RATE: Reviewer Profiling and Annotation-free Training for Expertise Ranking in Peer Review Systems

Weicong Liu* **Zixuan Yang*** **Yibo Zhao** **Xiang Li†**
School of Data Science and Engineering, East China Normal University

Abstract

Reviewer assignment is increasingly critical yet challenging in the LLM era, where rapid topic shifts render many pre-2023 benchmarks outdated and where proxy signals poorly reflect true reviewer familiarity. We address this evaluation bottleneck by introducing LR-bench, a high-fidelity, up-to-date benchmark curated from 2024–2025 AI/NLP manuscripts with five-level self-assessed familiarity ratings collected via a large-scale email survey, yielding 1,055 expert-annotated paper–reviewer–score annotations. We further propose a reviewer-centric ranking framework that distills each reviewer's recent publications into compact keyword-based profiles and fine-tunes an embedding model with weak preference supervision constructed from heuristic retrieval signals, enabling to match each manuscript against a reviewer profile directly. Across the LR-bench and the CMU gold-standard dataset, our approach consistently achieves state-of-the-art performance, outperforming strong embedding baselines by a clear margin. We release LR-bench at https://huggingface.co/datasets/Gnociew/LR-bench, and an github repository at https://github.com/Gnociew/RATE-Reviewer-Assignment.

1 Introduction

As a cornerstone of modern scientific research, the peer review system plays a crucial role in helping scientists evaluate submissions, providing constructive feedback, and safeguarding academic integrity (Black et al., 1998; Thurner and Hanel, 2011; Bianchi and Squazzoni, 2015). An expert reviewer can substantially improve a manuscript; however, if a reviewer lacks relevant domain expertise, the process may waste time for both authors and reviewers. Consequently, the selection of reviewers requires significant care. With the rapid growth of computer science, especially in the area of artificial intelligence, the number of submissions to conferences and journals has surged (Ribeiro et al., 2023; Shah, 2022), making manual reviewer assignment increasingly impractical. Therefore, developing effective reviewer-assignment algorithms has become essential.

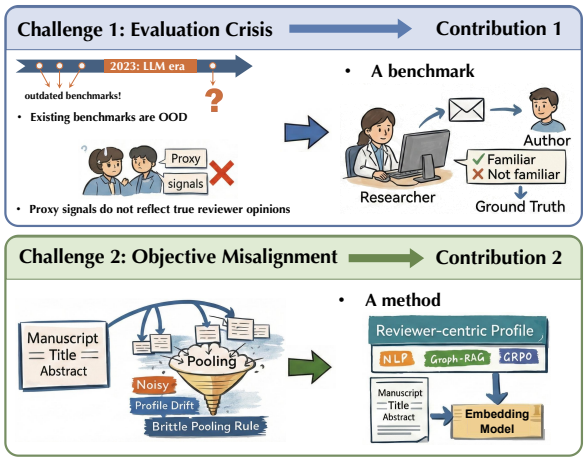

Figure 1: Challenges of reviewer assignment.

Despite its significance, the field of reviewer assignment currently faces two critical challenges. As illustrated in Figure 1, the first challenge lies in the evaluation crisis: **the lack of high-fidelity, open-source, and up-to-date benchmarks.** The exponential surge in AI and NLP research, particularly in the Large Language Model (LLM) era, has created a severe temporal shift. Existing benchmarks (Stelmakh et al., 2025; Singh et al., 2023), most of which were created before 2023, are increasingly out-of-distribution (OOD) and fail to reflect contemporary research topics. Further, many datasets (Mimno and McCallum, 2007; Zhang et al., 2025b) rely on third-party annotations rather than direct expert feedback, which often fail to capture the nuanced expertise.

The second challenge is the misalignment be-

*Equal contribution.
†Corresponding Author: xiangli@dase.ecnu.edu.cn

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43978–43996
July 2-7, 2026

tween training objectives and the inference goal in prior work. Many methods (Stelmakh et al., 2025; Zhang et al., 2025b; Hsieh et al., 2024) are optimized to retrieve papers that are most similar to the target manuscript, rather than to retrieve reviewers who are truly suitable for it. As a result, they do not directly model a reviewer's expertise; instead, they approximate reviewer relevance by aggregating similarities between the manuscript and the reviewer's publications using simple pooling rules (e.g., mean, max, or percentile). This design is sensitive to noise in a reviewer's publication list, leading to "profile drift", and it also relies on manually choosing a pooling strategy, which is often brittle and dataset-dependent. For example, a reviewer whose primary expertise is retrieval-augmented generation (RAG) may have coauthored a paper on graph learning without being a domain expert in graphs. Under max pooling, this single off-topic publication can dominate the aggregated score, causing the reviewer to be incorrectly ranked highly for a graph-focused manuscript.

To address the lack of high-fidelity, up-to-date benchmarks, we first introduce LR-Bench, a high-fidelity and up-to-date benchmark specifically designed to reflect the contemporary research landscape. Our benchmark centers on manuscripts curated from leading AI and NLP conferences within the last two years (2024–2025), directly bridging the content gap created by the recent surge in LLM research. To ensure high data quality, we employ a multi-stage curation process: for each manuscript, we retrieve a candidate pool of reviewers via content-based filtering, and subsequently collect five-level self-assessed Likert familiarity ratings through a large-scale email survey. This approach yields 1,055 high-fidelity paper-reviewer-score pairs, providing a gold standard reflecting real-world expert judgment.

Besides the benchmark, we propose a novel reviewer profiling-ranking algorithm that moves beyond the limitations of heuristic pooling. To overcome the limitations of heuristic pooling and profile drift, we propose a novel LLM-augmented reviewer profiling and self-supervised ranking framework. Specifically, we utilize LLMs to distill core keywords from a reviewer's publication history, synthesizing them into a structured natural language profile that captures their essential expertise. To train a robust assignment model without manual labels, we develop an automated data construction scheme based on pseudo-labeling. For any target manuscript, we retrieve potential candidates via semantic similarity and employ BM25(Robertson and Zaragoza, 2009; Li et al., 2021; Fensore et al., 2025) scores between the manuscript and reviewer profiles as a weak supervision signal to identify positive and hard negative pairs. Through contrastive learning on these synthesized pairs, our model learns to bridge the gap between explicit keyword matching and deep semantic expertise alignment. Experimental results on both LR-Bench and the CMU gold standard dataset (Stelmakh et al., 2025) demonstrate that our approach achieves state-of-the-art performance, surpassing many existing methods that rely on expensive annotated data.

In summary, our contributions are as follows:

- **A high-fidelity contemporary benchmark**: We release a benchmark of 1,055 expert-verified paper-reviewer-score pairs from 2024-2025, establishing a high-fidelity ground truth for the rigorous evaluation of modern assignment systems.
- **An LLM-based profiling method**: We synthesize reviewer expertise into structured natural language profiles using LLM-distilled keywords, effectively eliminating the profile drift issues in traditional heuristic approaches.
- **Zero-annotation training paradigm**: We design a self-supervised training strategy using BM25-guided pseudo-labeling, which allows the model to learn deep semantic expertise alignment without the need for any human-annotation.

2 Related Work

2.1 Reviewer Assignment Benchmarks

Existing benchmarks for reviewer assignment are primarily divided into two categories based on their annotation sources. The first category leverages proxy signals, such as authorship and keyword similarity (OpenReview, 2022; Karimzadehgan et al., 2008; Singh et al., 2023), or manual labels provided by third-party annotators (Mimno and McCallum, 2007; Zhang et al., 2025b). While these approaches are more accessible, they often introduce significant noise and suffer from limited annotation fidelity. The second category relies on reviewer self-assessments (Dumais and Nielsen, 1992; Rodriguez and Bollen, 2008; Stelmakh et al., 2025), which are generally regarded as the "gold standard" due to their high reliability. However,

these datasets are difficult to collect and many existing ones have become outdated, thereby failing to capture recent research trends. In this work, we follow the second paradigm and introduce a contemporary dataset and construction pipeline that continuously scales while preserving high-fidelity, up-to-date relevance labels.

