizes the external evaluation setup.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16208
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4. External expertise-scoring evaluation setup.

External Evaluation Item	**Value**
Benchmark	Stelmakh/OpenReview reviewer-paper expertise benchmark
Splits	d_20_1 to d_20_10
Reviewers per split	58
Submissions per split	463
Evaluated pairs per split	477
Metric	Weighted Kendall tau loss
Main external variant	SPECTER2 max pooling

Local End-to-End Evaluation

The complete deterministic backend was evaluated on the local OpenAlex citation-based proxy dataset. For each manuscript, each method produced a ranked list of 20 reviewer candidates. The 500 manuscripts were divided into a fixed 300/100/100 split. The validation split was used to select among predefined BERT-first weighted and RRF configurations, and the held-out test split was used once for final comparison against TF-IDF, BM25, BERT, SPECTER2-only, and the previous profile-heavy backend. The recommended reviewers were compared against the proxy relevant reviewer set derived from citing authors. The proxy relevant reviewer set for manuscript m is defined as:

$$G_m = (CitingAuthors(m) - Authors(m)) \cap A_{DB} \quad (7)$$

where G_m denotes the proxy relevant reviewer set for manuscript m, $CitingAuthors(m)$ denotes authors of works that cite manuscript m, $Authors(m)$ denotes the original authors of manuscript m, and A_{DB} denotes authors available in the local database. This formulation makes the evaluation transparent because proxy relevance is derived from citation relations, not from actual editorial reviewer assignment. Table 5 summarizes the six methods compared in the local end-to-end evaluation, spanning lexical baselines, dense embedding baselines, and profile-ranking ablation variants.

Table 5. Methods compared in the local end-to-end evaluation

Method	**Description**
TF-IDF	Lexical cosine similarity baseline over title and abstract
BM25	Lexical probabilistic retrieval baseline
BERT	General-purpose dense embedding baseline using all-MiniLM-L6-v2
SPECTER2-only	Scientific embedding retrieval without profile ranking
Previous Profile-Heavy backend	Earlier SPECTER2-based profile-ranking backend using frequency, similarity, h-index, recency, and author-position weighting
BERT-first RRF Semantic-Profile Backend	Complete backend selected on validation split using BERT-first RRF, SPECTER2 evidence, frequency, h-index, and recency

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16208
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Figure 2 summarizes the two evaluation tracks used in this study. The external track evaluates semantic expertise scoring using the Stelmakh/OpenReview benchmark. The local track evaluates end-to-end recommendation using citation-based proxy relevance from OpenAlex.

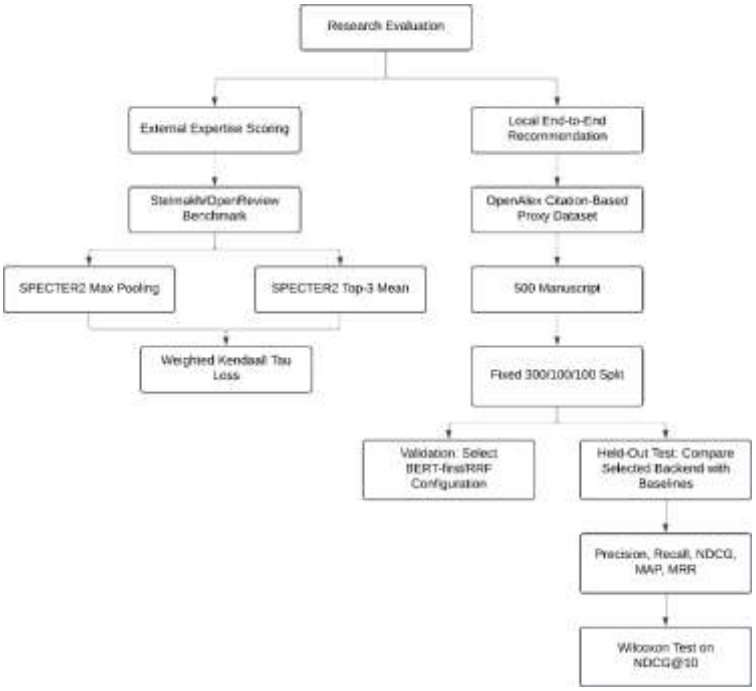

Fig. 2 Evaluation design for external expertise scoring and local end-to-end reviewer recommendation.

