

MERIT: Matching Expertise via Rubric-Informed Training for Reviewer Assignment

Zixuan Yang Yibo Zhao Weicong Liu Xiang Li*

School of Data Science and Engineering, East China Normal University

Abstract

Matching submissions with suitable reviewers at scale is a growing challenge for major venues, yet existing approaches either rely on coarse proxy signals that conflate general relatedness with true suitability, or require expensive human annotations that are difficult to scale for training. We propose MERIT, a two-stage framework that bridges this gap by converting criterion-level expertise matching into scalable suitability supervision. In the first stage, we train a reviewer assessor via reinforcement learning to identify the expertise dimensions a paper requires, match them against the reviewer’s prior work, and produce a suitability decision, with rewards provided by an LLM judge guided by paper-specific expertise rubrics. In the second stage, we distill the assessor’s predictions into an embedding-based retriever for efficient large-scale assignment. Experiments show that our 4B reviewer assessor outperforms larger general-purpose LLMs on suitability classification, and the resulting retriever achieves state-of-the-art performance across LR-Bench and the CMU Gold dataset. Our code is available at <https://github.com/Luli3220/MERIT>.

1 Introduction

The quality of peer review hinges on matching each submission with reviewers who have the right expertise (Fu et al., 2025; Long et al., 2013). As submission volumes grow—major artificial intelligence venues now receive over ten thousand papers per cycle (Gupta and Sarkar, 2025; Kim et al., 2025)—manually identifying suitable reviewers has become infeasible, making automatic reviewer assignment an increasingly critical problem.

Most existing methods estimate paper-reviewer affinity from scalable but coarse proxy signals: textual similarity (Charlin and Zemel, 2013; Zhang et al., 2020), topical overlap (Anjum et al., 2019;

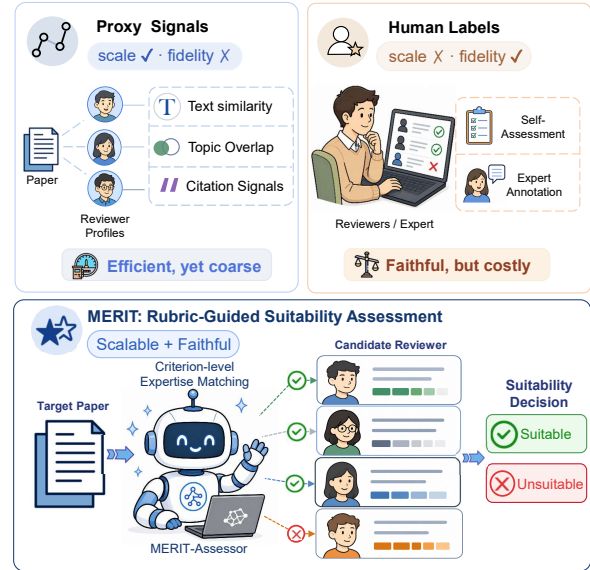


Figure 1: Existing reviewer assignment approaches rely on either scalable but coarse proxy signals (left) or faithful but costly human labels (right). MERIT (bottom) trains a reviewer assessor with paper-specific expertise rubrics for criterion-level suitability matching.

Jin et al., 2017), citation relations (Karimzadehgan et al., 2008; Pradhan et al., 2021), or combinations thereof (Zhang et al., 2025b). These signals collapse all overlap between a reviewer’s prior work and the target paper into a single relevance score, treating every shared element as equally important regardless of its role in the paper. For instance, consider a paper whose core contribution is a novel policy-optimization algorithm evaluated on instruction-following benchmarks. A benchmark specialist may overlap with the paper’s evaluation setting, while an optimization expert may overlap with its algorithmic contribution; an aggregate affinity score may fail to distinguish these roles, even though the latter expertise is more central to evaluating the paper’s contribution. The root issue is that a reviewer is truly *suitable* only when their expertise covers the dimensions most critical

*Corresponding Author: xiangli@dase.ecnu.edu.cn

to the paper, yet proxy signals provide no paper-specific mechanism for distinguishing critical expertise from incidental overlap.

An alternative is to use higher-fidelity signals such as reviewer self-assessments (Stelmakh et al., 2025; Liu et al., 2026) or expert annotations (Zhang et al., 2025b; Mimno and McCallum, 2007). These signals more directly capture whether a reviewer is suitable for a given paper, but they are expensive to collect and difficult to scale. This creates a gap: *scalable signals lack the granularity to capture suitability, while high-fidelity signals cannot be obtained at the scale required for training.*

Recent work has explored using large language models (LLMs) to assess reviewer suitability directly, but general-purpose LLMs remain unreliable for this task (Stelmakh et al., 2025; Liu et al., 2026), tending to produce broad judgments of topical relevance rather than evaluating expertise along individual dimensions. This may stem from the absence of an explicit specification of what expertise a paper demands: each paper requires expertise along a distinct set of dimensions, some critical to evaluating its core contribution and others secondary. Without such a specification, LLMs have no basis for distinguishing central expertise from incidental overlap. We address this by decomposing each paper’s expertise requirements into a set of weighted criteria—what we call a *paper-specific expertise rubric*. Each criterion specifies one expertise dimension and is assigned an importance weight reflecting how critical it is to the paper.

Building on this formulation, we propose MERIT, a two-stage framework. In the first stage, we train a reviewer assessor with reinforcement learning. Given a paper and a reviewer’s publication history, the model identifies the expertise dimensions the paper requires, evaluates whether the reviewer’s publications demonstrate expertise in each dimension, and produces a suitability decision. The reward signal is provided by an LLM judge that evaluates each model output against an automatically generated expertise rubric for the target paper, along two axes: whether the output addresses each criterion with grounded evidence (*rubric coverage*), and whether the final decision is consistent with the evidence (*decision consistency*).

Applying the reviewer assessor to every candidate reviewer is prohibitively costly at conference scale. In the second stage, we therefore use the trained assessor to annotate a large set of candidate pairs and distill the supervision into an embedding-

based retriever for efficient large-scale assignment.

In summary, our contributions are as follows:

- We formulate reviewer assignment as criterion-level expertise matching and introduce paper-specific expertise rubrics as a supervision signal that combines the scalability of proxy signals with the granularity of expert annotations.
- We propose MERIT, a two-stage framework: (i) training a reviewer assessor with rubric-guided reinforcement learning, and (ii) distilling its predictions into an embedding-based retriever for efficient large-scale assignment.
- Experiments show that our 4B reviewer assessor outperforms larger general-purpose LLMs on reviewer assessment, and that the resulting retriever achieves state-of-the-art performance across LR-Bench and the CMU Gold dataset.

2 Related work

Automatic Reviewer Assignment. The quality of reviewer assignment depends largely on paper-reviewer affinity scores (Aksoy et al., 2023; Wang et al., 2010), and prior work has explored a range of signals for estimating these scores. One line of work relies on scalable proxy signals. Early methods compute affinity from TF-IDF similarity (Charlin and Zemel, 2013) or latent topic models (Conry et al., 2009; Dumais and Nielsen, 1992) between papers and reviewer publications. More recent methods use citation-informed document encoders such as SPECTER (Cohan et al., 2020) and SciNCL (Ostendorff et al., 2022) to capture richer semantic relatedness, while CoF (Zhang et al., 2025b) combines semantic, topical, and citation signals into a unified scoring framework. RATE (Liu et al., 2026) uses proxy-derived scores to construct weak supervision for training a dedicated retrieval model. Despite their scalability, these methods estimate affinity from aggregate paper-reviewer overlap without distinguishing which expertise dimensions are most relevant to a given paper.

Another line of work obtains supervision from reviewer self-assessments (Stelmakh et al., 2025; Liu et al., 2026) or expert annotations (Mimno and McCallum, 2007; Zhang et al., 2025b). These signals more directly reflect suitability but demand substantial human effort, limiting their use as training data. Our work bridges these two lines by using LLMs to perform criterion-level expertise matching, producing supervision that captures fine-grained suitability while remaining scalable.

Rubric-based Rewards. Rubric-based evaluation decomposes the assessment of complex outputs into criterion-level judgments, and has been used to evaluate LLM outputs in open-ended generation (Liu et al., 2023; Kim et al., 2024), instruction following (Ye et al., 2024; Qin et al., 2024), and domain-specific tasks (Arora et al., 2025; Yang et al., 2026; Shi et al., 2026). More recently, such criterion-level judgments have been aggregated into reward signals for LLM post-training (Gunjal et al., 2025; He et al., 2025; Liu et al., 2025). In all these settings, rubrics define what constitutes a good *output*. We apply the same rubric-based framework to a different object of evaluation: our rubrics define what constitutes a suitable *reviewer*, specifying the expertise dimensions a paper requires and their relative importance.

3 Methodology

Given a target paper p and a pool of candidate reviewers $\mathcal{R}$, where each reviewer $r \in \mathcal{R}$ is represented by their publication history $H_r = \{h_1, \dots, h_{|H_r|}\}$, our goal is to identify reviewers whose expertise covers the dimensions most critical to evaluating p , efficiently and at scale. We propose MERIT, a two-stage framework for Matching Expertise via Rubric-Informed Training for reviewer assignment, as illustrated in Figure 2.

In the first stage, we train a reviewer assessor with reinforcement learning to perform criterion-level expertise matching and produce suitability decisions, with rewards provided by an LLM judge guided by automatically generated, paper-specific expertise rubrics (Section 3.2). In the second stage, we distill the trained assessor’s predictions into an embedding-based retriever for efficient large-scale assignment (Section 3.3). We refer to the trained assessor and retriever as the MERIT-Assessor and MERIT-Retriever, respectively. Both stages operate on paper-reviewer pairs constructed via a shared candidate retrieval procedure (Section 3.1).