2.2 Reviewer Assignment Methods

Prior work on reviewer assignment is typically framed as a two-stage process: estimating paper–reviewer relevance and allocating reviewers to papers under practical constraints. In the relevance estimation stage, existing methods can be broadly categorized into three classes: (i) explicit-feedback approaches (Benferhat and Lang, 2001; Tayal et al., 2014; Fiez et al., 2020) that leverage human-provided signals such as reviewer interests and bids; (ii) content-based approaches (Tan et al., 2021; Aitymbetov and Zorbas, 2025; Zhang et al., 2025b) that compute similarity between reviewer profiles and the submitted manuscript using lexical matching, statistical models, or embedding-based representations; and (iii) network-based approaches (Rodriguez and Bollen, 2008; Xu and Du, 2013) that exploit relational signals from citation, co-authorship, or collaboration graphs. These approaches are often complementary and are frequently integrated into multi-stage retrieval-and-reranking pipelines. Following relevance estimation, the allocation stage focuses on optimizing assignments to satisfy various objectives such as aggregate similarity (Goldsmith and Sloan, 2007; Charlin and Zemel, 2013), topic coverage (Karimzadehgan and Zhai, 2009), or fairness (Garg et al., 2010; Stelmakh et al., 2021). These frameworks generally model the problem as a constrained optimization task to limit reviewer workload and meet review requirements. In this work, we focus on the stage of estimating paper-reviewer relevance and introduce a novel algorithm that provides more accurate scores for matching.

3 Dataset Construction

To address the challenge of lacking high-fidelity, open-source, and up-to-date benchmarks, we propose LR-Bench, which consists of (i) a large unlabeled arXiv corpus used to construct paper/author metadata and candidate reviewer pools, and (ii) a labeled benchmark subset with self-reported expertise ratings collected via email outreach.

3.1 Data Source & Preprocessing

We construct an up-to-date, unlabeled corpus by crawling recent papers from arXiv, the largest publicly accessible and continuously updated preprint repository with broad coverage across computer science. We focus on five representative sub-areas: Artificial Intelligence (`cs.AI`), Computation and Language (`cs.CL`), Computer Vision and Pattern Recognition (`cs.CV`), Information Retrieval (`cs.IR`), and Machine Learning (`cs.LG`). Collectively, these sub-areas cover major CS/AI research areas with broad topical diversity and abundant recent submissions, providing a practical testbed for reviewer assignment in modern CS venues. To stay aligned with fast-evolving research trends in the LLM era and to reduce temporal drift in topics and expertise signals, we restrict the collection to papers whose last revision date falls within a two-year window (Oct 2023-Oct 2025).

For each paper, we retrieve the external metadata, including the title, arXiv identifier, and PDF URL via the arXiv API. Subsequently, we download the corresponding PDFs and employ GROBID (Lopez, 2009) to extract the title, abstract, author list, affiliations (when available) and email addresses (when available). To ensure data integrity, we perform a cross-source consistency check: a paper is discarded if its PDF-extracted title deviates significantly from the arXiv metadata. Furthermore, we apply rigorous filtering to exclude entries with missing essential fields (title, abstract, or authors) or corrupted text. For papers with multiple revisions, only the most recent version within our temporal window is retained. The corpus comprises 161,228 unique papers with high-fidelity metadata.

To ensure consistent author identities for downstream reviewer assignment, we perform author disambiguation via a precision-oriented hierarchical strategy. Following metadata normalization, we reconcile author entries by prioritizing exact email matches, followed by exact affiliation matches when emails are unavailable. For non-exact matches (e.g., institution variants or abbreviations), we employ an LLM to semantically verify identity. Pairs are merged only upon exact metadata alignment or LLM confirmation; otherwise, they remain distinct. While this conservative approach may leave some cross-email identities split, it effectively mitigates homonym conflation under sparse metadata. The resulting corpus identifies 513,877 unique authors, providing a robust foundation for

paper-author relations.

3.2 Query Sampling & Candidate Recall

Based on the full unlabeled corpus, we sample 4,000 query papers for ground-truth collection using equal-size stratified sampling across the five sub-fields (800 papers per sub-area). For each paper p in our corpus, we formulate its textual representation x_p by concatenating its title and abstract. We then employ a pre-trained Sentence-BERT encoder to map each x_p into a dense embedding space. All paper embeddings are indexed in a high-performance vector database faiss (Douze et al., 2025) to facilitate efficient search.

To construct a recall set of potential reviewers for each query paper, we implement a two-stage pipeline consisting of optimistic dense vector retrieval and Conflict of Interest (COI) (Tang et al., 2010; Wu et al., 2018; Leyton-Brown et al., 2024) filtering. To identify potential reviewers for a given query paper q, we retrieve all papers from the database with a similarity score exceeding a predefined threshold τ. The initial candidate reviewer set, denoted as $\mathcal{C}_q$, is formed by the authors of these retrieved papers. To ensure the integrity and objectivity of the assignment process, we implement a rigorous COI filtering mechanism. Specifically, we exclude (i) any author of the query manuscript itself and (ii) any individual who shares a direct co-authorship history with the authors of q. This procedure yields the final refined candidate reviewer set $\mathcal{C}_q^{\text{COI}}$ for the subsequent ground-truth collection.

3.3 Expertise Ground Truth Collection

We collect ground-truth via email outreach: we contact candidates in the reviewer pool and ask them to rate their familiarity/expertise for the query papers they are matched to by our retrieval pipeline, after reading the title and abstract of each paper. As an annotator-screening measure, we only contacted candidates who had published at least three papers in the past two years, thereby prioritizing active experts and improving the reliability of self-assessed familiarity labels. This aims to capture reviewer expertise directly rather than relying on weak proxies such as topical overlap. To help reduce subjectivity and encourage consistent interpretation across participants, we adopt a five-level Behaviorally Anchored Rating Scale (BARS) (Garine, 2014; Kell et al., 2017) ranging from top expert to no expertise. Details of the scale are provided in Appendix C.1.

To minimize disruption to candidate reviewers, we aggregate and consolidate assignments for each candidate across queries and cap the workload at six query papers per contacted candidate. Each outreach email explicitly states the purpose and intended use of the collected feedback; the full email template is included in the Appendix B. By December 19, 2025, we received 407 responses, yielding 1,069 reviewer–paper ratings. We then performed data cleaning and quality control to reduce the impact of low-effort or unreliable responses. First, we removed duplicate ratings for the same reviewer–paper pair, keeping only the most recent submission (8 duplicate ratings removed). Next, we computed summary statistics to identify potentially suspicious patterns, such as near-zero rating variance, highly imbalanced scores that insensitive to paper content, or cases where multiple responses from the same annotator were submitted within an unusually short time interval. This process flagged 23 ratings for further inspection, and manual review led to the removal of 6 additional ratings. Finally, we obtain 1,055 reviewer-paper ratings.

3.4 Benchmark Construction & Statistics

Finally, we convert the labeled subset into an evaluation benchmark and report key statistics.

Using arXiv categories as a reproducible proxy, we summarize the topical composition of the final labeled data by the category of the query paper. The 1,055 retained reviewer–paper ratings are distributed across the five sampled categories as follows: AI: 206, CL: 212, CV: 264, IR: 157, and LG: 216. This suggests that the collected labels are not overly concentrated in any single category.

The resulting benchmark consists of pointwise and pairwise supervision derived from the collected ratings. Each raw annotation is a pointwise record (p, r, y), where p is a query paper, r is a candidate reviewer, and $y \in \{1, \ldots, 5\}$ is the BARS expertise rating. From these, we derive pairwise preference tuples to support ranking-based training and evaluation: (i) paper-centric tuples (p, r^+, r^-), where r^+ is rated as more expert than r^- for the same paper, and (ii) reviewer-centric tuples (r, p^+, p^-), where the reviewer indicates higher familiarity with p^+ than p^-. In both cases, we form a preference whenever the two ratings differ, treating the higher-rated item as preferred and discarding ties. In total, the benchmark comprises 212 paper-centric and 1,184 reviewer-centric preference tuples.