The local evaluation used ranking metrics commonly applied in information retrieval and recommender systems. Precision@K measures the proportion of relevant reviewers among the top-*K* recommendations, while Recall@K measures coverage of the relevant set. NDCG@10 applies logarithmic discounting to penalize relevant items appearing at lower ranks, making it the primary metric for this ranking task. MAP summarizes average precision across rank positions, and MRR captures how early the first relevant reviewer appears in the list.

Statistical Testing

Statistical significance was tested using the paired Wilcoxon signed-rank test on per-manuscript NDCG@10 scores. The selected BERT-first RRF Semantic-Profile Backend was treated as the primary method and compared against each baseline on the held-out test split. The test used a one-sided alternative hypothesis that the primary method produced higher NDCG@10 than the compared baseline. A result was considered statistically significant at $p < 0.05$.

Implementation Environment

The backend was implemented using FastAPI and Python. PostgreSQL stores paper, author, and authorship metadata, while Qdrant stores SPECTER2 paper embeddings for vector search. SQLAlchemy asynchronous sessions are used by the application backend, while the standalone ground-truth construction script uses a synchronous PostgreSQL connection for reliability. The evaluation scripts use NumPy, SciPy, scikit-learn, sentence-transformers, rank-bm25, and custom deterministic ranking scripts for baseline implementation, RRF-based scoring, validation selection, and statistical testing. LLM-assisted reranking exists as an optional application feature, but it was excluded from the deterministic evaluation because it was not reproducible under the available API constraints.

RESULT

The evaluation was conducted in two stages. The first stage assessed reviewer-paper expertise scoring on the external Stelmakh/OpenReview benchmark. The second stage evaluated end-to-end reviewer recommendation on the local OpenAlex citation-based proxy dataset using validation-based configuration selection and held-out test reporting.

External Expertise-Scoring Result

The Stelmakh/OpenReview benchmark measures how well a scoring method orders reviewer-paper expertise judgments based on human self-reported expertise labels (Stelmakh et al., 2025). This result is reported separately from the local evaluation because it evaluates the semantic expertise-scoring component, not the complete reviewer discovery pipeline.

Table 6. External expertise-scoring on the Stelmakh/OpenReview benchmark

Method	Weighted Kendall Tau Loss	95% CI	Note
SPECTER2 Max (ours)	0.22	[0.18; 0.26]	Best evaluated variant in this study
SPECTER2 Top-3 Mean (ours)	0.23	[0.19; 0.29]	Slightly worse than max pooling
Public SPECTER2 Max baseline	0.22	[0.18; 0.26]	Closest public baseline

The SPECTER2 max-pooling variant obtained a weighted Kendall tau loss of 0.22 with a 95% confidence interval of [0.18; 0.26]. Its result was practically identical to the public SPECTER2 max baseline, which also reported a loss of 0.22 with the same confidence interval. In contrast, the top-3 mean variant obtained a loss of 0.23 with a 95% confidence interval of [0.19; 0.29]. The gap between max pooling and top-3 mean was 0.01, indicating only a small difference between the two pooling strategies in this benchmark. This result should be interpreted as evidence for the externally evaluated expertise-scoring component only.

Local End-to-End Recommendation Result

The local evaluation used the OpenAlex citation-based proxy dataset with a fixed validation and held-out test split. The validation split was used to select among predefined BERT-first weighted and RRF configurations. The selected configuration was then evaluated once on the held-out test split against lexical baselines, dense embedding baselines, and the previous profile-heavy backend. This design was used to avoid selecting the final configuration directly from the held-out test results.

Table 7. Validation split ranking for BERT-first and RRF configuration selection.

