Beyond Paper-to-Paper: Structured Profiling and Rubric Scoring for Paper-Reviewer Matching

Yicheng Pan[1,†], Zhiyuan Ning[2,†], Ludi Wang[3], Yi Du[1,3,*]
[1]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
[2]Westlake University, Hangzhou, China
[3]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
panyicheng25@mails.ucas.ac.cn, ningzhiyuan@westlake.edu.cn, {wld, duyi}@cnic.cn
[†]These authors contributed equally. [*]Corresponding author.

***Abstract*—As conference submission volumes continue to grow, accurately recommending suitable reviewers has become a challenge. Most existing methods follow a "Paper-to-Paper" matching paradigm, implicitly representing a reviewer by their publication history. However, effective reviewer matching requires capturing multi-dimensional expertise, and textual similarity to past papers alone is often insufficient. To address this gap, we propose P2R, a training-free framework that shifts from implicit paper-to-paper matching to explicit profile-based matching. P2R uses general-purpose LLMs to construct structured profiles for both submissions and reviewers, disentangling them into Topics, Methodologies, and Applications. Building on these profiles, P2R adopts a coarse-to-fine pipeline to balance efficiency and depth. It first performs hybrid retrieval that combines semantic and aspect-level signals to form a high-recall candidate pool, and then applies an LLM-based committee to evaluate candidates under strict rubrics, integrating both multi-dimensional expert views and a holistic Area Chair perspective. Experiments on NeurIPS, SIGIR, and SciRepEval show that P2R consistently outperforms state-of-the-art baselines. Ablation studies further verify the necessity of each component. Overall, P2R highlights the value of explicit, structured expertise modeling and offers practical guidance for applying LLMs to reviewer matching.**

***Index Terms*—Reviewer Matching; Expertise Modeling; Large Language Models**

I. Introduction

Peer review is central to trustworthy science, yet rising submission volumes make fully manual paper–reviewer assignment increasingly impractical. This motivates *Automatic Paper–Reviewer Matching*: given a submission, the goal is to identify the most suitable and competent experts to ensure efficient and high-quality reviewing [1], [2]. The key question is how to represent reviewer expertise in a way that reflects the actual decision criteria used by Area Chairs, while remaining scalable to large conferences.

Prior work has explored a range of strategies. Early methods typically concatenate a reviewer's historical publications into a single document and compute keyword similarity (e.g., TF-IDF) against the submission [3]. Graph-based methods construct networks from co-authorship or citations and propagate expertise scores through the graph structure [4]. Embedding-based approaches encode papers into dense vectors to measure semantic proximity [5]–[7]. Most recently, Chain-of-Factors (CoF) learns factor-aware representations to retrieve similar papers and aggregates paper-level scores to rank reviewers [8]. Despite their differences, these approaches share a common core: they model relationships between papers and implicitly represent a reviewer as a composite of their publications.

This work was supported in part by the Natural Science Foundation of China under Grant No. T2322027 and 62442204, and the Key Research Program of the Chinese Academy of Sciences under Grant No. RCJJ-145-24-20.

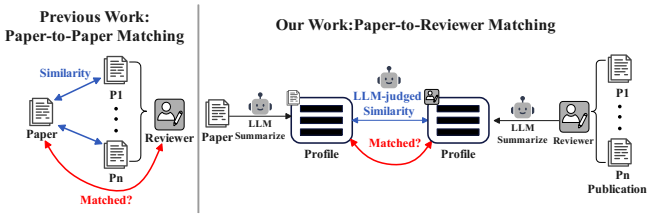


Fig. 1. **Paradigm comparison.** The left panel illustrates traditional **Paper-to-Paper** matching, which relies on similarities between a submission and historical publications. In contrast, the right panel depicts our **Paper-to-Reviewer** framework (P2R). P2R leverages LLMs to synthesize structured profiles, enabling direct and interpretable expertise alignment.

However, reviewer assignment is a fundamentally multi-factor decision. In practice, Area Chairs must assess the specific *paper–reviewer fit*, which depends on explicit dimensions such as the research topics, the methodological toolkit, and the target application setting [7]. A simple paper-to-paper match often misses these critical differences. For example, a reviewer may publish extensively on the same topics as a submission but primarily conduct theoretical analysis rather than the empirical evaluation required by the paper, or work in a different application domain. Such structural mismatches are difficult to resolve by aggregating paper similarities. This motivates moving from paper-based proxies toward a multi-dimensional characterization of both submissions and reviewers.

Recent Large Language Models (LLMs) have demonstrated strong capabilities in long-context understanding, instruction following, and reasoning over scientific text. Leveraging these capabilities, we propose **P2R**, a multi-stage framework that shifts from implicit paper-to-paper matching to explicit profile-based matching. P2R constructs structured profiles for submissions and reviewers and then performs a committee-style assessment under explicit rubrics. Concretely, P2R consists of three stages: (1) **Structured Profiling**, which generates

profiles along three dimensions—Topics, Methodologies, and Applications; (2) **Hybrid Retrieval**, which combines profile matching with embedding retrieval to efficiently shortlist high-recall candidates; and (3) **Rubric-driven Committee Scoring**, where an LLM committee consisting of three dimension-specific evaluators and a holistic Area Chair scores candidates under explicit rubrics to produce the final ranking.