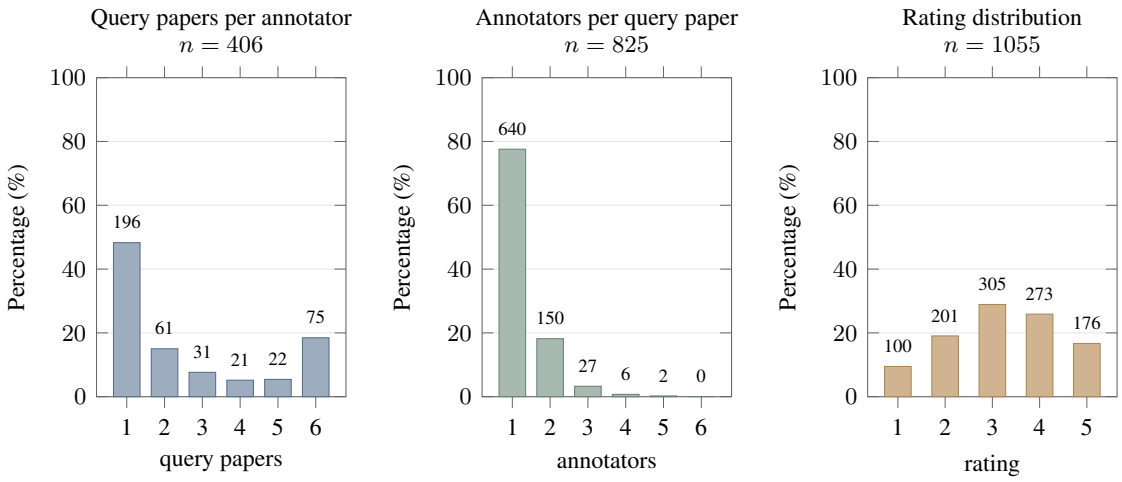

Figure 2: Coverage and supervision statistics of the benchmark. From left to right, the panels show the distributions of query papers per annotator, annotators per query paper, and rating distribution. The y-axis shows percentages, and raw counts are annotated above the bars.

We report coverage and supervision density in Figure 2, including (1) queries per participant (capped at 6), (2) participants per paper, and (3) the $\{1, \ldots, 5\}$ rating distribution, exposing the benchmark's sparsity and long-tail behavior.

4 Reviewer-centric Ranking

To address the challenge of misalignment between training objectives and the inference goal, we introduce RATE, an annotation-free training pipeline to rank candidate reviewers for a query paper. We formalize this task as a ranking problem. Given a query manuscript q and a candidate reviewer set $\mathcal{C}_q^{\text{COI}}$, our goal is to learn a scoring model f_θ. This model computes a relevance score between each candidate reviewer $r \in \mathcal{C}_q^{\text{COI}}$ and the query q, which is subsequently used to rank the candidates in descending order of their predicted affinity.

4.1 Reviewer Profiling

To better capture a reviewer's expertise and address the profile drift issue, we depart from the traditional approach that computes pairwise similarity between the query manuscript and each of the reviewer's historical publications before aggregating these individual scores. Instead, we leverage the extensive internal knowledge and zero-shot capabilities of LLMs to synthesize the reviewer's entire publication history into a cohesive profile. Then, we can obtain a single, holistic embedding for this synthesized profile to perform the final matching. This "synthesize-then-embed" strategy allows the model to grasp a unified research trajectory rather than relying on paper-level comparisons.

Specifically, for each reviewer r and their associated publication history $\mathcal{P}_r$, we employ an LLM to distill a list of salient keywords from each individual paper $p \in \mathcal{P}_r$. These keyword lists are then aggregated into a single comprehensive keyword collection. Notably, we deliberately retain duplicate keywords during this aggregation, which ensures that the frequency of recurrence for specific terms serves as a proxy for the reviewer's level of expertise and familiarity within those particular research sub-domains. Finally, we linearize this frequency-preserving collection into a natural-language sentence by appending the keywords, joined by commas, to a fixed prefix (e.g., "*The reviewer's research keywords include:*"). This yields a unified textual representation x_r for the reviewer profile r. With this textual profile, reviewer-paper matching is cast as a heterogeneous text retrieval problem, where reviewer expertise descriptions and paper abstracts play distinct semantic roles but are required to be aligned in a shared embedding space.

4.2 Dual-view Heuristic Preference Alignment

Despite being framed within a text-based retrieval interface, reviewer–paper matching fundamentally involves heterogeneous and semantically asymmetric text types, which general-purpose embedding models often struggle to align effectively in specialized domains. Therefore, we propose a dual-view, annotation-free preference data construction pipeline to construct training data and adapt these

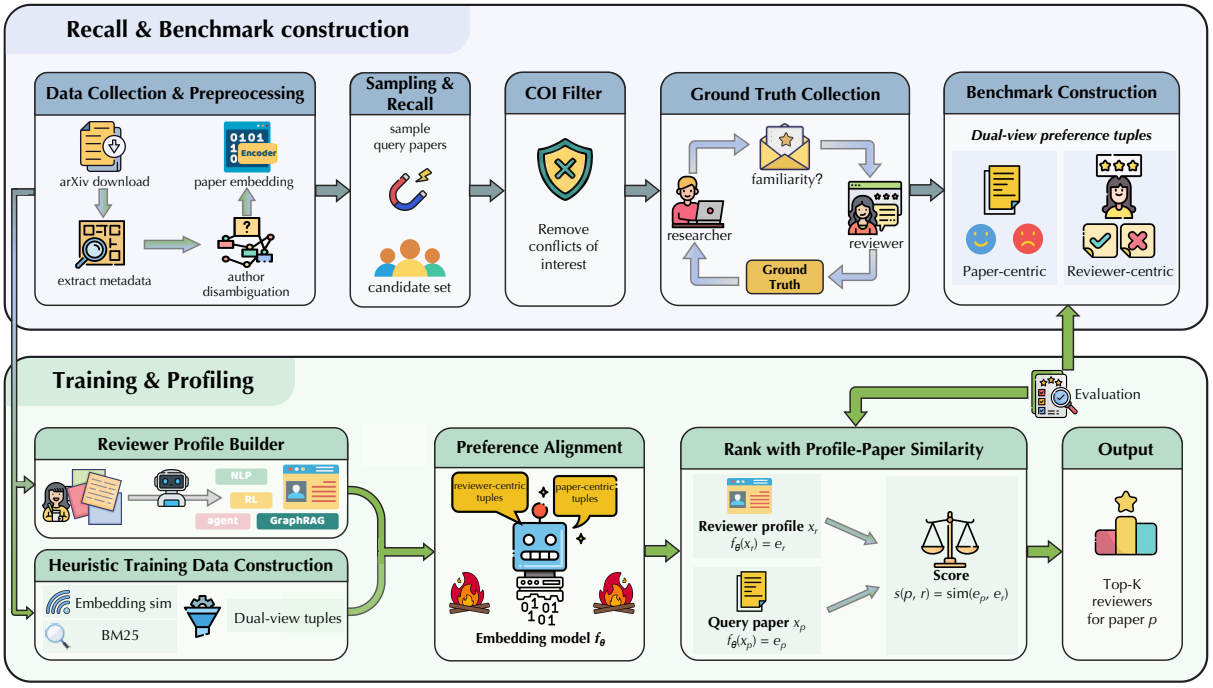

Figure 3: Overview of pipline.

models to this specific task. To unify the optimization process, we define each training instance as a preference triplet (a, c^+, c^-), where a represents the anchor and c represents a candidate. This formulation covers two symmetric views: Paper-centric view ($a = q, c = r$): Finding the better-matched reviewer r^+ over r^- for a manuscript q. Reviewer-centric view ($a = r, c = q$): Identifying the query paper q^+ that better aligns with a reviewer r's expertise than q^-. This dual-perspective optimization helps the model learn a more robust representation of expertise space (Li et al., 2021).

To enable the model to recognize that keyword frequency directly reflects a reviewer's expertise in a sub-domain, we employ BM25, a lexical retrieval method sensitive to term frequency, to construct training data. Our core philosophy is to prioritize high precision over high recall; we prefer to exclude potentially noisy hard samples to ensure the cleanliness of the training signal. Specifically, we apply this logic to both paper-centric and reviewer-centric pairs. For any given anchor (either a query paper or a reviewer profile), we first obtain a semantically relevant candidate set using embedding similarity, and then rank the candidates within this set using BM25. We select the top-ranked candidate as the positive sample (r^+ or q^+). To control difficulty, we choose candidates with scores approximately one-tenth and one-third of the positive score as easy and hard negatives, respectively. This symmetric strategy provides high-quality, annotation-free training data from both perspectives.

Based on the unified triplets (a, c^+, c^-), we optimize the model f_θ using a multi-task loss function. The first component is a pairwise ranking loss $\mathcal{L}_{\text{pair}}$, which encourages the anchor a to be closer to the positive candidate c^+ than to the negative c^- in the embedding space. Formally:

$$\mathcal{L}_{\text{pair}} = -\log \sigma \left(\frac{s(a, c^+) - s(a, c^-)}{\tau} \right), \quad (1)$$

where $s(\cdot, \cdot)$ denotes the cosine similarity score and τ is a temperature hyper-parameter. The second component is a contrastive cross-entropy loss $\mathcal{L}_{\text{ce}}$, which enhances the model's discriminative power by pulling the anchor and positive candidate together while pushing away other candidates in the batch. Formally:

$$\mathcal{L}_{\text{ce}} = -\log \frac{\exp(s(a, c^+)/\tau)}{\sum_{c \in \mathcal{B}} \exp(s(a, c)/\tau)}, \quad (2)$$

where $\mathcal{B}$ denotes all the candidates within the mini-batch. The final training objective is a weighted sum of two terms: $\mathcal{L} = \mathcal{L}_{\text{pair}} + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}}$. To ensure efficient domain adaptation, we implement f_θ by fine-tuning a pre-trained embedding model via Low-Rank Adaptation (LoRA) (Hu et al., 2022).