Rank	Configuration	NDCG@10
1	BERT-first RRF 80-15-03-01-01	0.2355
2	BERT-first RRF 90-05-03-01-01	0.2329
3	BERT-first RRF Union 85-10-03-01-01	0.2322
4	BERT-first RRF Union 85-10-03-01-01	0.2321
5	BERT-first Weighted 85-08-04-02-01	0.2206
6	BERT-first Weighted 85-10-03-01-01	0.2204
7	BERT-first Weighted 80-10-05-03-02	0.2204
8	BERT-first Weighted Union 85-10-03-01-01	0.2204
9	BERT-first Weighted 90-05-03-01-01	0.2201
10	BERT-first Minimal Profile	0.2199
11	BERT-first Weighted 75-15-05-03-02	0.2197
12	BERT-first SPECTER2 Tie-break	0.2180

13	BERT-first Semantic Only	0.2149

The validation results show that rank-level fusion variants outperformed direct weighted-score variants. The best validation configuration was BERT-first RRF 80-15-03-01-01, which achieved NDCG@10 of 0.2355. This configuration assigns weights of 0.80 to BERT RRF evidence, 0.15 to SPECTER2 RRF evidence, 0.03 to retrieval frequency, 0.01 to h-index, and 0.01 to recency. It was selected as the final revised backend for held-out test evaluation.

Table 8. Held-out test results on the OpenAlex citation-based proxy dataset.

Method	**P@5**	**P@10**	**NDCG@10**	**MAP**	**MRR**
TF-IDF	0.0960	0.1270	0.1161	0.0120	0.1944
BM25	0.1280	0.1520	0.1453	0.0127	0.2443
BERT	0.2520	0.2150	0.2290	0.0163	0.4226
SPECTER2-only	0.2000	0.1920	0.1980	0.0134	0.3453
Previous Profile-Heavy Backend	0.1680	0.1580	0.1688	0.0110	0.3483
BERT-first RRF Semantic-Profile Backend	0.2700	0.2540	0.2621	0.0192	0.4352

On the held-out test split, the selected BERT-first RRF Semantic-Profile Backend achieved the highest performance across the primary metric and most supporting metrics. It obtained NDCG@10 of 0.2621, compared with 0.2290 for BERT, 0.1980 for SPECTER2-only, 0.1688 for the previous profile-heavy backend, 0.1453 for BM25, and 0.1161 for TF-IDF. The revised BERT-first RRF Semantic-Profile Backend also achieved the highest P@5, P@10, MAP, and MRR among the evaluated methods. These results indicate that rank-level fusion of BERT and SPECTER2 evidence improved reviewer candidate ranking compared with both single dense retrieval and the earlier profile-heavy scoring configuration.

Statistical Test Result

The Wilcoxon signed-rank test compared the BERT-first RRF Backend against each baseline using paired manuscript-level NDCG@10 scores. The one-sided alternative tested whether the BERT-first RRF Backend produced higher NDCG@10 than each compared method.

Table 9. Wilcoxon signed-rank test result comparing BERT-first RRF Backend against each method on per-manuscript NDCG@10 scores.

Comparison	**Mean Difference**	**p-value**	**Significant**
BERT-first RRF Backend vs TF-IDF	+0.145920	<0.001	Yes
BERT-first RRF Backend vs BM25	+0.116717	<0.001	Yes
BERT-first RRF Backend vs BERT	+0.033059	<0.001	Yes
BERT-first RRF Backend vs SPECTER2-only	+0.064073	<0.001	Yes
BERT-first RRF Backend vs Previous Profile-Heavy Backend	+0.093298	<0.001	Yes

The Wilcoxon signed-rank test confirmed that the selected BERT-first RRF Semantic-Profile Backend significantly outperformed all compared baselines on held-out test NDCG@10. The improvement over BERT was +0.0331 with p = 0.000122, showing that the revised rank-fusion strategy provided a statistically significant gain over the strongest dense baseline. The backend also significantly outperformed SPECTER2-only, the previous

profile-heavy backend, BM25, and TF-IDF. These results directly address the earlier limitation that the previous profile-heavy backend did not outperform dense embedding baselines.