3.1 Data Construction

Training on randomly sampled paper-reviewer pairs would yield mostly trivial negatives with little topical overlap, offering limited signal for learning fine-grained suitability distinctions. We instead retrieve candidates using an existing affinity model (RATE; Liu et al., 2026), so that candidates have relatively high topical overlap with the target paper. This makes the resulting pairs challenging: candi-

dates are topically related to the target paper but vary in whether they are truly suitable to review it.

For each target paper, we augment its title and abstract with the introduction section to provide richer context for both rubric generation and suitability assessment. We then retrieve the top-ranked reviewers as candidates, yielding paper-reviewer pairs that are split into two disjoint subsets: one for training the MERIT-Assessor (Section 3.2) and another for generating pseudo-labels for the MERIT-Retriever (Section 3.3). Detailed construction procedures, including filtering criteria and dataset statistics, are provided in Section 4.1.

3.2 Reviewer Assessor

Given a target paper p and a candidate reviewer r , the MERIT-Assessor produces a binary label $\hat{y} \in \{0, 1\}$ indicating whether r is suitable to review p . Specifically, the model generates a structured chain-of-thought with four parts: (1) *required expertise*, a list of expertise dimensions identified from the target paper; (2) *evidence clues*, specific publications from the reviewer’s history linked to each dimension; (3) *reasoning summary*, a synthesis of the evidence; and (4) *final label*, the suitability decision. This structure grounds each prediction in explicit evidence and makes the decision process transparent. Since gold suitability labels are unavailable at scale, we train the model with reinforcement learning using rewards from an LLM judge guided by paper-specific expertise rubrics. We first describe the rubric construction procedure, then detail the reward design.

Rubric construction. Each paper-specific expertise rubric specifies the expertise dimensions required to review the target paper and their relative importance. Formally, a rubric is defined as $u_p = \{(c_j, w_j)\}_{j=1}^k$, where each criterion $c_j = (t_j, d_j)$ consists of a title t_j and a description d_j specifying one expertise dimension, and w_j denotes its importance weight. Criteria are divided into two tiers: *Core* criteria ($w_j = 5$) represent expertise essential for evaluating the paper’s central contribution, while *Secondary* criteria ($w_j \in \{3, 4\}$) capture supporting knowledge. Each rubric contains 3–6 criteria with 1–2 Core entries, as a small number of well-differentiated dimensions yields more reliable LLM-judge evaluation than a long list of overlapping items (Gunjal et al., 2025). Rubrics are generated by prompting an LLM with the paper’s title, abstract, and introduction, along with

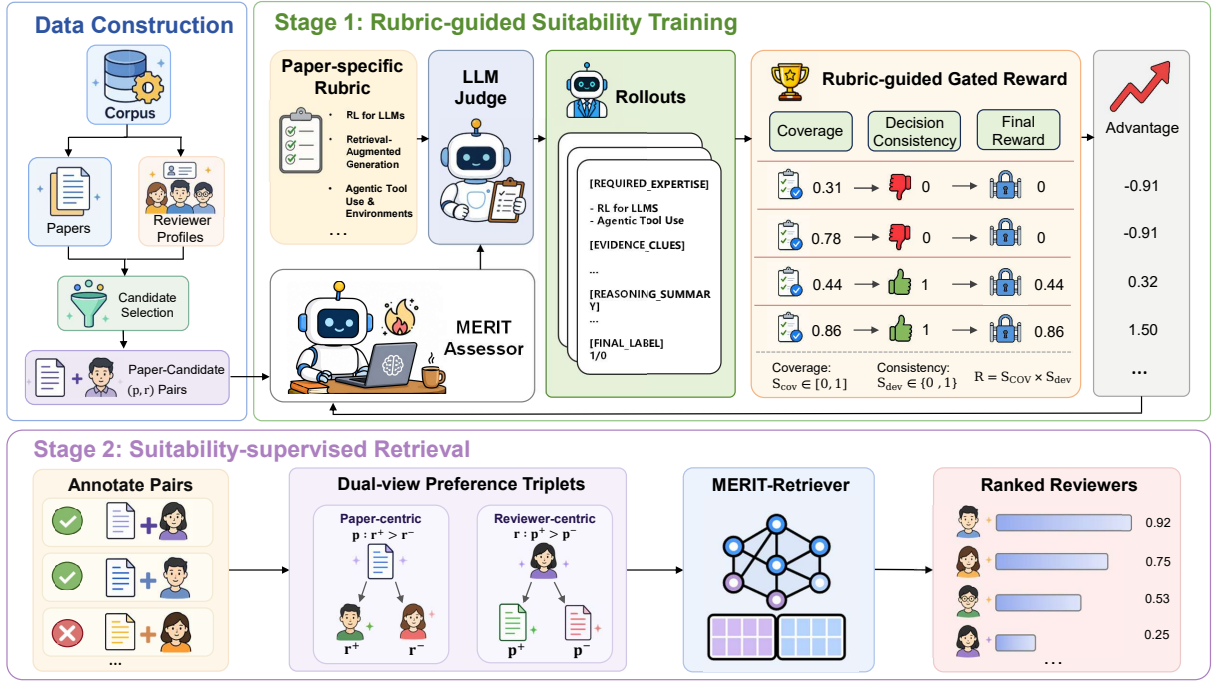


Figure 2: Overview of the full pipeline of our proposed method.

human-written exemplars. Prompts and a rubric example are provided in Appendices C.1 and D.

Reward design. The MERIT-Assessor generates a structured chain-of-thought before producing a final label. Two failure modes can arise in this process: the analysis may fail to identify the expertise dimensions critical to the paper, or may assert relevance without citing specific publications as evidence; and the final label may contradict the analysis itself. The reward signal is designed to penalize both. We decompose it into two components—*rubric coverage*, which measures how thoroughly the analysis addresses the rubric criteria with evidence from the reviewer’s publications, and *decision consistency*, which measures whether the final label is consistent with the evidence presented in the analysis—and combine them through a gating mechanism so that the model is rewarded only when its decision is supported by a thorough, evidence-grounded assessment. Both components are scored by the LLM judge in a single pass over the model output and rubric.

Rubric coverage. For each rubric entry $(c_j, w_j) \in u_p$, the LLM judge assigns a binary indicator $m_j \in \{0, 1\}$, setting $m_j = 1$ if the output addresses the expertise dimension specified by c_j with grounded evidence—either by citing relevant publications from the reviewer’s history or by explicitly noting the absence of such evidence. The

overall coverage score is the importance-weighted fraction of covered criteria:

$$s_{cov} = \frac{\sum_{j=1}^k w_j m_j}{\sum_{j=1}^k w_j}. \quad (1)$$

Decision consistency. A positive prediction ($\hat{y} = 1$) requires evidence that the reviewer covers at least one Core criterion and at least one additional criterion, ensuring that the decision reflects both central expertise and sufficient breadth; a negative prediction ($\hat{y} = 0$) is consistent when this condition is not met. The LLM judge assigns $s_{dec} \in \{0, 1\}$, with $s_{dec} = 1$ if $\hat{y}$ is consistent with this rule and $s_{dec} = 0$ otherwise.

Gated reward. The final reward gates coverage on consistency:

$$R = s_{dec} \cdot s_{cov}. \quad (2)$$

When $s_{dec} = 0$, the reward is zero regardless of coverage, ensuring that no credit is given to outputs whose final label contradicts their own analysis. The full prompt templates for both the policy model and the LLM judge are provided in Appendix C.1.

RL training. We optimize the MERIT-Assessor with GRPO (Shao et al., 2024). For each input prompt x^i , the LLM policy π_θ samples a group of G outputs $\{o_j^i\}_{j=1}^G$, and each output receives a reward $r_j^i = R(o_j^i)$. We compute the group-wise

advantage as:

$$\hat{A}_j^i = \frac{r_j^i - \text{mean}(\{r_l^i\}_{l=1}^G)}{\text{std}(\{r_l^i\}_{l=1}^G)}. \quad (3)$$

The same advantage $\hat{A}_j^i$ is applied to every token in output o_j^i . Following DAPO (Yu et al., 2026), we adopt the clipped surrogate objective with the clip-higher strategy to ensure update stability while encouraging exploration. Specifically, for each token $o_{j,t}^i$, we define the policy ratio as:

$$\rho_{j,t}^i = \frac{\pi_\theta(o_{j,t}^i | x^i, o_{j,<t}^i)}{\pi_{\theta_{\text{old}}}(o_{j,t}^i | x^i, o_{j,<t}^i)}. \quad (4)$$

The token-level objective is then defined as:

$$l_{j,t}^i = \min\left(\rho_{j,t}^i \hat{A}_j^i, \text{clip}(\rho_{j,t}^i, 1 - \epsilon_l, 1 + \epsilon_h) \hat{A}_j^i\right). \quad (5)$$

The final RL objective averages the token-level surrogate over generated tokens and adds a KL regularization term toward the reference policy:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\{o_j^i\}_{j=1}^G \sim \pi_{\theta_{\text{old}}}} \left[\frac{\sum_{j=1}^G \sum_{t=1}^{|o_j^i|} l_{j,t}^i}{\sum_{j=1}^G |o_j^i|} \right] \\ - \beta \mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \end{aligned} \quad (6)$$

Here, $\mathcal{D}_{\text{KL}}$ is computed using the low-variance KL estimator (k3) between π_θ and π_{ref} .

3.3 MERIT-Retriever

Applying the MERIT-Assessor to every candidate is prohibitively expensive at conference scale. We therefore distill its predictions into an embedding-based retriever. Using a separate subset of candidate pairs (constructed as in Section 3.1), we annotate each pair with the trained assessor to obtain pseudo-labeled tuples $(p, r, \hat{y})$, which are then converted into preference triplets for training.