5 Experiments

Due to space limitations, we provide the ablation study in Appendix A.1, extended comparisons with graph- and recommendation-based methods in Appendix A.2, the analysis of pooling strategies in Appendix A.3, and the evaluation of SPECTER2 variants in Appendix A.4.

5.1 Experimental Setup

Dataset. For our method, both training and validation data are generated from our data construction pipeline, without human annotation. The test data come from the LR-Bench and the CMU gold standard dataset. From both datasets, we derive pairwise preferences from sparse labels for ranking evaluation as stated in Section 3.4. To avoid data leakage, we utilize a paper-level holdout protocol: test papers are strictly excluded from all training data construction, while reviewers may overlap across splits, reflecting the practical setting of ranking known reviewers for unseen manuscripts.

Reviewer profiles. For each reviewer r, the profile text x_r is synthesized from their publications over the preceding two years. In our approach, x_r is a keyword-based profile generated via the pipeline detailed in Section 4.1, utilizing Qwen3-Max[1] and GLM4.6[2] as backbone Large Language Models (LLMs). To maintain experimental consistency and ensure a fair comparison, the publication history for all baseline methods is restricted to the same two-year window with ours.

Training details. We train Qwen3-Embedding-8B and Qwen3-Embedding-0.6B (Zhang et al., 2025a) with LoRA adapters on 3,000 dual-view heuristic preference tuples. Detailed hyperparameter settings are provided in Appendix D.

5.2 Baseline Methods

To provide a comprehensive evaluation, we compare our proposed methods against several state-of-the-art baselines, categorized into three groups:

(i) **Statistical-based Method**. We include TPMS (Charlin and Zemel, 2013) that calculates relevance using TF-IDF similarity between a reviewer's publication history and the target paper. (ii) **Embedding-based Models**. These approaches leverage dense embeddings to capture semantic relevance: BERTScore (Zhang* et al., 2020) utilizes contextual embeddings to measure semantic similarity via soft alignment. SciBERT (Beltagy et al., 2019) is a BERT-based model pre-trained on scientific corpora. The SPECTER family, including SPECTER (Cohan et al., 2020), SciNCL (Ostendorff et al., 2022), and SPECTER 2 (Singh et al., 2023), enhances scientific representations by leveraging citation links and sophisticated sampling strategies. CoF (Zhang et al., 2025b) is a factor-aware framework that employs instruction tuning and a coarse-to-fine search strategy. ACL (Graham Neubig and Cohn., 2021) uses contrastive training on non-contiguous abstract segments to identify similarity. (iii) **LLMs**. We also evaluate DeepSeek-V3.2 (DeepSeek-AI et al., 2025) and Qwen3-Max (Qwen, 2025) by prompting them to score reviewer-paper compatibility in zero-shot.

For completeness, additional details and specific prompts are provided in Appendices E and F.3. We tune hyperparameters of all baselines to maximize their performance on our evaluation datasets.

5.3 Evaluation Protocol and Metrics

To evaluate our framework, we employ expertise-aligned loss, accuracy (Acc.), and human evaluation to capture ranking quality and practical utility.

Following Stelmakh et al. (2025), we use a normalized ranking loss $\mathcal{L} \in [0, 1]$ as our primary metric. We unify author-centric and paper-centric perspectives by constructing preference pairs (x, y) where the ground-truth familiarity $\epsilon_x > \epsilon_y$. The loss penalizes misordered predictions by the magnitude of their label difference:

$$\mathcal{L} = \frac{\sum_{(x,y)\in\mathcal{P}} \mathcal{I}(s_x < s_y) \cdot |\epsilon_x - \epsilon_y|}{\sum_{(x,y)\in\mathcal{P}} |\epsilon_x - \epsilon_y|}, \quad (3)$$

where $\mathcal{P}$ is the set of all valid pairs, $\mathcal{I}$ is the indicator function, and s is the predicted similarity. This metric represents the ratio of the model's error to that of a worst-case adversarial ranker. In addition to the weighted loss, we report accuracy, which measures the ratio of pairs where the model correctly predicts the expertise ordering, providing a direct assessment of the model's accuracy.

5.4 Main Results

As illustrated in Table 1, when utilizing Qwen3-Max as the backbone LLM for reviewer profiling

[1]https://help.aliyun.com/zh/model-studio/qwen-api-reference

[2]https://docs.bigmodel.cn/cn/guide/models/text/glm-4.6

Table 1: **Main Results.** Comparison of our method against state-of-the-art baselines, where our variants are denoted as [Profiling LLM] + [Embedding Model]. The **best results** are bold, and the runner-up results are underlined. For embedding-based methods, we report results using the pooling strategy that achieves the best results.

	Loss (↓)				Acc. (↑)			
Algorithm	**LR-PC**	**LR-RC**	**Gold**	**Avg.**	**LR-PC**	**LR-RC**	**Gold**	**Avg.**
Statistical-based Methods								
TPMS	0.2646	0.2333	0.2811	0.2597	70.28%	72.30%	71.89%	71.49%
Embedding-based Methods								
ACL	0.3338	0.3038	0.3163	0.3180	61.32%	65.96%	68.37%	65.22%
CoF	0.2939	0.2218	0.2564	0.2574	65.57%	73.31%	74.36%	71.08%
BERTScore	0.2846	0.339	0.3216	0.3153	65.57%	62.08%	67.84%	65.16%
SciBERT	0.4016	0.4410	0.3505	0.3977	57.55%	55.15%	64.95%	59.22%
SciNCL	0.2354	0.2114	0.2141	0.2203	69.34%	73.90%	78.59%	73.64%
SPECTER	0.2048	0.2171	0.2672	0.2297	73.58%	73.31%	73.28%	73.39%
SPECTER2 PRX	0.1902	0.2176	0.2144	0.2074	74.06%	72.89%	78.56%	75.17%
LLM-based Methods								
DeepSeek-V3.2	0.2779	0.2351	0.2237	0.2456	50.00%	53.89%	77.36%	60.42%
Qwen3-max	0.2713	0.2289	0.2246	0.2416	47.17%	55.32%	77.54%	60.01%
GLM-4.6 + RATE-0.6B	0.2008	0.1989	0.2350	0.2116	74.53%	74.83%	76.50%	75.29%
Qwen3-Max + RATE-0.6B	0.1955	0.2035	0.2378	0.2123	75.47%	74.32%	76.22%	75.34%
GLM-4.6 + RATE-8B	0.1875	**0.1895**	0.2125	0.1965	75.94%	**75.51**%	78.05%	76.78%
Qwen3-Max + RATE-8B	**0.1795**	0.1926	**0.1991**	**0.1904**	**76.89**%	75.25%	**80.09**%	**77.41**%

and Qwen3-8B-Embedding as the pre-trained embedding model, our approach achieves an average accuracy of 77.41% across two datasets, setting a new state-of-the-art (SOTA) performance. Furthermore, our method demonstrates robust generalizability across various backbone profiling LLMs and pre-trained embedding models, consistently maintaining accuracy levels above 75%. In contrast, among all evaluated baselines, only SPECTER2 PRX exceeds the 75% threshold, further underscoring the superiority and versatility of our framework.

Regarding the simple word-frequency-based TPMS method, we surprisingly observe that despite its algorithmic simplicity, it exhibits remarkable stability, achieving an accuracy exceeding 70% on both datasets. Notably, its average accuracy even outperforms that of modern Large Language Models (LLMs) with vast knowledge bases, suggesting that precise term matching may still be a dominant factor in reviewer assignment, potentially outweighing the complex semantic reasoning provided by general-purpose LLMs.

Regarding embedding-based approaches, methods that incorporate scientific citation network information—such as SciNCL and the SPECTER family—outperform pre-trained embedding models that rely solely on semantic content. This finding underscores the pivotal role of citation relationships in reviewer assignment. Specifically, SPECTER2 PRX achieves the highest performance among these, reaching an average accuracy of 75.17%. However, it relies heavily on complex aggregation strategies for paper similarity, posing significant challenges for practical deployment.