We validate the effectiveness of P2R through experiments on three benchmarks: NeurIPS, SIGIR, and SciRepEval. P2R achieves the best overall performance across all three datasets, surpassing strong pretrained language model baselines and CoF in a *training-free* manner, in contrast to approaches that rely on task-specific training [7], [8]. Comprehensive ablation studies further show that each stage is indispensable and contributes complementary gains, supporting the value of explicit profiling and rubric-based evaluation.

Overall, P2R advances reviewer matching by directly modeling paper–reviewer fit with multi-dimensional signals rather than relying solely on paper representations. Beyond improved accuracy, our framework offers a practical recipe for applying LLMs to reviewer matching: use LLMs to build structured expertise profiles, retrieve candidates with high recall, and enforce decision criteria via rubrics during final scoring.

To support reproducibility, the source code is publicly available at: https://github.com/kg4sci/P2R.

II. RELATED WORK

A. Automatic Paper–Reviewer Matching

Automatic paper–reviewer matching has been widely studied to reduce the manual burden in large-scale peer review. Most pipelines separate (i) learning a relevance function $f(p, r)$ between a submission p and a reviewer r, and (ii) solving a constrained assignment problem with capacity and conflict-of-interest constraints [9], [10]. We focus on the first part and aim to build a relevance function that is both accurate and interpretable. Early systems rely on lexical similarity and topic models, while graph-based methods further exploit co-authorship and citation links to propagate expertise signals [4], [11]. Production systems such as TPMS adopt human-in-the-loop workflows: they provide ranked lists to Area Chairs and use bids to refine relevance estimates [12]. These approaches are strong baselines, but expertise is often reduced to coarse textual or topical similarity [13]. Some work further formulates reviewer assignment as constraint-based optimization and studies robustness under sparse or noisy evidence [3], [14].

B. Scientific Document Representations for Matching

Recent work formulates matching as semantic retrieval. Domain PLMs such as SciBERT provide scientific representations, and citation-informed models such as SPECTER and SciNCL further improve embeddings with citation graphs [5], [6], [15]. They are commonly evaluated on benchmarks such as SciRepEval [7]. To address the limits of single-vector matching, Chain-of-Factors (CoF) combines semantic, topics, and citation signals and applies instruction tuning in a coarse-to-fine framework [8]. However, it still depends on implicit embedding similarity, which can blur distinct expertise dimensions and hide structural mismatches. We instead use explicit multi-dimensional profiles to make mismatches visible.

C. LLMs for Reviewer Recommendation and Decision Making

LLMs increasingly drive reviewer recommendation [16]–[18]. Recent benchmarks use LLMs as zero-shot rankers with pairwise prompting, showing strong direct preference judgments [19]. LLMs also perform well as rubric-based judges, motivating committee-style deliberation rather than a single similarity score [20], [21]. This is supported by the emerging capabilities of LLMs in complex reasoning and multi-step decision making, often elicited through advanced prompting strategies. We follow this direction but enforce explicit structure. We replace generic summaries with schema-consistent profiling that separates *Topics*, *Methodologies*, and *Applications*. We then apply rubric scoring with multiple expert perspectives, which moves from latent correlation to explicit decision simulation.

III. PROPOSED METHOD

A. Problem Setup

Given a submission paper p with text x_p (title and abstract) and a reviewer r with n_r historical publications, let $x_{r,i}$ denote the concatenation of the title and abstract of the i-th historical paper. We further define the reviewer digest as $\bar{x}_r = \mathrm{concat}(x_{r,1}, \ldots, x_{r,n_r})$. The goal is to produce a ranked list of reviewers for each submission. In this work, we focus on designing a relevance scoring function $f(p, r)$ for ranking, while leaving conflict constraints to downstream assignment.

B. Granularity-aware Expertise Profiling

Dense embeddings often collapse distinct competence dimensions into a single scalar, obscuring structural mismatches (e.g., a shared topic but differing methodologies). To explicitly represent expertise, we leverage the instruction-following capabilities of LLMs to construct explicit, multi-dimensional reviewer profiles. We define a structured profile schema $\mathcal{D} = \{\textsc{Topics}, \textsc{Methodologies}, \textsc{Applications}\}$, corresponding to the critical dimensions of paper-reviewer fit.