Preference Data Construction Each training instance is a preference triplet (a, c^+, c^-) , where a is an anchor, c^+ a suitable candidate ($\hat{y} = 1$), and c^- an unsuitable candidate ($\hat{y} = 0$). For each anchor, every positive candidate is paired with every negative candidate. We construct triplets under two complementary views: (1) in the *paper-centric* view, the anchor is a target paper and the candidates are reviewers; (2) in the *reviewer-centric* view, the anchor is a reviewer and the candidates are papers. Both views share the same triplet format and are trained jointly for more robust representations (Liu et al., 2026; Li et al., 2021).

Paper-conditioned reviewer profiles. Training on the preference triplets above requires encoding both papers and reviewers into a shared embedding space. Since the MERIT-Assessor evaluates whether a reviewer’s publications collectively cover multiple expertise dimensions, the retriever’s reviewer representation should capture the same cross-publication patterns. However, prior methods (Zhang et al., 2025b; Stelmakh et al., 2025) score each publication against the target paper independently and aggregate the results (e.g., via percentile pooling), discarding inter-publication dependencies. We instead construct a paper-conditioned reviewer profile by selecting the K publications from H_r most relevant to p , ranked by cosine similarity under the backbone embedding model f_0 (before fine-tuning):

$$s_p(h_m) = \cos(f_0(p), f_0(h_m)). \quad (7)$$

Restricting to the top K publications keeps the input compact while filtering out unrelated work. The selected publications are concatenated with a task-specific instruction prefix (e.g., “For the Author (Query): Represent this author’s publication history and expertise for finding relevant academic papers.”) and encoded as a single sequence to produce the reviewer representation. We use the frozen f_0 for publication selection throughout both training and evaluation, ensuring consistency independent of the retriever’s learned parameters.

Dual-view Preference Alignment We optimize an embedding model f_θ , initialized from f_0 , with a multi-objective loss over the preference triplets from both views. The matching score between an anchor a and a candidate c is defined as:

$$s(a, c) = \cos(f_\theta(a), f_\theta(c)). \quad (8)$$

The pairwise ranking loss encourages the model to score the suitable candidate higher than the unsuitable one:

$$\mathcal{L}_{\text{pair}} = -\log \sigma\left(\frac{s(a, c^+) - s(a, c^-)}{\tau}\right), \quad (9)$$

where τ is a temperature hyperparameter. The in-batch contrastive loss provides additional negative signal by contrasting the positive candidate against all candidates in the mini-batch $\mathcal{B}$:

$$\mathcal{L}_{\text{ncc}} = -\log \frac{\exp(s(a, c^+)/\tau)}{\sum_{c \in \mathcal{B}} \exp(s(a, c)/\tau)}. \quad (10)$$

Table 1: Reviewer suitability classification results on LR-Bench. Best and second-best results are marked in bold and underlined, respectively.

Model	LR-Bench				
	Acc. (↑)	B.Acc. (↑)	P. (↑)	R. (↑)	F1 (↑)
Direct prompting					
Qwen3.5-Plus	59.18%	64.41%	50.52%	95.67%	66.12%
DeepSeek-V3.2	63.11%	66.93%	53.40%	<u>89.76%</u>	66.96%
Expertise-aware prompting					
Qwen3-4B	67.08%	68.81%	57.61%	78.45%	66.15%
Qwen3.5-Plus	67.87%	70.39%	57.71%	85.43%	<u>68.89%</u>
DeepSeek-V3.2	<u>69.51%</u>	<u>71.00%</u>	<u>60.06%</u>	79.92%	68.58%
MERIT-Assessor-4B	71.64%	73.11%	62.09%	81.89%	70.63%

The final objective combines both terms:

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \lambda_{\text{nce}} \mathcal{L}_{\text{nce}}. \quad (11)$$

We fine-tune f_θ with LoRA (Hu et al., 2022) for parameter-efficient adaptation while preserving the pretrained representations. Training hyperparameters are provided in Appendix B.2.

4 Experiments

Our experiments address three questions. (Q1) Does rubric-guided reward training improve suitability classification over general-purpose LLM prompting? (Section 4.4.1) (Q2) Does distilling the MERIT-Assessor’s predictions into an embedding-based retriever improve reviewer retrieval? (Section 4.4.2) (Q3) How does the quality of the pseudo-label source affect downstream retrieval? (Section 4.5) We additionally report sensitivity to reviewer profile size in Appendix A.1, cost analysis in Appendix A.2 and human validation of the generated rubrics and LLM judge in Appendix A.3.

4.1 Experimental Setup

Training data. All training data are drawn from the RATE corpus (Liu et al., 2026), with papers appearing in any evaluation benchmark removed to prevent data leakage. We use two disjoint splits for the two stages of our framework, constructed by a shared procedure: we recover each paper’s introduction from ar5iv (Ginev, 2024) and discard papers for which no introduction is available; we restrict the reviewer pool to authors with at least three publications and exclude authors whose publication history contains the target paper itself; and for each target paper, we rank all eligible reviewers with the RATE retrieval model and retain the top 5 as candidates.

For training the MERIT-Assessor (Section 3.2), we randomly sample 2,000 papers, of which 1,497 have available introductions. We generate one expertise rubric per paper using Qwen3-Max-Thinking, yielding 7,485 rubric-equipped training tuples (p, r, u_p) . For training the MERIT-Retriever (Section 3.3), we sample a separate set of 6,000 papers (5,374 with available introductions) and construct 26,870 candidate paper-reviewer pairs using the same procedure. These pairs are annotated by the trained suitability model and converted into dual-view preference triplets following Section 3.3, yielding 8,720 paper-centric and 7,208 reviewer-centric triplets.

Training details. For the MERIT-Assessor, we initialize the policy model from Qwen3-4B (Team, 2025) and optimize with GRPO using the verl framework (Sheng et al., 2024), with DeepSeek-V3.2-Thinking (DeepSeek-AI et al., 2025) as the LLM judge. For the MERIT-Retriever, we initialize from Qwen3-Embedding-8B (Zhang et al., 2025a) and fine-tune with LoRA (Hu et al., 2022). The same embedding backbone is used to construct paper-conditioned reviewer profiles during both training and evaluation, following Section 3.3. Detailed experimental settings are provided in Appendix B.

Evaluation benchmarks. For suitability classification, we use LR-Bench (Liu et al., 2026), where each reviewer-paper pair carries a self-assessed expertise rating on a 1–5 scale. We binarize these ratings by treating scores of 4 and 5 as positive and the remaining scores as negative. To ensure consistency with the training input, we also recover each target paper’s introduction and discard examples where it is unavailable, yielding 810 labeled pairs

Table 2: Reviewer retrieval results on LR-Bench and CMU Gold. We report expertise-aligned loss and pairwise accuracy. Best and second-best results are marked in bold and underlined, respectively.

Algorithm	Loss ($\downarrow$)				Acc. ($\uparrow$)			
	LR-PC	LR-RC	Gold	Avg.	LR-PC	LR-RC	Gold	Avg.
Statistical-based Methods								
TPMS	0.2920	0.2322	0.2811	0.2684	66.22%	72.01%	71.89%	70.04%
LLM-based Methods								
DeepSeek-V3.2	0.2736	0.2348	0.2237	0.2440	50.34%	54.04%	77.36%	60.58%
Qwen3-Max	0.2698	0.2289	0.2246	0.2411	47.65%	55.01%	77.54%	60.07%
Embedding-based Methods								
BERTScore	0.2691	0.3380	0.3414	0.3162	66.22%	62.36%	65.86%	64.81%
CoF	0.2996	0.2175	0.2564	0.2578	64.19%	73.34%	74.36%	70.63%
SPECTER	0.2118	0.2396	0.2851	0.2455	<u>72.97%</u>	71.29%	71.49%	71.92%
SPECTER2 PRX	0.1966	0.2175	0.2436	0.2192	<u>72.97%</u>	72.26%	75.64%	73.62%
SciNCL	0.2042	0.1938	0.2663	0.2214	72.30%	75.03%	73.37%	73.57%
RATE-8B	<u>0.1851</u>	<u>0.1857</u>	<u>0.1991</u>	<u>0.1900</u>	75.00%	<u>75.63%</u>	<u>80.09%</u>	<u>76.91%</u>
MERIT-Retriever-8B	0.1698	0.1783	0.1842	0.1774	75.00%	76.84%	81.58%	77.81%

split into 200 for validation and 610 for testing.

For reviewer retrieval, we evaluate on LR-Bench and the CMU gold standard dataset (Stelmakh et al., 2025). Both datasets provide sparse expertise ratings based on human annotation; we derive pairwise preferences from these ratings following their original evaluation protocols. LR-Bench yields both paper-centric (PC) and reviewer-centric (RC) pairs, while CMU yields only reviewer-centric pairs. We reserve 30% of LR-Bench preference pairs for validation and use the remaining 70% for testing. CMU is used entirely for testing.

4.2 Metrics

For suitability classification, we report accuracy (Acc.), balanced accuracy (B.Acc.), precision (P.), recall (R.), and F1 score.

For reviewer retrieval, following Stelmakh et al. (2025) and Liu et al. (2026), we adopt a normalized ranking loss $\mathcal{L} \in [0, 1]$ as our primary metric. Given a set of preference pairs $\mathcal{P}$, where each pair (x, y) shares the same anchor and satisfies $\epsilon_x > \epsilon_y$ under ground-truth expertise labels, the loss penalizes misordered pairs in proportion to the gap between the two expertise ratings:

$$\mathcal{L} = \frac{\sum_{(x,y) \in \mathcal{P}} \mathcal{I}(s_x < s_y) \cdot |\epsilon_x - \epsilon_y|}{\sum_{(x,y) \in \mathcal{P}} |\epsilon_x - \epsilon_y|}, \quad (12)$$

where $\mathcal{I}$ is the indicator function, and s denotes the model-predicted score, and the loss is normalized to $[0, 1]$ by the sum of all pairwise expertise gaps.