We further explore the effectiveness of LLMs in this task. Paradoxically, we find that even the most advanced models with superior reasoning capabilities struggle to fully capture the intricacies of reviewer profiling. While their average loss remains relatively moderate at approximately 0.24, their accuracy performance is underwhelming, hovering around only 60% ranking as the second and third lowest among all evaluated methods. This discrepancy suggests that while LLMs can effectively distinguish relatively easy samples, they fail to differentiate between hard samples.

We conduct a **human evaluation** to assess the real-world utility of the assignments. We randomly sample 100 papers and task our algorithm and the baselines with retrieving the top-3 candidates from a potential reviewer pool, consistent with the methodology in Section 3.2 (more details are provided in Appendix C.2). We invite human experts to perform blind, pair-wise comparisons

Table 2: Human evaluation results on LR-Bench. Win rate indicates the proportion of cases where our method was preferred over the baseline.

Method	Baseline	LR-Bench		
		Win	Lose	Tie
RATE	TPMS	42%	8%	50%
	SciNCL	35%	17%	48%
	SPECTER2 PRX	44%	12%	44%

of these recommendation lists, determining which algorithm provides more qualified matches. We report the **Win Rate**, defined as the percentage of cases where our algorithm is judged superior to the baseline.

The results in Table 2 further validate the practical utility of our approach. Due to human effort and resource constraints, we focus this study on TPMS and the two top performing embedding based baselines. Our method achieves higher win rates against all selected baselines, specifically reaching a 42% win rate against TPMS and 44% against SPECTER2 PRX. The consistently low lose rates in human trials ranging from 8% to 17% underscore that our algorithm provides reliable reviewer recommendations that are well aligned with senior researchers' professional judgment.

5.5 Qualitative Analysis

We further conduct a qualitative analysis on paper-centric disagreement cases between RATE and the strongest baseline, SPECTER2 PRX. Among these cases, RATE agrees with the expert preference in 29 cases where SPECTER2 PRX does not, whereas SPECTER2 PRX agrees with the expert preference in 23 cases where RATE does not. Inspecting these disagreement cases helps reveal the recurring strengths and failure modes of the two approaches.

We observe two recurring patterns in RATE's wins. First, paper-level semantic matching can be misled by keyword or buzzword overlap, over-rewarding broad topical proximity even when the paper's core requirement is a narrower methodological expertise. Second, semantic matching may overweight domain-specific tokens, favoring reviewers familiar with the application area rather than the underlying method. In contrast, RATE more often aligns reviewers with the core expertise axes required by the submission, including method, evaluation setting, and domain, rather than topical proximity alone.

A representative failure mode of RATE arises when the most relevant expertise evidence is not sufficiently salient in a reviewer's profile, especially for multi-topic profiles. In such cases, adjacent but non-central cues may dominate the ranking, causing RATE to prefer a reviewer whose profile matches the application setting but not the paper's true technical focus. Representative examples are provided in Appendix G.

6 Conclusion

In this paper, we identify the challenges: lack of high-fidelity, up-to-date benchmarks and the misalignment between common training objectives and the goal of reviewer assignment. To address the former, we introduced LR-Bench, a contemporary benchmark curated from recent CS manuscripts with five-level self-assessed familiarity ratings. To address the latter, we introduce RATE, a keyword-based profiling and dual-view preference optimization framework that fine-tunes an embedding model using weak supervision derived from heuristic retrieval signals. Experiments showed consistent gains over strong embedding baselines and prior methods, and ablations supported the benefits of dual-view training.

Limitations

Our method does not explicitly model author-order signals in collaboration-based evidence (e.g., first/last author vs. middle author), which may weaken proxy signals in fields where author order reflects contribution. In addition, the reviewer profile is built from LLM-extracted keywords, which can be noisy or unstable across domains and may propagate errors to ranking. Finally, the approach may be less reliable for cold-start or sparsely published reviewers.

Ethical Consideration

Our work involves collecting human feedback via an email survey in which researchers self-report familiarity/expertise ratings for a set of query manuscripts. This survey is independent of any real conference or journal review process: the collected ratings are used solely for research evaluation and do not reveal or affect any double-blind reviewing decisions. Participation was voluntary,

and respondents could skip questions or stop at any time without any consequences.

Privacy is a primary concern. We contacted potential annotators using email addresses that are publicly available in their published papers. For data release, we will not distribute email addresses or other direct identifiers; instead, we anonymize annotators with random IDs and release only the information necessary for research replication, including manuscript metadata (e.g., title and abstract), the associated ratings, and the list of each annotator's publications from the past two years used to construct reviewer profiles. Before release, we conduct a privacy audit to detect and remove any direct identifiers or uniquely identifying fields that may appear in the collected data (e.g., email headers, names if present, affiliations), and ensure that only the specified anonymized fields are released. We do not solicit free-form textual responses; any unexpected sensitive or offensive content will be filtered or redacted prior to release. All manuscript/publication metadata are obtained from publicly available records and used in a manner consistent with their original access conditions, and we will respect any third-party restrictions on redistribution.

We note that releasing recent publication lists may still enable re-identification via linkage to public bibliographic records; we therefore avoid releasing any additional identifying attributes (e.g., affiliations) and explicitly inform participants of this residual risk and the intended research-only use. Upon publication, we will distribute the released artifacts with an explicit license and terms of use (research-only; no re-identification), and ensure compliance with any third-party source terms; when redistribution is restricted, we will release only identifiers/links and data-construction scripts instead of redistributing the underlying content. The released benchmark and derived artifacts are intended only for research on reviewer–manuscript matching (e.g., evaluation and reproducibility), and must not be used for operational reviewer selection or other non-research purposes.

Finally, automated reviewer assignment systems may be misused (e.g., to manipulate reviewer selection) or may amplify existing biases (e.g., favoring highly visible institutions or prolific researchers and disadvantaging early-career authors). We view our system strictly as a decision-support tool rather than a replacement for human oversight. Any deployment should incorporate standard safeguards such as conflict-of-interest checks and program-committee review of final assignments.

References

Nurmukhammed Aitymbetov and Dimitrios Zorbas. 2025. Autonomous machine learning-based peer reviewer selection system. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 199–207, Abu Dhabi, UAE. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Salem Benferhat and Jérôme Lang. 2001. Conference paper assignment. *International Journal of Intelligent Systems*, pages 1183–1192.

Federico Bianchi and Flaminio Squazzoni. 2015. Is three better than one? simulating the effect of reviewer selection and behavior on the quality and efficiency of peer review. In *2015 Winter Simulation Conference (WSC)*, pages 4081–4089.

N. Black, S. van Rooyen, F. Godlee, R. Smith, and S. Evans. 1998. What makes a good reviewer and a good review for a general medical journal? *JAMA*, pages 231–233.

Laurent Charlin and Richard S. Zemel. 2013. The toronto paper matching system: An automated paper-reviewer assignment system.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, page 7–10, New York, NY, USA. Association for Computing Machinery.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong,

Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Susan T. Dumais and Jakob Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, page 233–244, New York, NY, USA. Association for Computing Machinery.

Chase Fensore, Kaustubh Dhole, Joyce C Ho, and Eugene Agichtein. 2025. Evaluating hybrid retrieval augmented generation using dynamic test sets: Livrag challenge. *Preprint*, arXiv:2506.22644.

Tanner Fiez, Nihar Shah, and Lillian Ratliff. 2020. A super* algorithm to optimize paper bidding in peer review. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 580–589.

Naveen Garg, Telikepalli Kavitha, Amit Kumar, Kurt Mehlhorn, and Julián Mestre. 2010. Assigning papers to referees. *Algorithmica*, pages 119–136.

Armando Garrido Filipe Garine. 2014. The comprehensive assessment of team member effectiveness (catme): personality predicting teamwork competencies.

Judy Goldsmith and {Robert H.} Sloan. 2007. The ai conference paper assignment problem. In *Preference Handling for Artificial Intelligence - Papers from the 2007 AAAI Workshop, Technical Report*, AAAI Workshop - Technical Report, pages 53–57. 2007 AAAI Workshop ; Conference date: 22-07-2007 Through 22-07-2007.

Arya McCarthy Amanda Stent Natalie Schluter Graham Neubig, John Wieting and Trevor Cohn. 2021. Acl reviewer matching. `https://github.com/acl-org/reviewer-paper-matching`. Accessed: 2025-12-23.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 1725–1731. AAAI Press.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 639–648, New York, NY, USA. Association for Computing Machinery.