For any scientific text x, we employ a general-purpose LLM (Qwen3) [22] as a profiler to extract a structured profile $\phi(x)$:

$$\phi(x) = \big\{\mathcal{L}_d(x) \mid d \in \mathcal{D}\big\}, \tag{1}$$

where $\mathcal{L}_d(x)$ represents a list of concise phrases for dimension d (e.g., specific algorithms for METHODOLOGIES). **Reviewer profiling.** For each reviewer r, we query an instruction-following LLM (Qwen3) on the reviewer digest $\bar{x}_r$ with a structured instruction, e.g., *"Analyze the following publication history of a NeurIPS reviewer. Extract the following fields into a JSON object: 1. 'topics' (3-5 high-level research topics); 2. 'methodologies' (3-5 algorithmic approaches); 3. 'applications' (3-5 specific domains)..."*. This prompts the model to output a JSON-only profile $\phi(\bar{x}_r)$ under a strict schema constraint. If a field cannot be supported by evidence, the

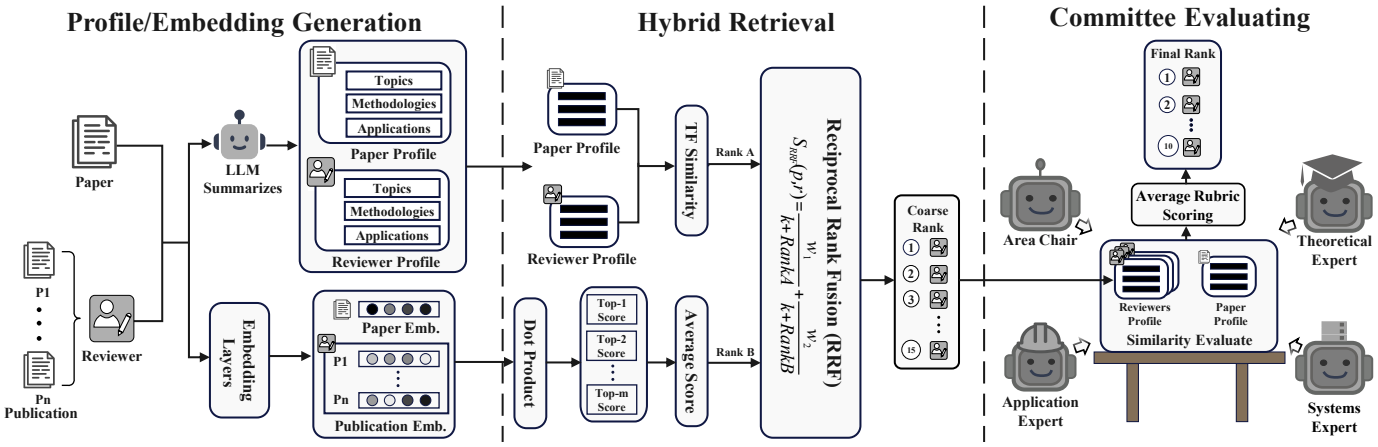


Fig. 2. Overview of P2R. The pipeline has three stages: (1) Profile and embedding generation, where an LLM summarizes structured aspect profiles for both the submission and each reviewer (from their historical publications), and an embedding model encodes the submission and all reviewer-history papers into dense vectors; (2) Hybrid retrieval, which fuses discrete profile matching and semantic matching via RRF to obtain a coarse ranking; and (3) Committee evaluation, where an LLM committee scores candidates with a rubric from multiple expert perspectives to produce the final ranking.

profiler returns an empty list, reducing hallucination risk and ensuring schema consistency.

Submission profiling. Similarly, for each submission p, we apply the same profiler to the title and abstract to obtain $\phi(x_p)$. This creates a shared structured representation space between papers and reviewers, enabling explicit aspect-level matching beyond latent semantic similarity.

Implementation details. P2R is training-free: it does not finetune any model on target datasets. Profiling is performed via Qwen3 API with JSON-only prompting; we use a low temperature and repeat each run three times to mitigate stochasticity, reporting mean performance.

C. Hybrid Retrieval with Reciprocal Rank Fusion

P2R performs coarse candidate selection by fusing two complementary signals: (i) discrete matching over structured profiles, and (ii) continuous similarity via embeddings. Let $\mathcal{R}$ denote the set of candidate reviewers.

Stream A: Discrete profile matching. To enable precise lexical matching, we linearize the structured profile into a discrete string $z(x)$ by joining all phrases across dimensions $\mathcal{D}$ with a space separator:

$$z(x) = \mathrm{concat}\Big(\big(\mathrm{concat}(\mathcal{L}_d(x))\big)_{d\in\mathcal{D}}\Big), \tag{2}$$

where $\mathrm{concat}(\cdot)$ denotes string concatenation and dimensions are concatenated in a fixed order: TOPICS, METHODOLOGIES, then APPLICATIONS.

We thus obtain the linearized submission profile $z_p = z(x_p)$ and reviewer profile $z_r = z(\bar{x}_r)$. We compute TF-cosine similarity:

$$s_{\mathrm{tf}}(p,r) = \cos\big(\mathrm{tf}(z_p), \mathrm{tf}(z_r)\big). \tag{3}$$

$\mathrm{tf}(\cdot)$ denotes raw term-frequency vectors over the profile vocabulary, ℓ_2-normalized before cosine similarity. This stream emphasizes exact overlaps of explicit expertise facets (e.g., application domains or method names), which are often diluted in dense embeddings.

Stream B: General-purpose LLM embeddings over reviewer history. To obtain a robust semantic signal without relying on scientific-domain Pre-trained Language Model, we use NV-Embed as a general-purpose dense embedding model [23]. NV-Embed is a decoder-only LLM-based embedder designed for retrieval and broad text embedding tasks, which employs a dedicated pooling mechanism (latent attention) and contrastive training recipes to produce strong sentence representations. In our framework, we do not perform any additional instruction tuning or finetuning; we directly use the released NV-Embed checkpoint as an off-the-shelf encoder.