In addition, we report pairwise accuracy, the fraction of pairs where the model correctly orders the two candidates by ground-truth rating.

4.3 Baselines

Suitability classification. We consider two categories of LLM-based baselines. **(1) Direct prompting** predicts reviewer suitability from the target paper and reviewer profile without explicit expertise decomposition. **(2) Expertise-aware prompting** follows the same structured assessment process as our policy model—expertise identification, evidence matching, and final prediction—but without RL training. We evaluate direct prompting with Qwen3.5-Plus and DeepSeek-V3.2, and expertise-aware prompting with Qwen3-4B, Qwen3.5-Plus, and DeepSeek-V3.2. All baselines use greedy decoding with thinking mode enabled.

Reviewer retrieval. We compare against three categories of baselines. **(1) Statistical method.** We include TPMS (Charlin and Zemel, 2013), a standard reviewer-assignment baseline based on TF-IDF similarity between the target paper and the reviewer’s publications. **(2) Embedding-based models.** We compare with BERTScore (Zhang* et al., 2020), a common text similarity metric based on contextual token embeddings. We also include SPECTER (Cohan et al., 2020), SciNCL (Osten-dorff et al., 2022), and SPECTER2 with the PRX adapter (Singh et al., 2023), all of which are scientific document encoders trained on citation links

between papers. In addition, we compare with CoF (Zhang et al., 2025b), a factor-aware reviewer-assignment framework that combines semantic, topical, and citation signals; and RATE-8B (Liu et al., 2026), a weakly supervised reviewer retriever that combines high-confidence proxy signals with keyword-based reviewer profiling. (3) **LLM-based methods.** We prompt DeepSeek-V3.2 and Qwen3-Max to score paper-reviewer affinity in a zero-shot setting, and use the resulting scores for pairwise ranking. Additional implementation details and LLM-baseline prompts are provided in Appendices B.3 and C.2.

4.4 Main Results

4.4.1 Reviewer suitability classification (Q1)

As shown in Table 1, the MERIT-Assessor achieves the best accuracy, balanced accuracy, precision, and F1 score, despite using a smaller 4B backbone than several larger prompting baselines. The most direct comparison is with Qwen3-4B under the same expertise-aware format but without training: our model improves Acc. from 67.08% to 71.64% and F1 from 66.15% to 70.63%, confirming that rubric-guided reinforcement learning teaches the model to make more discriminative judgments beyond what the structured format alone provides.

The prompting results also reveal a systematic positive prediction bias in direct LLM judgments. Without structured expertise decomposition, models tend to predict nearly all candidates as suitable — Qwen3.5-Plus reaches 95.67% recall but only 50.52% precision. Expertise-aware prompting mitigates this bias by requiring the model to identify expertise dimensions and match them against the reviewer’s publications before deciding. This decomposition alone, without any training, lifts Qwen3-4B above larger direct-prompting models in both accuracy and precision, indicating that the structured assessment format provides a useful inductive bias. These results establish the trained assessor as a stronger pseudo-label source; we next examine whether this improvement translates into better downstream retrieval.

4.4.2 Reviewer retrieval (Q2)

Table 2 shows that the MERIT-Retriever achieves the best or tied-best performance on every metric, reducing average expertise-aligned loss from 0.1900 to 0.1774 and raising average pairwise accuracy from 76.91% to 77.81% compared with RATE-8B, the strongest baseline. Since both mod-

els share the same embedding backbone and retriever architecture, this improvement suggests that higher-fidelity criterion-level suitability supervision provides more effective training signals than RATE’s proxy-derived weak supervision.

Among embedding-based baselines, citation-informed encoders (SPECTER, SPECTER2-PRX, SciNCL) consistently outperform BERTScore, confirming that citation-aware pretraining provides a stronger foundation for this task. RATE-8B further improves over these encoders through dedicated training with high-confidence proxy-derived weak supervision and keyword-based reviewer profiling, but its reliance on proxy signals limits the fidelity of its supervision. Notably, TPMS remains competitive despite relying solely on TF-IDF similarity without any learned representations, achieving 70.04% average accuracy—higher than both zero-shot LLM methods and close to CoF, which integrates semantic, topical, and citation signals.

Zero-shot LLM scoring exhibits a distinctive pattern: DeepSeek-V3.2 and Qwen3-Max achieve relatively low expertise-aligned loss (0.2440 and 0.2411) but low pairwise accuracy (60.58% and 60.07%). This gap indicates that while LLMs can identify clearly unsuitable reviewers, they struggle with fine-grained pairwise ordering among topically related candidates—precisely the regime where criterion-level matching matters most. Their accuracy also varies sharply across benchmarks, exceeding 77% on CMU Gold but falling below 55% on LR-Bench, suggesting sensitivity to the distribution of candidate difficulty.

4.5 Effect of Pseudo-label Quality (Q3)

Since Stage-2 retrieval relies on pseudo-labels generated by the Stage-1 annotator, we study how pseudo-label quality affects downstream retrieval performance. To examine this, we vary the Stage-1 annotator while keeping the retrieval training pipeline and candidate pairs unchanged. We compare three pseudo-label sources: (1) the base Qwen3-4B model with expertise-aware prompting but without RL training, (2) DeepSeek-V3.2, the strongest general-purpose prompting baseline (Table 1), and (3) the trained MERIT-Assessor. All variants are evaluated on LR-Bench.

As shown in Table 3, downstream retrieval performance improves monotonically as the annotator becomes stronger. Although different annotators produce different label distributions and therefore different numbers of preference triplets, Qwen3-4B

Table 3: Effect of the Stage-1 pseudo-label source on retrieval performance (LR-Bench). All variants annotate the same candidate paper-reviewer pairs and use the same retriever training procedure. Avg. denotes the macro-average over PC and RC.

Pseudo-label Source	#Triplets (PC, RC)	Loss ($\downarrow$)			Acc. ($\uparrow$)		
		PC	RC	Avg.	PC	RC	Avg.
Qwen3-4B	(12077, 9259)	0.2004	0.1879	0.1942	72.30%	74.79%	73.55%
DeepSeek-V3.2	(7037, 6112)	<u>0.1813</u>	<u>0.1805</u>	<u>0.1809</u>	<u>74.32%</u>	<u>75.87%</u>	<u>75.10%</u>
MERIT-Assessor-4B	(8720, 7208)	0.1698	0.1783	0.1741	75.00%	76.84%	75.92%

generates the most triplets yet yields the weakest retriever, suggesting that retrieval performance is more closely associated with pseudo-label quality than quantity. Overall, the consistent trend indicates that higher-fidelity suitability predictions provide more effective supervision for the retriever.

5 Conclusion

We identify a key challenge in automatic reviewer assignment: scalable supervision signals lack fidelity to true reviewer suitability, while high-fidelity signals are difficult to scale. To address this, we proposed MERIT, a two-stage framework that trains a reviewer assessor with paper-specific expertise rubrics and distills its predictions into an efficient embedding-based retriever for large-scale assignment. Experiments showed that the trained 4B assessor produces more discriminative suitability judgments than larger general-purpose LLMs, and that distilling its predictions yields consistent retrieval gains across two benchmarks. These results demonstrate that rubric-guided, criterion-level supervision can better balance fidelity and scalability in automatic reviewer assignment.

6 Limitations

Despite the promising results, our work has several limitations. First, the construction of the training set depends on RATE’s recall for candidate mining, which may bias the learned suitability boundaries when RATE fails to retrieve relevant candidates.

Second, our reviewer profiles treat all publications equally without modeling author-order signals such as first- or last-author versus middle-author contributions. In fields where author position is a proxy for contribution depth, this may lead the model to overestimate reviewer expertise based on publications with peripheral involvement.

Third, our evaluation focuses on expertise matching in isolation and does not account for practical constraints in real-world reviewer assignment, such

as reviewer availability, workload balance, conflicts of interest, and expertise scarcity.

Fourth, all experiments are conducted on computer science venues, and the rubric structure—a small number of weighted, tiered criteria—reflects the expertise patterns typical of this domain. Whether this structure generalizes to fields with different reviewing norms remains to be validated.

Ethical Consideration

Potential Risks. MERIT is designed to assist reviewer assignment in academic peer review. Although the system aims to improve matching quality, we recognize several potential risks. First, the suitability model and rubric generator may inherit biases from their underlying LLMs, potentially favoring certain research paradigms or methodologies over others. Second, we emphasize that MERIT provides ranked suggestions, not final decisions, and is intended to assist rather than replace human judgment in reviewer assignment. Third, aggregating a researcher’s publication history into a structured expertise profile constitutes an automated capability assessment. All publications used are publicly available and drawn from existing research datasets released for academic use, and no private or personally identifiable information beyond public authorship metadata is involved.

Licenses, Intended Use, and Sensitive Information.

All data used in this study are derived from publicly available academic resources. Our training data are drawn from the RATE corpus, which is constructed from public publications on arXiv. The evaluation benchmarks, LR-Bench and CMU Gold, are publicly released for research purposes. These datasets contain no sensitive personal information beyond publicly available publication metadata. For model training, we initialize from Qwen3-4B and Qwen3-Embedding-8B, both released under open-source licenses. We use the

verl framework (Apache 2.0 License) for GRPO training. For repositories where a specific license was not explicitly provided, we have used them strictly in accordance with their intended research purposes. Our code will be released under an open-source license to facilitate reproducibility.