DISCUSSIONS

The external benchmark result confirms that the SPECTER2 max-pooling implementation reproduced the expected behavior of a domain-specific scientific document representation model for reviewer-paper expertise scoring. The implementation achieved a weighted Kendall tau loss of 0.22 on the Stelmakh/OpenReview benchmark, consistent with the public SPECTER2 max baseline, which validates the expertise-scoring component rather than claiming a new advance (Singh et al., 2023; Stelmakh et al., 2025). The local evaluation showed that BERT was a strong dense baseline, while SPECTER2-only remained useful as a scientific-document retrieval signal. The revised BERT-first RRF backend combines these complementary signals at the rank level and achieved the highest held-out test NDCG@10. This indicates that BERT and SPECTER2 capture different retrieval evidence, and that rank-level fusion can be more effective than relying on either dense retriever alone.

The BERT-first RRF backend results also clarify the role of profile-based evidence. In the earlier profile-heavy configuration, author-level signals changed ranking behavior but did not consistently improve NDCG@10 over dense retrieval alone. The BERT-first RRF backend therefore uses profile signals only as lightweight supporting evidence. Frequency, h-index, and recency receive small weights, while semantic rank evidence remains dominant. This design improved the held-out test NDCG@10 from 0.2290 for BERT to 0.2621 for the selected backend and significantly outperformed BERT under the paired Wilcoxon test. Authorship position was excluded because earlier ablation showed that it reduced NDCG@10, which is consistent with concerns that byline order is an imperfect proxy for reviewer expertise or contribution patterns (Baum et al., 2023).

Several limitations constrain the interpretation of these findings. The citation-based proxy is not equivalent to true editorial reviewer assignment: citing authors may be topically related to a manuscript but are not necessarily qualified, available, independent, or free of conflict. The local corpus is restricted to OpenAlex Computer Science papers from 2020 to 2025, so generalization to other disciplines or time periods is not established. The affiliation-based conflict filtering component was not evaluated due to missing manuscript-author affiliation metadata in the proxy dataset. In addition, the RRF configuration was selected from predefined candidates on a validation split rather than learned from human editorial labels. Despite these constraints, the BERT-first RRF backend demonstrates practical value as an explainable decision-support tool: rank-level dense retrieval fusion improves candidate ranking, while frequency, h-index, and recency provide interpretable evidence that editors can inspect during reviewer selection. Therefore, the system is better positioned as a human-in-the-loop decision-support tool than as an automatic replacement for editorial judgment. Future work should evaluate the pipeline using human-annotated reviewer assignment data, test the RRF configuration across additional disciplines, and further validate conflict-of-interest modeling in real editorial settings.

CONCLUSION

This research developed an automated reviewer recommendation pipeline that integrates dense semantic retrieval, Reciprocal Rank Fusion, and lightweight profile-based ranking evidence. The external evaluation showed that SPECTER2 max pooling achieved a weighted Kendall tau loss of 0.22 on the Stelmakh/OpenReview benchmark, confirming the correctness of the expertise-scoring component. In the local end-to-end evaluation, the selected BERT-first RRF Semantic-Profile Backend achieved the highest held-out test NDCG@10 of 0.2621 and significantly outperformed BERT, SPECTER2-only, BM25, TF-IDF, and the previous profile-heavy backend. These findings answer the research objective by showing that rank-level fusion of BERT and SPECTER2, supported by frequency, h-index, and recency signals, can improve reviewer candidate ranking while preserving interpretable evidence for editorial decision support. However, the evaluation still relies on citation-based proxy relevance rather than true editorial reviewer assignments, so future studies should validate the pipeline using human-annotated reviewer data, broader disciplinary datasets, and more reliable conflict-of-interest metadata.