Given a submission paper p, we compute its embedding e_p from its title and abstract. Similarly, for each reviewer r, we compute embeddings $\{e_{r,i}\}$ for all their historical papers. We define the reviewer–paper semantic score as the mean cosine similarity over the Top-m most similar history papers:

$$s_{\mathrm{emb}}(p,r) = \frac{1}{m}\sum_{i\in\mathrm{Top}\text{-}m}\big(e_p^\top e_{r,i}\big), \tag{4}$$

where all vectors are ℓ_2-normalized and, to be consistent with prior work, we set $m = 3$. This Top-m aggregation effectively reduces noise from irrelevant historical publications.

Rank Fusion and Candidate Generation. To combine these heterogeneous signals, we convert scores to rankings. Let RankA and RankB denote the rank of reviewer r for submission p derived from s_{tf} and s_{emb}, respectively. We compute the coarse retrieval score via Reciprocal Rank Fusion (RRF) [24]:

$$s_{\mathrm{rrf}}(p,r) = \frac{w_1}{k+\mathrm{RankA}} + \frac{w_2}{k+\mathrm{RankB}}, \tag{5}$$

where k is a smoothing constant and w_1, w_2 are stream weights. We retain the Top-M reviewers as the candidate set $\mathcal{C}_p$ for the subsequent committee evaluation stage. We set $M = 15$ to ensure a sufficient candidate pool for the committee evaluation.

D. Rubric-driven Committee Evaluation

Although hybrid retrieval effectively aggregates lexical and semantic signals to optimize recall, it relies on generalized similarity scores that may mask mismatches in specific dimensions. To better ensure expertise matching, we leverage LLMs to simulate a multi-perspective Area Chair committee. Instead of relying on a single generic score, we query the model under C distinct personas (i.e., Area Chair, Theoretical Expert, Application Expert, and Systems Expert) to ensure a comprehensive evaluation.

Aspect-based Rubric Scoring. Each committee member c evaluates the alignment between profiles $\phi(x_p)$ and $\phi(\bar{x}_r)$ against an explicit rubric, e.g., *"You are a SIGIR Area Chair... Three-dimension rubric: Topics / Methodologies / Applications... For each paper tag, find its best match among the reviewer tags: strong match* $= 1.0$ *(near-synonymous), related match* $= 0.5$*, otherwise* 0*; sum these weights to obtain the per-dimension overlap* overlap_d*..."*. Let $\mathcal{L}_p^d = \mathcal{L}_d(x_p)$ and $\mathcal{L}_r^d = \mathcal{L}_d(\bar{x}_r)$ denote the tag lists of the paper and reviewer in dimension d, respectively. The LLM computes one-to-one overlap and assigns per-dimension scores using coverage and softJaccard thresholds (e.g., where for a dimension d, $\text{coverage}_d = \text{overlap}_d/|\mathcal{L}_p^d|$ and $\text{softJaccard}_d = \text{overlap}_d/(|\mathcal{L}_p^d| + |\mathcal{L}_r^d| - \text{overlap}_d)$; a high match is assigned (e.g., $s_d = 3$) if coverage_d and $\text{softJaccard}_d \geq 2/3$, and a lower s_d otherwise). We use coverage_d to ensure sufficient requirement coverage and softJaccard_d to control profile specificity, preventing high scores from reviewers with overly generic tag sets. The final score is computed as a weighted sum of aspect scores:

$$s_{\text{cont}}^{(c)}(p,r) = \alpha \cdot s_T + \beta \cdot s_M + \gamma \cdot s_A, \tag{6}$$

where s_T, s_M, and s_A denote the rubric scores for TOPICS, METHODOLOGIES, and APPLICATIONS, respectively, and α, β, γ represent their corresponding importance coefficients. To match the annotation schema, we map the continuous score $s_{\text{cont}}^{(c)}(p,r)$ into the discrete label space $\{0,1,2,3\}$ using three thresholds $\tau_1 < \tau_2 < \tau_3$, assigning labels 0, 1, 2, and 3 to the intervals $(-\infty, \tau_1)$, $[\tau_1, \tau_2)$, $[\tau_2, \tau_3)$, and $[\tau_3, \infty)$, respectively. This step is used only to align predictions with the ground truth label space.

Committee aggregation. Let $s_{\text{llm}}^{(c)}(p,r)$ be the final discrete score assigned by the c-th committee member. We aggregate the judgments from all C members by computing the mean:

$$s_{\text{llm}}(p,r) = \frac{1}{C}\sum_{c=1}^{C} s_{\text{llm}}^{(c)}(p,r). \tag{7}$$

Final fusion with retrieval. We define the final relevance function as $f(p,r) = s_{\text{final}}(p,r)$. Let $\text{Rank}_{\text{llm}}(p,r)$ denote the rank of reviewer r for submission p derived from the committee score s_{llm}. We treat the committee ranking as an additional RRF stream with weight w_3 :

$$s_{\text{final}}(p,r) = s_{\text{rrf}}(p,r) + \frac{w_3}{k + \text{Rank}_{\text{llm}}(p,r)}. \tag{8}$$

This preserves high-recall retrieval signals while allowing rubric-based evaluation to correct misranked candidates.