References

- Meltem Aksoy, Seda Yanik, and Mehmet Fatih Amasyali. 2023. [Reviewer assignment problem: A systematic review of the literature](#). *J. Artif. Int. Res.*, 76.
- Omer Anjum, Hongyu Gong, Suma Bhat, Wen-Mei Hwu, and JinJun Xiong. 2019. [PaRe: A paper-reviewer matching approach using a common topic space](#). 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 518–528, Hong Kong, China. Association for Computational Linguistics.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *ArXiv*, abs/2505.08775.
- Laurent Charlin and Richard S. Zemel. 2013. [The toronto paper matching system: An automated paper-reviewer assignment system](#).
- 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.
- Don Conry, Yehuda Koren, and Naren Ramakrishnan. 2009. [Recommender systems for the conference paper assignment problem](#). In *Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09*, page 357–360, New York, NY, USA. Association for Computing Machinery.
- 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.
- 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.
- Mei Rosemary Fu, Steven R. Chesnut, Anita Skarbek, and Sarah E. Patel. 2025. [High-quality peer review for scientific manuscripts](#). *Journal of Transcultural Nursing*, 36:473 – 474.
- Deyan Ginev. 2024. ar5iv:04.2024 dataset, an html5 conversion of arxiv.org. hosted at <https://sigmathling.kwarc.info/resources/ar5iv-dataset-2024/>. SIGMathLing – Special Interest Group on Math Linguistics.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. 2025. [Rubrics as rewards: Reinforcement learning beyond verifiable domains](#). In *NeurIPS 2025 Workshop on Efficient Reasoning*.
- Sunanda Gupta and Aruddha Sarkar. 2025. [Peer review of scientific studies: Problems and potential solutions](#). *Cureus*, 17.
- Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Selina Peng, Beibin Li, Shengjie Bi, Shishir Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujun Kumar Gonugondla, Hunter Lang, Yue Yu, and 6 others. 2025. [Advancedif: Rubric-based benchmarking and reinforcement learning for advancing llm instruction following](#). *ArXiv*, abs/2511.10507.
- Jhih-Yi Hsieh, Aditi Raghunathan, and Nihar B Shah. 2025. [Vulnerability of text-matching in ML/AI conference reviewer assignments to collusions](#). In *Championing Open-source DEvelopment in ML Workshop @ ICML25*.
- 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*.
- Jian Jin, Qian Geng, Qian Zhao, and Lixue Zhang. 2017. [Integrating the trend of research interest for reviewer assignment](#). In *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion*, page 1233–1241, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- 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.
- Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025. [Position: The AI conference peer review crisis demands](#)

- author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *The Twelfth International Conference on Learning Representations*.
- 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.
- Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2025. [Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment](#). *ArXiv*, abs/2510.07743.
- Weicong Liu, Zixuan Yang, Yibo Zhao, and Xiang Li. 2026. [RATE: Reviewer profiling and annotation-free training for expertise ranking in peer review systems](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43978–43996, San Diego, California, United States. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. 2013. [On good and fair paper-reviewer assignment](#). *2013 IEEE 13th International Conference on Data Mining*, pages 1145–1150.
- 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.
- 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tribikram Pradhan, Suchit Sahoo, Utkarsh Singh, and Sukomal Pal. 2021. [A proactive decision support system for reviewer recommendation in academia](#). *Expert Systems with Applications*, 169:114331.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. [InFoBench: Evaluating instruction following ability in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. [Hybridflow: A flexible and efficient rlhf framework](#). *arXiv preprint arXiv:2409.19256*.
- Yuzhen Shi, Huanghai Liu, Yiran Hu, Gaojie Song, Xinran Xu, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Di Feng, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Yuanyang Qiu, Yuemeng Qi, Jingwen Zhang, and 11 others. 2026. [Plawbench: A rubric-based benchmark for evaluating llms in real-world legal practice](#). *Preprint*, arXiv:2601.16669.
- 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, John Frederick Wieting, Yang Xi, Graham Neubig, and Nihar B Shah. 2025. [A gold standard dataset for the reviewer assignment problem](#). *Transactions on Machine Learning Research*.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Fan Wang, Ning Shi, and Ben Chen. 2010. [A comprehensive survey of the reviewer assignment problem](#). *International Journal of Information Technology & Decision Making (IJITDM)*, 9(04):645–668.
- Zhichao Yang, Sepehr Janghorbani, Dongxu Zhang, Jun Han, Qian Qian, Andrew Ressler, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. 2026. [Health-score: Towards scalable rubrics for improving health-llms](#). *ArXiv*, abs/2601.18706.
- Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. [FLASK: Fine-grained language model evaluation based on alignment skill sets](#). In *The Twelfth International Conference on Learning Representations*.

- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2026. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Dong Zhang, Shu Zhao, Zhen Duan, Jie Chen, Yanping Zhang, and Jie Tang. 2020. [A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation](#). *ACM Trans. Inf. Syst.*, 38(1).
- 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](#). *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.

Appendix

A Additional Experiments

A.1 Sensitivity to profile size.

We analyze sensitivity to K , the number of reviewer publications used to construct paper-conditioned profiles, holding all other settings fixed and applying the same K during both training and evaluation. As shown in Table 4, performance follows an inverted-U pattern: accuracy rises from $K=1$ (72.93%) to a peak at $K=3$ (75.92%, 0.1741 loss), then declines at $K=4$ and $K=5$ (73.04%). Too few publications leave reviewer expertise underspecified, while too many introduce marginally relevant work that dilutes the profile signal. We adopt $K=3$ as the default setting.

Table 4: Sensitivity to the number of selected reviewer publications K . We report expertise-aligned loss and pairwise accuracy on LR-Bench.

K	Loss ($\downarrow$)			Acc. ($\uparrow$)		
	PC	RC	Avg.	PC	RC	Avg.
1	0.2118	0.1930	0.2024	70.95%	74.91%	72.93%
2	0.2195	0.1805	0.2000	70.95%	76.56%	73.76%
3	0.1698	0.1783	0.1741	75.00%	76.84%	75.92%
4	0.1813	0.1953	0.1883	75.00%	75.51%	75.26%
5	0.2004	0.2078	0.2041	<u>72.97%</u>	73.10%	73.04%

A.2 Cost Analysis

To assess the practical efficiency of MERIT, we analyze its cost along two axes: one-time offline training and deployment.

Table 5: End-to-end offline cost of the MERIT pipeline.

Stage	Compute	API cost
RATE candidate mining	≈ 21 GPU-h	—
Rubric generation	—	\$23.58
Rubric-guided judging	—	\$78.57
GRPO training	≈ 245 GPU-h	—
Assessor pseudo-labeling	≈ 114 GPU-h	—
Retriever distillation	≈ 6 GPU-h	—
Total	≈ 386 GPU-h	\$102.15

Offline training cost. Table 5 breaks down the full offline pipeline of MERIT. GRPO training dominates the compute budget, while rubric-guided judging accounts for most of the API cost. However, the entire cost is incurred once: after the retriever is distilled, no further LLM calls or policy-model inference is required at deployment.

Table 6: Deployment cost on 1,000 randomly sampled preference triplets using a single NVIDIA A800-80GB GPU.

	Prep. (s)	Infer. (s)	LLM cost	GPU cost	Total cost
RATE	679	220	\$1.57	\$0.06	\$1.63
MERIT	60	713	\$0.00	\$0.20	\$0.20

Deployment cost. At deployment, MERIT runs only the embedding retriever, whereas RATE-8B additionally requires generative-LLM-based keyword extraction for reviewer profiles. We benchmark both methods on the same 1,000 randomly sampled preference triplets using one NVIDIA A800-80GB GPU.¹ As shown in Table 6, RATE’s keyword generation step dominates its wall time and cost, with the generative-LLM API fee accounting for over 96% of the total. MERIT eliminates this dependency, reducing the deployment cost to GPU rental alone—an $8\times$ reduction overall.

A.3 Human Validation

Since MERIT relies on automatically generated expertise rubrics and an LLM judge to provide training supervision, we conduct two human validation studies to assess the reliability of these components.

Annotation protocol. The evaluation team consisted of five master’s and doctoral students in computer science with relevant research experience. All annotators were compensated at a rate exceeding the local minimum wage and gave informed consent before participation. Annotators received task instructions and reference examples demonstrating the application of the evaluation criteria. All judgments were made independently, and human labels were determined by majority vote (at least 3 of 5 annotators).

Rubric quality. To validate the generated expertise rubrics, we randomly sampled 100 papers from the Stage-1 training set, yielding 405 rubric items (189 Core and 216 Secondary). For each rubric item, annotators applied the evaluation criterion corresponding to its rubric tier:

- **Core:** whether it captures expertise essential for evaluating the paper’s central contribution.
- **Secondary:** whether it captures supporting expertise relevant to evaluating the paper.

¹ GPU rental cost is estimated at \$0.93 per GPU-hour based on prevailing cloud rates at the time of writing.

As shown in Table 7, 89.6% of rubric items were validated by majority vote, with Core and Secondary items reaching 91.0% and 88.4%, respectively. The high validation rates across both tiers suggest that the generated criteria generally align with their intended roles.

Table 7: Human validation of generated expertise rubrics. Validation labels are determined by majority vote among five annotators.

Type	Items	Validated	Not validated
Core	189	172 (91.0%)	17 (9.0%)
Secondary	216	191 (88.4%)	25 (11.6%)
Overall	405	363 (89.6%)	42 (10.4%)

Judge-human alignment. To assess the reliability of the LLM judge’s evaluations, we sampled 100 paper-reviewer pairs from the training set and used the base Qwen3-4B with the training policy prompt to generate suitability assessments. The LLM judge and five annotators then independently evaluated these assessments for rubric coverage and decision consistency. We measured judge-human agreement using the human majority-vote labels as reference.