Jhih-Yi Hsieh, Aditi Raghunathan, and Nihar B. Shah. 2024. Vulnerability of text-matching in ml/ai conference reviewer assignments to collusions. *ArXiv*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Maryam Karimzadehgan and ChengXiang Zhai. 2009. Constrained multi-aspect expertise matching for committee review assignment. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 1697–1700, New York, NY, USA. Association for Computing Machinery.

Maryam Karimzadehgan, ChengXiang Zhai, and Geneva Belford. 2008. Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 1113–1122, New York, NY, USA. Association for Computing Machinery.

Harrison J. Kell, Michelle P. Martin-Raugh, Lauren M. Carney, Patricia A. Inglese, Lei Chen, and Gary Feng. 2017. Exploring methods for developing behaviorally anchored rating scales for evaluating structured interview performance. pages 1–26.

Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. 2024. Matching papers and reviewers at large conferences. *Artif. Intell.*, 331(C).

Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More robust dense retrieval with contrastive dual learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, page 287–296, New York, NY, USA. Association for Computing Machinery.

Weibin Liao, Yifan Zhu, Yanyan Li, Qi Zhang, Zhonghong Ou, and Xuesong Li. 2024. Revgnn: Negative sampling enhanced contrastive graph learning for academic reviewer recommendation. *ACM Trans. Inf. Syst.*, 43(1).

Cheng Liu, Chenhuan Yu, Ning Gui, Zhiwu Yu, and Songgaojun Deng. 2023. Simgcl: graph contrastive learning by finding homophily in heterophily. *Knowl. Inf. Syst.*, 66(3):2089–2114.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 473–474.

David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 500–509, New York, NY, USA. Association for Computing Machinery.

OpenReview. 2022. Paper-reviewer affinity modeling for openreview. https://github.com/openreview/openreview-expertise. Accessed: 2025-12-23.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688.

Qwen. 2025. Qwen3-max. https://qwen.ai/apiplatform. Accessed: 2025-12-23.

Ana Carolina Ribeiro, Amanda Sizo, and Luís Paulo Reis. 2023. Investigating the reviewer assignment problem: A systematic literature review. *Journal of Information Science*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Marko A. Rodriguez and Johan Bollen. 2008. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 319–328, New York, NY, USA. Association for Computing Machinery.

Nihar B. Shah. 2022. An overview of challenges, experiments, and computational solutions in peer review (extended version).

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Ivan Stelmakh, Nihar Shah, and Aarti Singh. 2021. Peerreview4all: fair and accurate reviewer assignment in peer review. 22(1).

Ivan Stelmakh, John Wieting, Sarina Xi, Graham Neubig, and Nihar B. Shah. 2025. A gold standard dataset for the reviewer assignment problem. *Preprint*, arXiv:2303.16750.

Shicheng Tan, Zhen Duan, Shu Zhao, Jie Chen, and Yanping Zhang. 2021. Improved reviewer assignment based on both word and semantic features. *Information Retrieval Journal*, pages 175–204.

Wenbin Tang, Jie Tang, and Chenhao Tan. 2010. Expertise matching via constraint-based optimization. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 34–41.

Devendra Kumar Tayal, P. C. Saxena, Ankita Sharma, Garima Khanna, and Shubhangi Gupta. 2014. New method for solving reviewer assignment problem using type-2 fuzzy sets and fuzzy functions. *Applied Intelligence*, pages 54–73.

S. Thurner and R. Hanel. 2011. Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B*, 84(4):707–711.

Siyuan Wu, Leong Hou U., Sourav S. Bhowmick, and Wolfgang Gatterbauer. 2018. Pistis: A conflict of interest declaration and detection system for peer review management. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1713–1716, New York, NY, USA. Association for Computing Machinery.

Yunhong Xu and Yuanwei Du. 2013. A three-layer network model for reviewer recommendation. In *2013 Sixth International Conference on Business Intelligence and Financial Engineering*, pages 552–556.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Yu Zhang, Yanzhen Shen, SeongKu Kang, Xiusi Chen, Bowen Jin, and Jiawei Han. 2025b. Chain-of-factors paper-reviewer matching. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1901–1910, New York, NY, USA. Association for Computing Machinery.

A Additional Experimental

A.1 Ablation Study

To quantify the contribution of our dual-view training strategy and the impact of preference-based fine-tuning, we evaluate the following four configurations: (1) **Pretrained (Zero-shot)** uses the original Qwen3-Embedding-8B model without any fine-tuning, serving as a zero-shot baseline to quantify the domain gap; (2) **Paper-centric only** fine-tunes the embedding using only paper-centric preference triples (p, r^+, r^-), corresponding to the conventional retrieval-to-reviewer ranking paradigm; (3) **Reviewer-centric only** fine-tunes the embedding model using only reviewer-centric preference triples (r, p^+, p^-), emphasizing modeling a reviewer's historical research trajectory; and (4) **Dual-view** is our complete model, trained with both paper-centric and reviewer-centric preferences under the unified objective described in Section 4.2.

Table 3 shows a substantial domain gap between the off-the-shelf embedding model and reviewer-assignment preferences: the zero-shot **Pretrained** baseline performs markedly worse than all fine-tuned variants. Both single-view fine-tuning settings yield substantial gains—**Paper-centric only** reaches 74.17% average accuarcy, and **Reviewer-centric only** reaches 75.26%—indicating that either perspective provides useful preference supervision for reranking. Importantly, **Dual-view** achieves the best performance across all evaluation subsets, reaching 77.41% average accuarcy and 0.1904 average loss. Compared to the best single-view variant, dual-view brings an additional +2.15 accuarcy points, suggesting that integrating both paper-centric and reviewer-centric preferences provides additional, complementary training signal beyond either view alone.

A.2 Comparison with Graph and Recommendation Models

To evaluate our semantic-driven approach against methods relying on structural and historical interaction data, we extend our comparison on the NIPS benchmark to include advanced recommendation systems (e.g., Wide&Deep (Cheng et al., 2016), DeepFM (Guo et al., 2017)) and graph-based models (e.g., LightGCN (He et al., 2020), SimGCL (Liu et al., 2023), RevGNN (Liao et al., 2024)). We adopt the exact evaluation protocol and baseline metrics from the original RevGNN paper to ensure a fair comparison.

As shown in Table 4, our purely content-based model, RATE-8B, outperforms all interaction-based baselines, achieving a 67.5% improvement in Recall@20 and 21.5% in NDCG@20 over the strongest baseline, RevGNN. By relying on pure semantic alignment rather than explicit historical edges, our approach effectively bypasses the data sparsity bottleneck inherent in standard graph and collaborative filtering models.

While RevGNN retains a slight advantage in Hit Ratio and Precision (HR@20, P@20) by explicitly exploiting network topological biases, RATE-8B remains highly competitive. Despite using zero structural signals, it ranks second-best in these metrics, demonstrating that our advanced semantic embeddings can robustly match or exceed topology-dependent models across all key retrieval metrics.

A.3 Impact of Pooling Strategies

To further investigate how different ways of aggregating reviewer expertise affect matching performance, we compared three pooling strategies: Mean [M], 75th Percentile [75], and Max [X]. Table 5 presents the detailed results across various embedding-based baselines using Loss and accuracy metrics.

A.4 Comprehensive Evaluation of SPECTER2 Variants

We evaluated five SPECTER2 adapter variants: Base, Adhoc Query, Classification, Proximity, and Regression. Each variant was tested across the three pooling strategies, with the full performance metrics presented in Table 6

B Expert Survey Email Template

To collect expert feedback for our dataset construction, we sent out standardized inquiry emails to candidates in the reviewer pool. The template used for this communication, which ensures transparency regarding the research purpose and data usage, is presented in Table 7.

C Human Evaluation

C.1 Behaviorally Anchored Rating Scale

To ensure a faithful ground truth beyond the limitations of administrative records, we collected expert

Table 3: **Ablation Study.** Impact of different training views and data settings on **Qwen3-Embedding-8B**. LR-PC and LR-RC represent our paper-centric and reviewer-centric benchmarks, while Gold refers to the CMU dataset. The average performance columns are highlighted in gray.

	Loss (↓)				Acc. (↑)			
Setting	**LR-PC**	**LR-RC**	**Gold**	**Avg.**	**LR-PC**	**LR-RC**	**Gold**	**Avg.**
(1) Pretrained (Zero-shot)	0.3064	0.3705	0.4307	0.3692	64.46%	61.73%	56.93%	61.04%
(2) Paper-centric Only	0.2247	0.2114	0.2181	0.2181	70.75%	73.56%	78.19%	74.17%
(3) Reviewer-centric Only	0.1955	0.2072	0.2141	0.2056	74.06%	73.14%	78.59%	75.26%
(4) **Dual-view**	**0.1795**	**0.1926**	**0.1991**	**0.1904**	**76.89%**	**75.25%**	**80.09%**	**77.41%**

Table 4: **Comparison with Advanced Recommendation and Graph-based Models.** Performance on the NIPS benchmark. All baseline results are directly sourced from the original RevGNN paper following their exact evaluation protocol.