TABLE I
DATASET STATISTICS

Dataset	#Papers	#Reviewers	Conference(s)	# Annotated Pairs
NeurIPS	34	190	NeurIPS 2006	393
SIGIR	73	189	SIGIR 2007	13797
SciRepEval	107	661	NeurIPS 2006, ICIP 2016	1729

IV. EXPERIMENTS

A. Experiment Setup

Datasets. Following prior works [7], [13], [25], We evaluate on three public reviewer-matching benchmarks with expert-curated relevance labels: NeurIPS [13], SIGIR [25], and SciRepEval [7]. Dataset statistics are summarized in Table I. NeurIPS uses a 0–3 relevance scale; SIGIR derives relevance from aspect-level overlap by discretizing the Jaccard similarity between paper and reviewer aspect labels into the same 0–3 scale; SciRepEval unifies augmented NeurIPS and ICIP ratings into the same 0–3 scale. **Baselines.** We compare P2R with the lexical matching baseline TPMS [12]. We include scientific PLM retrievers SciBERT [5], SPECTER [6], and SciNCL [15], as well as the general-purpose dense retriever COCO-DR [26]. We also evaluate SPECTER 2.0 [7] using both PRX and CLF variants under its adapter-based multi-task setting. Finally, we compare with the state-of-the-art Chain-of-Factors (CoF) [8], which combines semantic, topic, and citation factors and applies a coarse-to-fine ranking pipeline.

Evaluation Metrics. To quantitatively evaluate our model, we follow prior work [7]. We adopt Precision@N (P@N) as the primary evaluation metric, specifically reporting P@5 and P@10. For each submission paper p, let $\mathcal{R}_p$ denote the set of candidate reviewers for whom ground-truth relevance annotations exist. We rank all reviewers $r \in \mathcal{R}_p$ based on the relevance scores predicted by our model. Let $\hat{r}_{p,n}$ denote the reviewer ranked at the n-th position in this sorted list. The P@N scores are then calculated under two settings: **Soft** and **Hard**, defined as follows:

$$\begin{aligned} \text{Soft P@}N &= \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \frac{\sum_{n=1}^{N} \mathbb{I}\left(y_{p,\hat{r}_{p,n}} \geq 2\right)}{N}, \\ \text{Hard P@}N &= \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \frac{\sum_{n=1}^{N} \mathbb{I}\left(y_{p,\hat{r}_{p,n}} = 3\right)}{N}. \end{aligned} \tag{9}$$

Here, $|\mathcal{P}|$ denotes the total number of test papers, and $\mathbb{I}(\cdot)$ represents the indicator function. The term $y_{p,\hat{r}_{p,n}}$ denotes the ground-truth relevance score (ranging from 0 to 3) of the reviewer ranked at the n-th position for paper p. Notably, due to the limited number of ground-truth reviewers per submission in NeurIPS (often fewer than ten), we restrict our evaluation to P@5 for this benchmark. For readability, we report P@N as percentages.

TABLE II

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART BASELINES. WE EVALUATE AVERAGE P@N ACROSS NEURIPS, SIGIR, AND SCIREPEVAL BENCHMARKS, WHERE "AVG" DENOTES THE MEAN PERFORMANCE. P2R SURPASSES COMPETITIVE BASELINES ACROSS DIVERSE DOMAINS. BEST RESULTS ARE IN **BOLD**; SECOND-BEST ARE UNDERLINED.

	NeurIPS [13]			SIGIR [25]					SciRepEval [7]				
Method	**Soft P@5**	**Hard P@5**	**AVG**	**Soft P@5**	**Soft P@10**	**Hard P@5**	**Hard P@10**	**AVG**	**Soft P@5**	**Soft P@10**	**Hard P@5**	**Hard P@10**	**AVG**
TPMS [12]	49.41	22.94	36.18	39.73	38.36	17.81	17.12	28.26	62.06	53.74	31.40	24.86	43.02
SciBERT [5]	47.06	21.18	34.12	34.79	34.79	14.79	15.34	24.93	59.63	54.39	28.04	24.49	41.64
SPECTER [6]	52.94	25.29	39.12	39.73	40.00	16.44	16.71	28.22	65.23	**56.07**	32.34	25.42	44.77
SciNCL [15]	54.12	27.06	40.59	40.55	39.45	17.81	17.40	28.80	66.92	55.42	34.02	25.33	45.42
COCO-DR [26]	54.12	25.29	39.71	40.00	40.55	16.71	17.53	28.70	65.05	55.14	31.78	24.67	44.16
SPECTER 2.0 CLF [7]	52.35	24.71	38.53	39.45	38.63	16.16	16.30	27.64	64.49	55.23	31.59	24.49	43.95
SPECTER 2.0 PRX [7]	54.12	27.65	40.89	37.53	37.12	16.44	16.05	26.64	66.17	55.70	33.83	25.61	45.33
CoF [8]	55.68	28.24	41.96	45.57	41.69	**22.47**	17.76	31.87	**68.47**	55.89	34.52	25.33	46.05
P2R (Ours)	**59.41**	**33.38**	**46.40**	**46.85**	**43.29**	21.10	**18.22**	**32.37**	68.41	55.98	**35.05**	**27.58**	**46.76**