As shown in Table 8, the LLM judge achieves over 82% F1 on both rubric coverage and decision consistency. These results suggest that the judge provides a reasonably reliable supervision signal for training the MERIT-Assessor.

Table 8: LLM judge alignment with human evaluators on rubric coverage and decision consistency. Human reference labels are determined by majority vote among five annotators.

Type	Acc.	P.	R.	F1
Core coverage	82.7%	82.0%	87.2%	84.5%
Secondary coverage	79.5%	79.0%	83.8%	81.3%
Overall coverage	80.9%	80.4%	85.3%	82.8%
Decision consistency	79.0%	76.1%	91.1%	82.9%

B Detailed Experimental Settings

B.1 Computing Facilities

We train the MERIT-Assessor with GRPO on 4× NVIDIA A800-80GB GPUs and fine-tune the MERIT-Retriever on 2× NVIDIA A800-80GB GPUs. All evaluation experiments are conducted on a single NVIDIA A800-80GB GPU.

B.2 Training Hyperparameters

Tables 9 and 10 list the main hyperparameters for the MERIT-Assessor and the MERIT-Retriever, respectively.

B.3 Baseline details

Table 11 summarizes the implementation sources and model checkpoints used for retrieval baselines. All baselines are evaluated using their released checkpoints.

For baselines that score a reviewer through profile publications, we follow their original evaluation protocols whenever available, including the profile construction and score aggregation strategies, to ensure a fair comparison. BERTScore, SPECTER, SPECTER2-PRX, and SciNCL are paper-level encoders that produce similarities between individual document pairs. To obtain a paper-reviewer score, we compute similarities between the target paper and each of the reviewer’s publications and aggregate via 75th-percentile pooling (Hsieh et al., 2025). For CoF and RATE, which include their own reviewer-level scoring procedures—top-3 averaging and keyword-based profile construction, respectively—we follow their original protocols.

Table 9: Main hyperparameters for training the MERIT-Assessor.

Parameter	Training	Inference
clip_low	0.2	N/A
clip_high	0.28	N/A
learning_rate	$5e - 7$	N/A
max_prompt_length	5120	N/A
max_response_length	4096	4096
total_epochs	1	N/A
train_prompt_batchsize	16	N/A
group_size	8	N/A
train_prompt_mini_batchsize	8	N/A
temperature	1.0	0.0
top_p	1.0	N/A

C Large Language Model Prompts

We provide the prompt templates used in our method and LLM-based baselines. All prompts use placeholder variables (e.g., {{paper_title}}) that are filled at inference time.

C.1 Prompts for MERIT-Assessor Training

Stage 1 involves three prompts that operate in sequence: (1) a rubric generation prompt that pro-

Table 10: Main hyperparameters for fine-tuning the MERIT-Retriever.

Parameter	Value
Mixed precision	bf16
Attention implementation	FlashAttention-2
Batch size	16
Evaluation batch size	4
Gradient accumulation steps	2
Epochs	1
Learning rate	$4e - 5$
Warmup ratio	0.01
Temperature	0.04
λ_{nce}	0.5
Max query length	2048
Max document length	2048
Patience	100
Evaluation frequency	10 steps
Seed	42
LoRA rank	32
LoRA alpha	64
LoRA dropout	0.06

duces paper-specific expertise criteria from the target paper’s title, abstract, and introduction (Figure 3); (2) a policy-model prompt that elicits a structured chain-of-thought assessment of reviewer suitability (Figure 4); and (3) an LLM-judge prompt that evaluates each assessment against the generated rubric to compute the gated reward (Figure 5).

Table 11: Implementation sources and model checkpoints for retrieval baselines.

Method	Source / Checkpoint
TPMS	niharshah/reviewer-paper-match
BERTScore	Tiiiger/bert_score
SPECTER	allenai/specter
SciNCL	malteos/scincl
SPECTER2-PRX	allenai/specter2
CoF	yuzhimanhua/CoF
RATE-8B	Gnociew/RATE-Reviewer-Assignment

C.2 Prompts for LLM-Based Baselines

We use baseline prompts corresponding to the two evaluation settings. For reviewer suitability classification, we evaluate two prompting strategies: direct prompting predicts a binary label without explicit expertise decomposition (Figure 6), while expertise-aware prompting uses the same structured assessment template as our policy model (Figure 4) but without RL training. For reviewer retrieval, we use a zero-shot scoring template that elicits a 1–5 expertise rating for pairwise ranking (Figure 7).

D Example Expertise Rubric

Table 12 presents a human-written example of a paper-specific expertise rubric. This exemplar is provided to the LLM as an in-context demonstration during rubric generation (see the prompt in Figure 3). It illustrates the target structure: 1–2 Core criteria (weight 5) capturing expertise essential to the paper’s central contribution, and several Secondary criteria (weight 3–4) covering supporting knowledge. Criterion descriptions specify reviewer capabilities rather than paper-specific details, ensuring that the rubric generalizes to candidate reviewers with diverse publication backgrounds.

E Case Studies

We present case studies examining three aspects of our framework: the MERIT-Assessor’s prediction behavior (Section E.1), errors corrected by the MERIT-Retriever over the strongest baseline (Section E.2), and the retriever’s remaining failure modes (Section E.3).

E.1 Assessor Case Studies

Tables 13–16 cover the four prediction outcomes of the MERIT-Assessor. The correct predictions illustrate how the model grounds suitability decisions in paper-specific expertise requirements and evidence from candidate reviewers’ prior work. The error cases reveal a recurring pattern in evidence-transfer calibration: the model may over-transfer superficially related terminology across methodological contexts (false positive) or under-transfer relevant expertise when the candidate’s work differs in a specific technical variant (false negative).

E.2 Retriever Error Corrections

To understand what types of errors MERIT-Retriever corrects, we analyze LR-Bench preference pairs where RATE ranks incorrectly but MERIT ranks correctly. Among the 88 corrected pairs, 63 (71.6%) have an rating gap of 1, compared with 19 (21.6%) for gap 2 and 6 (6.8%) for gap 3. Among the errors corrected by MERIT, most involve adjacent expertise ratings, indicating that MERIT can resolve many close-call distinctions that RATE misses. Two representative cases are presented below.

Case 1: 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	0.7266	0.7364
MERIT	0.7627	0.6663

Interpretation. Although the anchor paper contains strong RAG-related surface terms, its core contribution is a hashing-based retrieval framework that learns binary hash codes for efficient storage and fine-grained retrieval. The positive reviewer has direct expertise in semantic hashing and hash-code learning, while the negative reviewer is primarily oriented toward general RAG and LLM-based retrieval. RATE ranks the negative reviewer higher, possibly due to broader retrieval-related topical overlap. MERIT instead gives greater weight to the paper’s methodological core—hashing-based retrieval—suggesting that suitability supervision derived from criterion-level assessment helps the retriever distinguish central expertise from broader topical overlap.

Case 2: Anchor paper. *DVLTA-VQA: Decoupled Vision-Language Modeling with Text-Guided Adaptation for Blind Video Quality Assessment*

Preference and model scores.

	Positive	Negative
Expert preference	5	4
RATE	0.7091	0.7140
MERIT	0.8468	0.7127

Interpretation. The anchor paper is broadly related to visual-quality modeling, but its central evaluation methodology requires expertise in video and image quality assessment together with VLM/CLIP-based quality modeling, rather than general image aesthetics or video restoration. The positive reviewer has published directly on CLIP-based video quality assessment, while the negative reviewer focuses on image aesthetics and video super-resolution. RATE captures broad visual-quality similarity but does not distinguish these roles; MERIT identifies the expertise more directly relevant to the paper’s central methodology.

These cases suggest that MERIT better distinguishes expertise aligned with a paper’s central contribution from broader topical overlap.

E.3 Retriever Failure Modes

We inspect MERIT-Retriever’s misranked LR-Bench pairs and find that remaining errors often occur when both candidates are strongly related to the anchor and the correct ordering depends on finer distinctions in expertise breadth, specialization, or strength. Two representative cases below each illustrate a different failure subtype.

Case 1: Anchor reviewer. *Generative Large Recommendation Models: Emerging Trends in LLMs for Recommendation; A Unified Framework for Adaptive Representation Enhancement and Inversed Learning in Cross-Domain Recommendation; Multi-granularity Interest Retrieval and Refinement Network for Long-Term User Behavior Modeling in CTR Prediction.*

Preference and model scores.

	Positive	Negative
Expert preference	5	4
MERIT	0.8556	0.8603

Interpretation. The reviewer’s profile spans multiple recommendation sub-areas, including LLM-based recommendation, cross-domain recommendation, and CTR prediction. The positive paper is a broad survey on data scarcity in recommendation systems, whose evaluation demands breadth across the field. The negative paper targets long sequence modeling in industrial recommenders, offering a more concentrated technique-level connection to the reviewer’s CTR work. MERIT assigns high scores to both but slightly prefers the technique-level match, reversing the expert’s preference for the survey. This suggests a breadth–depth calibration gap: MERIT may over-emphasize a concentrated methodological match when breadth across the field is itself the relevant expertise requirement.

Case 2: Anchor reviewer. *Federated Markov Imputation: Privacy-Preserving Temporal Imputation in Multi-Centric ICU Environments; Distribution-Controlled Client Selection to Improve Federated Learning Strategies; Improving Early Sepsis Onset Prediction Through Federated Learning.*

Preference and model scores.