Datasets	**NIPS**			
Metrics	**R@20**	**N@20**	**HR@20**	**P@20**
TPMS	0.0912	0.0583	0.2104	0.0169
Wide&Deep	0.1219	0.0610	0.2541	0.0203
DeepFM	0.1219	0.0610	0.2541	0.0203
ENMF	0.0070	0.0029	0.0247	0.0012
RecVAE	0.1045	0.0805	0.4890	0.0315
LightGCN	0.0945	0.0543	0.2226	0.0182
SGL	0.0995	0.0666	0.2253	0.0125
SimGCL	0.1408	0.0880	0.4411	0.0242
SHT	0.1142	0.0564	0.2396	0.0110
ApeGNN_HK	0.1267	0.0721	0.2962	0.0179
ApeGNN_PPR	0.1024	0.0754	0.2469	0.0154
RevGNN	0.1526	0.1246	**0.6099**	**0.0420**
RATE-8B	**0.2557**	**0.1514**	0.5000	0.0341

Table 5: **Impact of Pooling Strategies.** Detailed comparison of Mean [M], 75th Percentile [75], and Max [X] pooling strategies across different embedding-based baselines.

		Loss (↓)				Acc. (↑)			
Model	**Pool.**	**LR-PC**	**LR-RC**	**Gold**	**Avg.**	**LR-PC**	**LR-RC**	**Gold**	**Avg.**
BERTScore	[M]	0.2846	0.3398	0.3216	0.3153	65.57%	62.08%	67.84%	65.16%
	[75]	0.2832	0.3330	0.3414	0.3192	66.04%	62.84%	65.86%	64.91%
	[X]	0.3311	0.3152	0.3033	0.3165	61.32%	64.53%	69.67%	65.17%
SciBERT	[M]	0.4016	0.4410	0.3505	0.3977	57.55%	55.15%	64.95%	59.22%
	[75]	0.4668	0.4515	0.3630	0.4271	52.36%	54.05%	63.70%	56.70%
	[X]	0.4282	0.4546	0.3449	0.4092	55.66%	53.63%	65.51%	58.27%
SciNCL	[M]	0.2168	0.2077	0.2601	0.2282	71.70%	73.99%	73.99%	73.23%
	[75]	0.2061	**0.1994**	0.2663	0.2239	73.58%	**74.49%**	73.37%	73.81%
	[X]	0.2354	0.2114	**0.2141**	0.2203	69.34%	73.90%	**78.59%**	73.94%
SPECTER	[M]	0.2380	0.2505	0.3115	0.2667	70.28%	70.35%	68.85%	69.83%
	[75]	0.2247	0.2364	0.2851	0.2487	72.64%	71.96%	71.49%	72.03%
	[X]	0.2048	0.2171	0.2672	0.2297	73.58%	73.31%	73.28%	73.39%
SPECTER2	[M]	0.2141	0.2354	0.2507	0.2334	71.23%	71.88%	74.93%	72.68%
(PRX)	[75]	0.1981	0.2166	0.2436	0.2194	73.58%	73.06%	75.64%	74.09%
	[X]	**0.1902**	0.2176	0.2144	**0.2074**	**74.06%**	72.89%	78.56%	**75.17%**

Table 6: **Comprehensive Evaluation of SPECTER2 Variants.** Comparison of SPECTER2 with different adapters (Base, Adhoc Query, Classification, Proximity, and Regression) under Mean [M], 75th Percentile [75], and Max [X] pooling strategies.

		Loss (↓)				**Acc.** (↑)			
Adapter	**Pool.**	**LR-PC**	**LR-RC**	**Gold**	**Avg.**	**LR-PC**	**LR-RC**	**Gold**	**Avg.**
SPECTER2 (Base)	[M]	0.2035	0.2255	0.2594	0.2295	72.64%	72.72%	74.06%	73.14%
	[75]	0.2114	**0.2082**	0.2484	0.2227	71.70%	**73.65**%	75.16%	73.50%
	[X]	0.2074	0.2150	0.2307	0.2177	72.17%	73.48%	76.93%	74.19%
Adhoc Query	[M]	0.2434	0.2599	0.2817	0.2617	69.34%	69.34%	71.83%	70.17%
	[75]	0.2434	0.2390	0.3042	0.2622	68.87%	71.28%	69.58%	69.91%
	[X]	0.2646	0.2427	0.2717	0.2597	66.04%	70.35%	72.83%	69.74%
Classification	[M]	0.3045	0.2677	0.2709	0.2810	63.68%	68.33%	72.91%	68.31%
	[75]	0.2673	0.2469	0.2639	0.2594	67.45%	69.85%	73.61%	70.30%
	[X]	0.2527	0.2589	0.2893	0.2670	68.40%	69.51%	71.07%	69.66%
Proximity (PRX)	[M]	0.2141	0.2354	0.2507	0.2334	71.23%	71.88%	74.93%	72.68%
	[75]	0.1981	0.2166	0.2436	0.2194	73.58%	73.06%	75.64%	74.09%
	[X]	**0.1902**	0.2176	**0.2144**	**0.2074**	**74.06**%	72.89%	**78.56**%	**75.17**%
Regression	[M]	0.3697	0.3946	0.3751	0.3798	57.08%	59.88%	62.49%	59.82%
	[75]	0.3152	0.3559	0.3581	0.3431	59.91%	63.09%	64.19%	62.40%
	[X]	0.3298	0.3591	0.3465	0.3451	59.43%	62.50%	65.35%	62.43%

self-assessments via email using a five-level Behaviorally Anchored Rating Scale (BARS). The specific criteria provided to the participants are defined as follows:

- **5 - Top Expert:** *I am an active researcher in this sub-field; I have recently published work highly relevant to this paper, or I could write a similar paper myself.*
- **4 - Expert:** *I am very familiar with this field; I could reproduce the method in the paper and accurately judge the quality of its technical details.*
- **3 - Knowledgeable:** *I work or research in a related field; I understand the core concepts but have not published papers or worked on projects in this specific sub-direction.*
- **2 - Vague Familiarity:** *I have heard of this field and can understand the abstract, but I am unfamiliar with the specific methodologies or technical details.*
- **1 - No Expertise:** *I completely do not understand this field and cannot understand the terminology or core logic in the text.*

C.2 Human Preference Trial and Evaluation Details

We conducted a human preference study using pairwise comparisons to evaluate our algorithm against competitive baselines, reporting the results in terms of the **win rate**.

The evaluation team consisted of five Master's and PhD students with relevant domain expertise and English proficiency. To ensure ethical labor practices, all judges were compensated at a rate exceeding the local minimum wage. Before the study, we briefed them on the task and ensured they fully understood the objectives and provided informed consent. **To further align judgment standards, each judge was provided with several high-quality evaluation examples as references, demonstrating the application of our criteria (criteria detailed below)**.

Given the complexity of assessing research expertise, we provided the judges with specific guidelines to ensure consistency. When presented with a query paper and two sets of Top-3 recommended reviewers (alongside their publication histories), judges were instructed to select the set that better satisfied the following criteria:

- **Topic Alignment:** The degree of fit between the reviewers' research backgrounds and the query paper's core domain, keywords, and technical methodologies.
- **Expertise Depth:** Whether the reviewers have published high-quality work in relevant fields, ensuring they can evaluate technical contributions rather than just surface-level concepts.

Dear [Scholar Name],

We hope this email finds you well. We are a research team from the [Department/Lab Name] at [University/Institute Name].

With rapidly growing submissions to AI-related conferences, reviewer assignment has become increasingly challenging. To support research in this area, we are constructing a publicly available dataset to advance automated reviewer assignment systems.

Based on your published works (e.g., "*[Example Publication Title]* "), our algorithm identified you as a potential expert for the papers listed below.

Note: this is NOT a request to review a paper. We only seek to collect expert feedback to help construct the dataset, and the collected data will be used solely for academic research purposes. We would be very grateful if you could click the value in the circle below that best reflects your familiarity with each topic.

It takes less than 10 seconds per paper, and your feedback would be incredibly valuable to our study. Thank you very much for your time and help.

Best regards,

[Department/Lab Name]

[University/Institute Name]

Table 7: Standardized email template sent to scholars for expertise verification.