Hyperparameter Settings. For PLM baselines, we follow [7] and aggregate paper–paper relevance into paper–reviewer relevance by averaging the top-3 matched papers. For P2R, we set the RRF constant $k = 60$, Top-m aggregation $m = 3$, and candidate size $M = 15$. We set the committee rubric weights $(\alpha, \beta, \gamma) = (0.5, 0.3, 0.2)$. The discretization thresholds are $\mathcal{T} = \{\tau_1, \tau_2, \tau_3\} = \{0.35, 1.35, 2.35\}$. For RRF, we constrain weights to $w_i \in [0, 2]$.

B. Main results

Table II summarizes the comparative performance across all benchmarks. P2R establishes a new state-of-the-art, securing the top rank in 10 out of 13 metrics and ranking second in the remaining three. This superiority is consistent across diverse domains, including machine learning (NeurIPS), information retrieval (SIGIR), and the broad scientific landscape (SciRepEval). Notably, on the NeurIPS dataset, P2R achieves a substantial gain of +5.14 percentage points in Hard P@5 compared to the strongest baseline. Since Hard P@5 strictly measures the retrieval of "very relevant" experts, this gain indicates that explicit aspect profiling captures fine-grained expertise nuances that implicit embeddings often miss. Crucially, P2R achieves these results in a training-free manner, outperforming CoF which relies on task-specific fine-tuning. This suggests that leveraging LLMs for structured profiling and committee-based evaluation serves as a more effective paradigm than relying solely on latent representation learning.

C. Ablation experiment

Our framework introduces two primary technical innovations. First, we implement LLM-driven structured profiling. We summarize schema-consistent expertise profiles from publication histories, explicitly spanning topics, methodologies, and applications. Second, we propose rubric-based score fusion. This module employs an LLM judge to assign aspect-level scores, which we integrate with hybrid retrieval signals via Reciprocal Rank Fusion (RRF). We validate the contribution of these components through a comprehensive ablation study. Specifically, we examine the following variants:

TABLE III

ABLATION ANALYSIS. VALUES ARE MEANS OF SOFT AND HARD P@N. BOLD AND UNDERLINE DENOTE THE BEST AND SECOND-BEST RESULTS.

Model Variant	**NeurIPS**	**SIGIR**	**SciRepEval**
Individual Components			
Discrete Profile Matcher Only	37.28	32.02	43.26
Semantic Matcher Only	45.32	26.79	46.61
LLM Committee Only	39.05	31.13	43.53
Hybrid Retrieval (Profile + Semantic)			
Hybrid Strategy (Discrete + Semantic)	45.62	32.06	46.64
Final Fusion (Hybrid + Committee)			
P2R (Hybrid + LLM Committee)	**46.40**	**32.37**	**46.76**

We compare five variants. **Discrete Profile Matcher Only** ranks reviewers by lexical overlap between flattened structured profiles. **Semantic Matcher Only** ranks reviewers by cosine similarity of dense embeddings. **LLM Committee Only** skips retrieval and directly applies the LLM committee to the full reviewer pool, testing whether LLMs can serve as standalone rankers without pre-filtering. **Hybrid Retrieval** combines the Discrete and Semantic matchers with RRF to produce a candidate list, without the final committee fusing. **P2R (Full Framework)** uses Hybrid Retrieval for candidate generation and then applies the rubric-driven committee for precision ranking, forming the complete coarse-to-fine pipeline.

Based on Table III, we highlight three key observations: **Necessity of Pre-filtering.** The standalone LLM Committee significantly underperforms the full framework. Direct evaluation over the global reviewer pool proves ineffective. The vast search space introduces excessive noise, distracting the model from granular expertise signals. A pre-filtered, high-quality candidate pool is indispensable for precise evaluation. **Benefits of Hybrid Retrieval.** Integrating discrete and continuous signals yields consistent gains. The Hybrid Strategy outperforms both single-stream baselines. This confirms that the structural precision of profiles complements the semantic coverage of embeddings. **Effectiveness of Coarse-to-Fine**

TABLE IV
PERFORMANCE COMPARISON OF PROFILE REPRESENTATIONS. VALUES REPRESENT THE AVERAGE OF SOFT AND HARD P@N METRICS. BEST RESULTS ARE IN **BOLD**

Representation Strategy	NeurIPS	SIGIR	SciRepEval
Unstructured Description	45.81	25.79	46.52
Structured Aspects(Ours)	**46.40**	**32.37**	**46.76**

TABLE V
IMPACT OF PROFILE GRANULARITY. VALUES REPRESENT THE AVERAGE OF SOFT AND HARD P@N METRICS. BEST RESULTS ARE IN **BOLD**; SECOND-BEST ARE UNDERLINED.