	Positive	Negative
Expert preference	4	3
MERIT	0.7994	0.8136

Interpretation. The reviewer has direct prior work on client selection in federated learning. The positive paper proposes a Shapley-based client selection method, matching this specialization directly, while the negative paper addresses federated-learning heterogeneity more generally through similarity and complementarity. MERIT captures the shared federated-learning domain but fails to sufficiently prioritize the reviewer's more specific specialization in client selection.

These cases highlight a limitation of binary suitability supervision: it provides limited signal for finer distinctions in expertise breadth, specialization, and strength. More fine-grained preference supervision may help resolve such cases.

Table 12: Human-written expertise rubric for the paper *Search-R1*. Each row reports a criterion title t_j and description d_j , together with its Core/Secondary role and importance weight w_j .

Input paper <i>Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning</i>		
Criterion	Type	Description
RLHF & RL for LLMs	Core Weight 5	Strong expertise in applying reinforcement learning algorithms (specifically PPO or GRPO) to large language models, including understanding policy optimization stability and reward function design.
Retrieval-Augmented Generation (RAG)	Core Weight 5	Deep understanding of RAG architectures, particularly multi-turn retrieval, query generation strategies, and the integration of external search results into LLM context.
Agentic Tool Use & Environments	Secondary Weight 4	Familiarity with modeling LLMs as agents that interact with external environments (search engines) via special tokens or API calls in an interleaved manner.
Reasoning & Chain-of-Thought	Secondary Weight 3	Knowledge of advanced reasoning frameworks (e.g., Chain-of-Thought, DeepSeek-R1 styles) to evaluate the quality of the model’s generated reasoning trajectories.

Table 13: True positive case. The model correctly identifies the reviewer as suitable based on strong evidence for core GEC evaluation expertise.

Case study: Correct positive prediction
Input: Target paper Title: <i>CLEME2.0: Towards Interpretable Evaluation by Disentangling Edits for Grammatical Error Correction</i> TL;DR: The paper proposes an interpretable evaluation metric for grammatical error correction by disentangling edits into fine-grained correction types and evaluating consistency with human judgments .
Key fields from the policy-model output [REQUIRED_EXPERTISE] CORE Grammatical Error Correction evaluation metrics CORE Edit-level analysis and metric interpretability SECONDARY Reference-based vs. reference-free evaluation methods SECONDARY Adversarial testing, robustness, and human consistency of GEC metrics
[EVIDENCE_CLUES] <i>Rethinking Evaluation Metrics for Grammatical Error Correction:</i> Supports CORE (GEC evaluation), Secondary (human-aligned evaluation), and Secondary (benchmarking of automatic metrics). <i>Reliability Crisis of Reference-free Metrics:</i> Supports Secondary (adversarial testing), Secondary (reference-free metrics), and Secondary (robustness). <i>Improving Explainability of Sentence-level Metrics:</i> Supports Secondary (attribution methods) and Secondary (interpretability).
[REASONING_SUMMARY] The candidate has extensive expertise in GEC evaluation, including both reference-based and reference-free metrics, and their limitations. Their work on human consistency, benchmarking, and meta-evaluation aligns with the target paper’s evaluation on human-judgment datasets and multiple reference datasets. Their work on adversarial attacks further provides supporting expertise on the reliability limitations of automatic GEC metrics.
Policy-model decision SUITABLE Final label: 1.
Interpretation. This case illustrates how the policy model identifies paper-specific expertise requirements, links them to evidence from a reviewer’s prior work, and produces an evidence-grounded positive suitability decision.

Table 14: True negative case. The model correctly identifies partial topical overlap in graph-based recommendation but insufficient evidence for the paper’s core requirements in rule-driven attribute embedding.

Case study: Correct negative prediction	
Input: Target paper Title: <i>RAE: A Rule-Driven Approach for Attribute Embedding in Property Graph Recommendation</i> TL;DR: The paper proposes a rule-driven attribute embedding framework for property graph recommendation. It mines semantic dependency rules from property graphs, uses them to guide random walks for enriched attribute embeddings, and integrates these embeddings into GCN-based recommendation models to improve robustness under data sparsity and attribute missingness .	
Key fields from the policy-model output [REQUIRED_EXPERTISE] CORE Rule-driven attribute embedding in property graphs CORE Semantic rule mining and attribute dependency modeling SECONDARY GNN/GCN-based recommendation SECONDARY Data sparsity and attribute missingness	
[EVIDENCE_CLUES] <i>Federated Heterogeneous Graph Neural Network for Privacy-preserving Recommendation:</i> Provides partial overlap in HIN-based recommendation and graph neural networks , but does not address rule-driven attribute embedding , property-graph attribute dependencies , or semantic rule mining . <i>Federated Graph Condensation with Information Bottleneck Principles:</i> Focuses on graph condensation , federated graph learning , and privacy protection , rather than attribute embeddings or rule-guided representation learning . <i>Phantom Subgroup Poisoning: Stealth Attacks on Federated Recommender Systems:</i> Studies poisoning attacks in federated recommender systems, but not attribute-based recommendation , property graphs , or semantic dependency rules . <i>Rethinking Byzantine Robustness in Federated Recommendation:</i> Focuses on Byzantine robustness and sparse aggregation in federated recommendation, rather than attribute modeling or rule-based embeddings .	
[REASONING_SUMMARY] The candidate has related expertise in graph-based recommendation , federated learning , and robustness/security for recommender systems. However, the target paper’s main contribution is not federated recommendation or graph robustness, but a rule-driven attribute embedding method that mines semantic dependency rules from property graphs and uses them to guide representation learning. The candidate’s prior work provides only partial overlap through HINs, GNNs, and sparsity-aware recommendation, but does not provide direct evidence of expertise in property-graph attribute modeling , semantic rule mining , or rule-guided attribute embeddings .	
Policy-model decision UNSUITABLE Final label: 0.	
Interpretation. This case illustrates how the policy model distinguishes partial topical overlap from true reviewer suitability: although the candidate has related expertise in graph and federated recommendation, the model correctly identifies the lack of evidence for the target paper’s core requirements in rule-driven attribute embedding and semantic rule mining.	

Table 15: False positive case. The model incorrectly transfers superficially related terminology (counterfactual explanations, predictive multiplicity) to the target paper’s causal inference requirements.

Case study: False positive prediction	
Input: Target paper Title: <i>Self Balancing Neural Network: A Novel Method to Estimate Average Treatment Effect</i> TL;DR: The paper proposes a self-balancing neural network for estimating the average treatment effect in observational studies. It addresses bias from confounding variables and propensity score misspecification by learning pseudo propensity scores through a balancing network integrated into the outcome model.	
Key fields from the policy-model output [REQUIRED_EXPERTISE] CORE Causal inference in observational studies CORE Propensity score modeling and balancing techniques SECONDARY Machine learning methods for causal estimation SECONDARY Addressing model misspecification in causal models SECONDARY Neural network architectures for causal inference	
[EVIDENCE_CLUES] <i>Explainable Bank Failure Prediction Models:</i> Discusses counterfactual explanations , model interpretability , and the reliability of black-box predictive models. <i>Predictive Multiplicity in Survival Models:</i> Addresses predictive multiplicity , model disagreement, and uncertainty in survival-based predictive modeling. <i>Beyond the Single-Best Model: Rashomon Partial Dependence Profile:</i> Explores model multiplicity , explanation uncertainty , and trustworthy model interpretation.	
[REASONING_SUMMARY] The candidate’s work spans causal inference , model reliability , and predictive multiplicity , which are central to the target paper’s focus on addressing confounding and model misspecification. While the candidate does not explicitly mention self-balancing neural networks or propensity score modeling , their focus on model consistency, uncertainty, and robustness aligns with the paper’s technical contributions.	
Policy-model decision SUITABLE Predicted label: 1.	
Interpretation. This error case reveals a terminology-transfer error : the policy model transfers superficially related terms across different methodological contexts. In particular, counterfactual explanations in explainable AI do not provide direct evidence of expertise in observational causal inference, and predictive multiplicity or generic model reliability does not establish expertise in propensity score modeling or average treatment effect estimation .	

Table 16: False negative case. The model applies overly strict technique-specific matching, underestimating the reviewer’s transferable expertise in diffusion-based image editing.

Case study: False negative prediction	
Input: Target paper Title: <i>FramePainter: Endowing Interactive Image Editing with Video Diffusion Priors</i> TL;DR: The paper proposes FramePainter , an interactive image-editing method that reformulates image editing as an image-to-video generation task. It leverages video diffusion priors from Stable Video Diffusion and introduces matching attention to improve dense correspondence and visual consistency under large object motion.	
Key fields from the policy-model output [REQUIRED_EXPERTISE] CORE Video diffusion models for image editing SECONDARY Techniques for handling dynamic motion between frames SECONDARY Attention mechanisms for image-to-video generation SECONDARY Consistent object editing with diffusion models SECONDARY Image editing with spatial-temporal correspondence SECONDARY Evaluation of generalization in out-of-domain scenarios	
[EVIDENCE_CLUES] <i>PixelMan: Consistent Object Editing with Diffusion Models via Pixel Manipulation and Generation:</i> Supports Secondary (consistent object editing with diffusion models). <i>GOAt: Explaining Graph Neural Networks via Graph Output Attribution, Building Optimal Neural Architectures using Interpretable Knowledge, and EiG-Search: Generating Edge-Induced Subgraphs for GNN Explanation in Linear Time:</i> Treated as irrelevant to the target paper.	
[REASONING_SUMMARY] The candidate’s work is focused on graph neural networks , neural architecture search , and diffusion-based object editing . While <i>PixelMan</i> involves diffusion models for object editing , it is limited to static image manipulation and does not address video diffusion priors , dynamic motion handling , or image-to-video generation . The candidate lacks explicit expertise in video diffusion models , attention mechanisms for temporal correspondence , and the specific technical challenges of FramePainter , such as leveraging video diffusion priors for interactive editing and handling large motion between frames.	
Policy-model decision UNSUITABLE Predicted label: 0.	
Interpretation. This false negative reflects overly strict technique-specific matching : the policy model emphasizes the absence of direct evidence in video diffusion priors , image-to-video generation , and matching attention , but underestimates the candidate’s transferable expertise in diffusion-based image editing , consistent object manipulation , and appearance-preserving editing .	