REFERENCES

Aksoy, M., Yanik, S., & Amasyali, M. F. (2023). Reviewer Assignment Problem: A Systematic Review of the Literature. In *Journal of Artificial Intelligence Research* (Vol. 76). https://doi.org/10.1613/jair.1.14318

Anjum, O., Gong, H., Bhat, S., Xiong, J., & Hwu, W.-M. (2019). PaRe: A Paper-Reviewer Matching Approach Using a Common Topic Space. *Association for Computational Linguistics*, 518–528. https://doi.org/10.18653/v1/D19-1049

Baum, M. A., Braun, M. N., Hart, A., Huffer, V. I., Meßmer, J. A., Weigl, M., & Wennerhold, L. (2023). The first author takes it all? Solutions for crediting authors more visibly, transparently, and free of bias. *British Journal of Social Psychology*, *62*(4), 1605–1620. https://doi.org/10.1111/bjso.12569

Beltagy, I., Lo, K., & Cohan, A. (2019). SCIBERT: A Pretrained Language Model for Scientific Text. *Association for Computational Linguistics*, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Bruch, S., Gai, S., & Ingber, A. (2023). An Analysis of Fusion Functions for Hybrid Retrieval. *ACM Transactions on Information Systems*, *42*(1). https://doi.org/10.1145/3596512

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. *Association for Computational Linguistics*, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

Devlin, J., Chang, M.-W., Lee, K., & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Association for Computational Linguistics*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Jiang, C., Ma, X., Zeng, J., Zhang, Y., Yang, T., & Deng, Q. (2023). TAPRec: time-aware paper recommendation via the modeling of researchers' dynamic preferences. *Scientometrics*, *128*(6), 3453–3471. https://doi.org/10.1007/s11192-023-04731-4

Lin, S. C., & Lin, J. (2023). A Dense Representation Framework for Lexical and Semantic Matching. *ACM Transactions on Information Systems*, *41*(4). https://doi.org/10.1145/3582426

Momeni, F., Mayr, P., & Dietze, S. (2023). Investigating the contribution of author- and publication-specific features to scholars' h-index prediction. *EPJ Data Science*, *12*(1). https://doi.org/10.1140/epjds/s13688-023-00421-6

Nugroho, A. S., Saikhu, A., & Anggraini, R. N. E. (2023). Development of Reviewer Assignment Method with Latent Dirichlet Allocation and Link Prediction to Avoid Conflict of Interest. *Jurnal RESTI*, *7*(4), 837–844. https://doi.org/10.29207/resti.v7i4.4900

Ostendorff, M., Blume, T., Ruas, T., Gipp, B., & Rehm, G. (2022, June 20). Specialized document embeddings for aspect-based similarity of research papers. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. https://doi.org/10.1145/3529372.3530912

Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., & Rehm, G. (2022). Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. *Association for Computational Linguistics*, 11670. https://doi.org/10.18653/v1/2022.emnlp-main.802

Otero, D., Parapar, J., & Barreiro, Á. (2023). Relevance feedback for building pooled test collections. *Journal of Information Science*, *51*, 1379–1396. https://api.semanticscholar.org/CorpusID:258954038

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. https://doi.org/10.48550/arXiv.2205.01833

Ribeiro, A. C., Sizo, A., & Reis, L. P. (2023). Investigating the reviewer assignment problem: A systematic literature review. *Journal of Information Science*. https://doi.org/10.1177/01655515231176668

Singh, A., Arcy, M. D. ', Cohan, A., Downey, D., & Feldman, S. (2023). SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. *Association for Computational Linguistics*, 5548–5566. https://doi.org/10.18653/v1/2023.emnlp-main.338

Stelmakh, I., Wieting, J., Xi, S., Neubig, G., & Shah, N. B. (2025). A Gold Standard Dataset for the Reviewer Assignment Problem. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=XofMHO5yVY

Tan, S., Duan, Z., Zhao, S., Chen, J., & Zhang, Y. (2021). Improved reviewer assignment based on both word and semantic features. *Information Retrieval Journal*, *24*(3), 175–204. https://doi.org/10.1007/s10791-021-09390-8

Yong, Y., Yao, Z., & Zhao, Y. (2021). A framework for reviewer recommendation based on knowledge graph and rules matching. *2021 IEEE International Conference on Information Communication and Software Engineering, ICICSE 2021*, 199–203. https://doi.org/10.1109/ICICSE52190.2021.9404099

Zhao, X., & Zhang, Y. (2022). Reviewer assignment algorithms for peer review automation: A survey. *Information Processing and Management*, *59*(5). https://doi.org/10.1016/j.ipm.2022.103028