Configuration	NeurIPS	SIGIR	SciRepEval
Best Subset ($j = 1$)	<u>46.10</u>	28.90	46.36
Best Subset ($j = 2$)	45.52	<u>30.82</u>	<u>46.68</u>
P2R (Full Profile)	**46.40**	**32.37**	**46.76**

Strategy. The full P2R model achieves optimal performance. Retrieval provides a high-quality candidate pool, and subsequent committee evaluation enhances precision. This validates our design choices. Combining hybrid retrieval for candidate generation and the LLM committee for final refinement yields the best performance.

These results confirm that each P2R component is indispensable. The framework effectively enhances traditional retrieval with LLM-based evaluating, leveraging their distinct strengths for accurate reviewer matching.

D. Impact of Profile Design

We investigate the optimal construction of expertise profiles from two perspectives: representation format and aspect granularity. **Structured vs. Unstructured Representation.** Table IV substantiates the superiority of *Structured Aspects* over *Unstructured Descriptions* (free-form summaries). While free-form summaries capture broader context, they are prone to introducing generic noise and hallucinations, particularly when reviewer histories are sparse. In contrast, structured profiling acts as a semantic filter. By enforcing a strict schema, the model is constrained to discard irrelevant information and extract precise technical keywords. This "denoising" effect explains the consistent performance gains across all datasets, validating structured profiling as a more robust representation for expertise matching. **Aspect Granularity.** We further analyze the contribution of different expertise dimensions. Table V reports the peak performance for aspect subsets of size j. A clear trend emerges: single-aspect matching is insufficient for accurate recommendation. The complete triad—*Topics*, *Methodologies*, and *Applications*—consistently yields the best performance. These results confirm that expertise matching requires coverage across multiple dimensions. *Topics* align the broad subject, *Methodologies* verify technical fit, and *Applications* define the target setting. These signals are additive. Dropping any single dimension removes critical context, causing the model to miss subtle mismatches.

TABLE VI
IMPACT OF DECISION STRATEGY. VALUES REPRESENT THE AVERAGE OF SOFT AND HARD P@N METRICS.

Strategy	NeurIPS	SIGIR	SciRepEval
Direct Ranking	42.58	31.59	46.27
Rubric Scoring (Ours)	**46.40**	**32.37**	**46.76**

TABLE VII
IMPACT OF COMMITTEE SIZE. VALUES REPRESENT THE AVERAGE OF SOFT AND HARD P@N METRICS. BEST RESULTS ARE IN **BOLD**

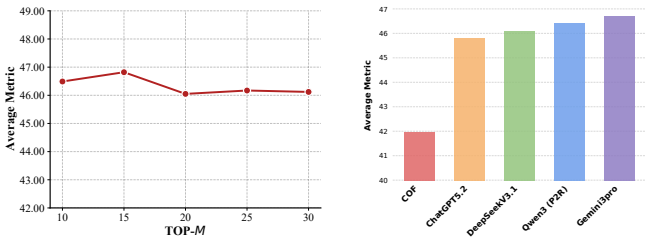
Configuration	NeurIPS	SIGIR	SciRepEval
Best Subset ($C = 1$)	45.80	32.12	46.54
Best Subset ($C = 2$)	<u>46.10</u>	32.10	46.47
Best Subset ($C = 3$)	45.81	<u>32.13</u>	<u>46.57</u>
P2R (Full, $C = 4$)	**46.40**	**32.37**	**46.76**

E. Committee Configuration Analysis

How to configure the committee critically impacts evaluation accuracy. Here, we examine these two fundamental design choices. **Ranking vs. Scoring.** Table VI shows that Rubric Scoring outperforms Direct Ranking across all datasets. This difference is primarily due to the cognitive load imposed by each method. Direct Ranking requires comparing multiple candidates at once, often leading to vague or imprecise judgments. In contrast, Rubric Scoring breaks the decision into individual checks, prompting the model to assess each candidate against specific criteria (e.g., "Does the methodology match?"). This approach ensures that decisions are based on clear, evidence-backed evaluation, reducing ambiguity and aligning better with human review standards. **Committee Composition.** Table VII confirms the necessity of a diverse committee. The full configuration consistently surpasses smaller subsets, indicating that distinct roles capture more sufficient signals for robust evaluation. For instance, a "Theorist" expert identifies mathematical mismatches that a generalist might overlook, while an "Application" expert ensures domain fit. A single perspective leaves information gaps, whereas the full committee provides a more complete and robust evaluation.