- **Complementary Coverage:** The extent to which the Top-3 set collectively covers different facets of the paper (e.g., for a paper on "RL in Healthcare," a mix of RL and medical informatics experts is preferred over three experts in only one area).

In cases where the two sets were equally qualified, they were instructed to report a **Tie**.

D Experiment Settings

D.1 Computing Facilities.

All experiments are conducted on a single NVIDIA A800-80G GPU.

D.2 Hyperparameter Settings

We fine-tune the Qwen3-Embedding-8B model using LoRA. The key hyperparameters are summarized in Table 8. The task prompts used for query (Q) and reviewer (R) retrieval are detailed in the following paragraph.

Table 8: Hyperparameter settings for fine-tuning.

Category	Hyperparameters (Value)
LoRA	r: 16, α: 32, Dropout: 0.1
Optimization	LR: 2.3e-05, Warmup: 0.05, Epochs: 15 Batch: 4, Accumulation: 1, Seed: 622 τ: 0.0634, Patience: 6
Input	Max Len (Q/R): 2048, Keywords: 512 Weights: CE (0.915), Pair (1.0)

Task Prompts. For the retrieval task, we use: (1) *Query:* "Given a submission title and abstract, retrieve reviewers whose expertise profile matches and who are familiar with the work." (2) *Reviewer-Centric:* "Given a reviewer profile, retrieve papers that match the reviewer's expertise."

E Baselines Methods.

To ensure the reproducibility of our experiments, we summarize the specific implementation sources and model checkpoints for all baseline methods in Table 9.

While most embedding-based models in our study are evaluated using general pooling strategies (i.e., Mean, Max, or Percentile) to aggregate paper-to-paper similarities, we adhere to the original aggregation rules for certain established baselines to ensure a fair comparison:

- **ACL Algorithm:** We follow the established calculation by identifying the top 3 most similar papers and summing their scores weighted by $1/n$, where n is the rank of the paper in terms of similarity.

- **CoF:** This method aggregates multiple relevance components, including semantic, topical, and citation-based features, to calculate the final expertise fit.

Table 9: Implementation sources and model checkpoints for baselines.

Method	Source / Checkpoint
TPMS	niharshah/goldstandard
SciBERT	allenai/scibert_uncased
BERTScore	Tiiiger/bert_score
SPECTER	allenai/specter
SciNCL	malteos/scincl
SPECTER 2.0	allenai/specter2
ACL Algo.	acl-org/matching
CoF	yuzhimanhua/CoF

F Large Language Model Prompts

F.1 Large Language Model Prompt for Author Disambiguation

To handle potential author ambiguity and ensure data quality, we utilize an LLM-based clustering approach for author disambiguation. The specific instructions and matching rules used for this task are presented in Table 10.

F.2 Large Language Model Prompt for Keyword Extraction

To construct comprehensive author profiles for downstream matching, we employ an LLM to extract domain-specific keywords from paper titles and abstracts. This process ensures that the response focus is captured through precise and representative topics. The prompt template is detailed in Table 11.

F.3 Large Language Model Prompt for Evaluation

For evaluating LLMs in a zero-shot setting, we use a structured prompt to guide the model in scoring reviewer expertise. The template is shown in Table 12.

G Representative Qualitative Cases

We inspect all paper-centric disagreement cases between RATE and SPECTER2 PRX. In total, there are 29 cases where RATE agrees with expert preference but SPECTER2 PRX does not, and 23 cases showing the reverse pattern. Below, we present two representative wins of RATE and one representative failure case.

G.1 Case 1: Keyword/Buzzword Trap

Anchor paper. *HASH-RAG: Bridging Deep Hashing with Retriever for Efficient, Fine Retrieval and Augmented Generation*

Preference and model scores.

	Positive	Negative
Expert preference	5	4
RATE	0.721	0.705
SPECTER2 PRX	0.898	0.943

Interpretation. Although the anchor paper contains strong RAG-related surface terms, its core contribution is the use of deep hashing for efficient large-scale retrieval. The most relevant expertise is therefore closer to hashing-based retrieval, nearest-neighbor search, and retrieval evaluation than to general familiarity with RAG. The positive reviewer shows evidence of semantic hashing and hash-based retrieval, whereas the negative reviewer is primarily oriented toward general RAG and retrieval systems. This case suggests that SPECTER2 PRX is influenced by broad topical proximity, while RATE better captures the narrower methodological expertise required by the paper.

G.2 Case 2: Domain-Token Overweighting

Anchor paper. *Abundance-Aware Set Transformer for Microbiome Sample Embedding*

Preference and model scores.

	Positive	Negative
Expert preference	3	1
RATE	0.542	0.471
SPECTER2 PRX	0.873	0.892

Interpretation. The technical core of this paper is a set-based embedding method that models abundance information and permutation invariance in microbiome samples. The most relevant expertise is therefore set modeling and representation learning, rather than broad familiarity with biology or omics terminology. The positive reviewer shows clear evidence of set modeling and permutation-invariant methods, whereas the negative reviewer's profile is dominated by biology, chemistry, and multi-omics topics. This case indicates that SPECTER2 PRX can overweight domain-

Task: Cluster author instances based on institutional similarity.

[Clustering Rules]

1. **Primary Rule:** Focus on *Institutional Similarity*. Instances sharing the same main institution (university, company, research institute) are considered the same author.
2. **Institution Matching:** Group name variations (e.g., "MIT" = "Massachusetts Institute of Technology") and different location formats (e.g., "Alibaba Group" = "Alibaba Inc.").
3. **Department vs. Institution:** Prioritize the main institution over specific departments (e.g., "Dept. of CS, Stanford" = "Stanford").
4. **Empty Institutions:** Instances with empty or null affiliation fields must be placed in separate, individual clusters and never grouped with non-empty ones.

[Input Data]
`Instances:` {instances}

[Output Requirements]
Return the results **strictly** in JSON format:

```
{ "clusters": [[1, 2, 3], [4]] }
```

Ensure all instance IDs are included. Output only the JSON without any explanatory text.

Table 10: Prompting template for LLM-based author disambiguation.

Role: You are an academic keyword extraction assistant.

Task: Extract precise, domain-representative keywords from the research paper's title and abstract. These keywords should represent research domains, subfields, or areas of study most relevant to the paper's focus.

[Input Data]
`Title:` {paper_title}
`Abstract:` {paper_abstract}

[Output Requirements]
Please extract N concise and specific keywords. You must output **ONLY** the keywords as a comma-separated list, with no additional text, explanations, or formatting.

Example Output: Large Language Models, Information Retrieval, Graph Neural Networks

Table 11: Prompting template for LLM-based academic keyword extraction.

Task: Score the reviewer expertise and fit for reviewing the target paper.

[Target Paper]
`Title:` {paper_title}
`Abstract:` {paper_abstract}

[Reviewer Profile]
`Name:` {reviewer_name}
`Recent Publications:` {reviewer_papers}

[Scoring Rubric]
Please choose one integer score from 1 to 5:
5 (Top Expert): Active researcher in this specific sub-field; published highly relevant work; capable of writing a similar paper.
4 (Expert): Very familiar with the general field; can reproduce methods and judge technical quality.
3 (Knowledgeable): Works in a related field; understands core concepts but no direct work in this sub-direction.
2 (Vague Familiarity): General awareness; can understand the abstract but unfamiliar with technical nuances.
1 (No Expertise): No background; cannot understand terminology or core logic.

At the end of your response, you MUST output only one integer from 1 to 5, and nothing else.

Table 12: Zero-shot prompting template for LLM-based evaluation.

heavy vocabulary, while RATE aligns more closely with the paper's methodological requirements.

G.3 Typical Failure Mode: Evidence Not Salient in the Reviewer Profile

Anchor paper. *A New Image Quality Database for Multiple Industrial Processes*

Preference and model scores.

	Positive	Negative
Expert preference	5	3
RATE	0.598	0.654
SPECTER2 PRX	0.899	0.887

Interpretation. A representative failure mode of RATE arises when the most relevant expertise signal is not sufficiently salient in a reviewer's profile, especially for multi-topic profiles. In this case, the positive reviewer is clearly aligned with image or video quality assessment and IQA-style datasets, whereas the negative reviewer's profile is dominated by ISP-related themes such as illumination decomposition, color constancy, and white balancing. RATE appears to over-emphasize these adjacent industrial imaging cues and therefore mis-ranks the pair, while SPECTER2 PRX better matches the anchor paper to IQA/VQA-related expertise. This example points to a concrete improvement direction via profile-salience reweighting and finer-grained subtask disambiguation.