Figure 3: Prompt template for paper-specific expertise rubric generation.

```
### System Prompt
You generate reviewer expertise rubrics for a research paper.

Task:
Infer the generalized expertise a qualified reviewer should have based on the paper title, abstract, and introduction.

Rules:
- Do not require expertise in the paper's specific proposed method, acronym, dataset, or baseline names.
- Abstract paper-specific details into transferable expertise such as domain knowledge, methodology, theory, or evaluation skills.
- Keep rubric items distinct and avoid overlap.
- Prefer the paper's core methodological or theoretical expertise over broad topic labels.
- Use the smallest sufficient set of expertise areas needed to evaluate the paper.
- Do not split closely related skills into separate items.
- Add an item only if it captures a distinct reviewer expertise that cannot be merged with another item.

Output:
Return only a JSON array of 3 to 6 items.
Use 1 or 2 CORE items only.
Use 2 CORE items only when the paper requires two distinct and indispensable expertise areas; otherwise use 1.
All remaining items must be Secondary.

Each item must have exactly:
- "title"
- "description"
- "weight"

Description prefix rules:
- If weight = 5, the description must begin with "CORE:".
- If weight = 3 or 4, the description must begin with "Secondary:".

Weight meaning:
- CORE Criteria (Weight: 5)
  - Definition: The indispensable expertise required to understand and evaluate the paper's main contribution.
- Secondary Criteria (Weight: 3-4)
  - Definition: Supporting expertise needed to verify important methodological, theoretical, empirical, or evaluation details.

Quality requirements:
- Titles should describe reviewer expertise, not paper modules.
- Descriptions should specify the scope of the required expertise and explain how it is needed for evaluating the paper.
- Avoid overly fine-grained or redundant criteria.
- Prefer concise descriptions.

### Examples
{examples_string}

-----

### User Prompt
[Paper Title]
{paper_title}

[Paper Abstract]
{paper_abstract}

[Paper Introduction]
{paper_introduction}

Generate the reviewer-expertise rubric as a JSON array.

Return JSON only.
```

Figure 4: Prompt template for the policy model’s reviewer suitability assessment. This template is also used by the expertise-aware prompting baselines.

```

### System Prompt
You are an expert Conference Area Chair. Your task is to make a final "Accept/Reject" decision on whether a Candidate Reviewer is qualified to review a specific Target Paper.

Input Data:
1. Target Paper (Title, Abstract, Introduction)
2. Candidate Author Profile (List of historical publications)

Decision Logic:
- Label 1 (Qualified): The candidate possesses at least ONE of the CORE (Primary) expertise requirements AND at least one additional relevant skill (Secondary or another Core). They have sufficient domain knowledge to evaluate the paper's key contributions.
- Label 0 (Unqualified): The candidate matches ONLY secondary aspects, OR matches a CORE concept in isolation without any supporting context (e.g., knows the general theory but lacks the specific application domain), OR has no relevance at all.

Instruction:
You must strictly follow the output structure below.

Step 1: Analyze the Target Paper. Extract 3 to 6 distinct expertise requirements.
- Constraint: Select the most critical topics. Do not list more than 6 items to avoid dilution.
- Mark the central problem/method as "(CORE)" and enabling technologies as "(Secondary)".
- Important: Define CORE as the underlying problem/method (e.g., "Diffusion Models") rather than the specific novelty.
Step 2: Scan the Candidate's profile. Select specific papers that serve as evidence.
- Crucial: Look for semantic matches.
Step 3: Synthesize the analysis. Evaluate the match based on the Combination Rule.
- Threshold: The candidate MUST demonstrate expertise in at least ONE CORE requirement, PLUS at least one additional requirement (which can be a Secondary or another CORE).
- Note: Do NOT demand the candidate to master ALL COREs if multiple exist, as interdisciplinary papers often require a panel of different experts.

Output Format (Strictly use these headers):
[REQUIRED_EXPERTISE]
<List 3 to 6 requirements here, one per line. Be granular but focus on capabilities. Example:
1. Adversarial Patch Attacks on Vision Transformers (CORE)
2. Quantization-Aware Training (Secondary)
3. Image Classification Benchmarks (Secondary)>

[EVIDENCE_CLUES]
<List specific paper titles and a brief tag indicating which requirement they support. Example:
- "DeepPatch: Attacking ViTs": Supports CORE (Adversarial Patch Attacks).
- "Q-ViT: Robust Quantization": Supports Secondary (Quantization).
If no relevant papers are found, write 'None'.>

[REASONING_SUMMARY]
<Provide a holistic analysis. Explicitly discuss:
1. Whether the candidate covers the (CORE) expertise.
2. Which (additional) expertise supports the CORE match.
3. Conclusion based on the "Core + 1 additional" rule.>

[FINAL_LABEL]
<0 or 1>

-----

### User Prompt
[[Target Paper]]
Title: {{paper_title}}
Content: {{paper_abstract}}

{{paper_introduction}}

[[Candidate Author Profile]]
(Recent Publications)
{{candidate_history_list}}

Based on the strict standard (1=Expert Match, 0=Everything Else), evaluate this pair.

```

Figure 5: Prompt template for the LLM judge used to compute rubric-guided gated rewards.

```

### System Prompt
You are the Senior Program Chair (SAC) for a top-tier Computer Science conference.
You are auditing a "Reviewer Suitability Analysis" to ensure the evaluation is fair, evidence-based, and logically sound.

Your Goal: Detect incompetent analysis, logical contradictions, or "gaming the system" (e.g., keyword stuffing, rejecting
qualified candidates to play it safe, or accepting unqualified ones).

You have access to:
1. Target Paper Context.
2. Candidate's Verified Publications.
3. Gold Rubrics.
4. The Report to be audited.

Output Structure:
Output **ONLY** the strict JSON object.

JSON Schema:
1. "rubric_breakdown": { "Rubric Title": boolean }
   - true: The Report **actively addresses and discusses** this requirement.
     * It goes beyond merely listing the string; it evaluates whether the candidate possesses or lacks this specific
     expertise (e.g., linking it to a specific paper or explicitly stating it is missing in the reasoning).
   - false: The requirement is ignored, OR it is merely "keyword stuffed" (copy-pasted into a list without being analyzed
   against the candidate's profile).

2. "logical": Boolean.
   - true: The decision is **sound, fair, and strongly supported** by the "1 Core + 1 additional requirement" threshold rule.
     * If Label 1: The Report accurately identifies at least ONE Core and ONE additional requirement, supported by valid
     cited papers.
     * If Label 0: The Report correctly proves the candidate lacks Core expertise or lacks additional requirement support.
   - false: The Analysis contains a **logical failure, contradiction, or unfairness**. Examples of logical failures:
     * **False Positive (Bad 1):** Giving Label 1 based on weak/tangential evidence (e.g., Candidate knows broad "AI", Paper
     needs "Diffusion" -> logical is false) or hallucinated evidence.
     * **False Negative (Bad 0):** Giving Label 0 despite the candidate clearly meeting the "1 Core + 1 additional" threshold
     (e.g., demanding a 100% perfect match on a specific novelty when the candidate has sufficient foundational expertise).

     * **Contradiction:** The text argues the candidate is a strong match but outputs Label 0 (or vice versa).

-----

### User Prompt
=== 1. Target Paper Context ===
Title:
{{paper_title}}

Abstract:
{{paper_abstract}}

Introduction:
{{paper_introduction}}

=== 2. Candidate Profile ===
{{candidate_history_list}}

=== 3. Gold Rubrics ===
{{rubrics_json}}

=== 4. Report (To be Audited) ===
{{actor_output_text}}

=== Task ===

Audit the Report fairly and objectively.
1. **Check for Keyword Stuffing:** Mark rubric items as false if they are just listed in the requirements section but never
actually evaluated against the candidate's history in the reasoning/evidence.
2. **Check for Logical Consistency:** Ensure the final label accurately reflects the evidence. Penalize over-claiming (
unjustified 1s) AND overly strict rejections (unjustified 0s when the candidate meets the 1 Core + 1 additional
requirement threshold).

Generate the strict JSON output:
1. "rubric_breakdown": { "Rubric Title": boolean }
2. "logical": boolean.

```

Figure 6: Prompt template for the direct prompting baseline in reviewer suitability classification.

```
### System Prompt
You are an expert Conference Area Chair. Determine whether a Candidate Reviewer is qualified to review a Target Paper.

Input:
1. Target Paper (Title, Abstract, Introduction)
2. Candidate Author Profile (historical publications)

Decision Rule:
- Label 1: The candidate has sufficient expertise to review the target paper.
- Label 0: The candidate lacks the expertise needed to provide a competent review.

Output Format:
[FINAL_LABEL]
<0 or 1>
-----

### User Prompt
[[Target Paper]]
Title: {{paper_title}}
Content: {{paper_abstract}}
{{paper_introduction}}

[[Candidate Author Profile]]
(Recent Publications)
{{candidate_history_list}}

Based on the candidate's research background, are they a suitable reviewer for this target paper?
```

Figure 7: Prompt template for zero-shot LLM scoring in reviewer retrieval evaluation.

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

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

[Reviewer Profile]
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 is 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.
```