F. Model Robustness

We further examine the stability of P2R from two perspectives: sensitivity to the candidate pool size and robustness to different LLM backbones. After the hybrid retrieval stage, we retain the Top-M reviewers as the candidate set $\mathcal{C}_p$ for the fine-grained committee evaluation. As shown in Fig. 3a, varying M leads to only minor fluctuations in the overall performance on SciRepEval, indicating that P2R is not sensitive to the choice of M. To assess the generalization capability of P2R, we conduct studies using various LLM backbones. We instantiate P2R with diverse backbones, including ChatGPT5.2, DeepSeekV3.1, Qwen3, and Gemini3Pro, and evaluate them on the NeurIPS dataset (Fig. 3b). We benchmark these variants against Chain-of-Factors (CoF) [8], the state-of-the-art baseline. Regardless of the underlying LLM backbone, P2R



(a) (b)

Fig. 3. (a) Top-M sensitivity on SciRepEval. (b) LLM-backbone robustness on NeurIPS. Values represent the average of soft and hard P@N metrics.

consistently surpasses CoF, a scientific document representation model fine-tuned for reviewer matching. Notably, various general-purpose models achieve strong results, confirming the framework's robustness. These results confirm that our performance gains derive from the structured profiling mechanism rather than model-specific capabilities.

V. Conclusion

In this paper, we proposed P2R, a training-free framework that addresses the lack of explicit and multi-dimensional modeling of reviewers in prevalent approaches. P2R summarizes structured profiles spanning Topics, Methodologies, and Applications for both submissions and reviewers. It employs an LLM-based, rubric-driven committee to emulate the multi-dimensional assessment of human Area Chairs. Experiments on NeurIPS, SIGIR, and SciRepEval verify the effectiveness of our framework. P2R consistently surpasses state-of-the-art baselines across these datasets. Comprehensive ablation studies validate that each module within our framework is effective and indispensable. Finally, this work not only provides an effective solution to the reviewer matching challenge but also offers valuable empirical insights for the application of LLMs in this specific domain.

References

[1] S. Price and P. A. Flach, "Computational support for academic peer review: a perspective from artificial intelligence," *Communications of the ACM*, vol. 60, no. 3, pp. 70–79, 2017.

[2] J. Wang, Y. Liu, H. Xu, K. Hu, S. Di, W. Ni, L. Yue, M.-L. Zhang, K. Ren, and L. Chen, "When AI reviews science: Can we trust the referee?" *The Innovation Informatics*, vol. 2, no. 1, p. 100030, 2026.

[3] X. Liu, T. Suel, and N. Memon, "A robust model for paper reviewer assignment," in *Proceedings of the 8th ACM Conference on Recommender systems*, 2014, pp. 25–32.

[4] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Sixth international conference on data mining (ICDM'06)*. IEEE, 2006, pp. 613–622.

[5] I. Beltagy, K. Lo, and A. Cohan, "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)*, 2019, pp. 3615–3620.

[6] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "Specter: Document-level representation learning using citation-informed transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2270–2282.

[7] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman, "Scirepeval: A multi-format benchmark for scientific document representations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5548–5566.

[8] Y. Zhang, Y. Shen, S. Kang, X. Chen, B. Jin, and J. Han, "Chain-of-factors paper-reviewer matching," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1901–1910.

[9] L. Charlin, R. S. Zemel, and C. Boutilier, "A framework for optimizing paper matching." in *UAI*, vol. 11, 2011, pp. 86–95.

[10] A. Schrijver *et al.*, *Combinatorial optimization: polyhedra and efficiency*. Springer, 2003, vol. 24, no. 2.

[11] M. Saveski, S. Jecmen, N. Shah, and J. Ugander, "Counterfactual evaluation of peer-review assignment policies," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 765–58 786, 2023.

[12] L. Charlin and R. S. Zemel, "The toronto paper matching system: An automated paper-reviewer assignment system," in *ICML 2013 Workshop on Peer Reviewing and Publishing Models*, 2013.

[13] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 500–509.

[14] W. Tang, J. Tang, and C. Tan, "Expertise matching via constraint-based optimization," in *2010 IEEE/WIC/aCM international conference on web intelligence and intelligent agent technology*, vol. 1. IEEE, 2010, pp. 34–41.

[15] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm, "Neighborhood contrastive learning for scientific document representations with citation embeddings," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11 670–11 688.

[16] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao, "Large language models are zero-shot rankers for recommender systems," in *European Conference on Information Retrieval*. Springer, 2024, pp. 364–381.

[17] V. M. Karan, S. McQuistin, R. Yanagida, C. Perkins, G. Tyson, I. Castro, P. Healey, and M. Purver, "A dataset for expert reviewer recommendation with large language models as zero-shot rankers," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 11 422–11 427.

[18] Q. Peng, C. Wang, Y. Wang, H. Liu, X. Guo, and W. Wang, "Frontier-revrec: A large-scale dataset for reviewer recommendation," *arXiv preprint arXiv:2510.16597*, 2025.

[19] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, and Z. Ren, "Is chatgpt good at search? investigating large language models as re-ranking agents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14 918–14 937.

[20] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on llm-as-a-judge," *The Innovation*, p. 101253, 2026.

[21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46 595–46 623, 2023.

[22] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[23] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "Nv-embed: Improved techniques for training llms as generalist embedding models," in *International Conference on Learning Representations*, 2025.

[24] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758–759.

[25] M. Karimzadehgan, C. Zhai, and G. Belford, "Multi-aspect expertise matching for review assignment," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 1113–1122.

[26] Y. Yu, C. Xiong, S. Sun, C. Zhang, and A. Overwijk, "Coco-dr: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 1462–1479.