Navigating Through Paper Flood: Advancing LLM-Based Paper Evaluation Through Domain-Aware Retrieval and Latent Reasoning

Wuqiang Zheng[1], **Yiyan Xu**[1*], **Xinyu Lin**[2], **Chongming Gao**[1], **Wenjie Wang**[1*], **Fuli Feng**[1]
[1]University of Science and Technology of China
[2]National University of Singapore
{qqqqqzheng, yiyanxu24, chongming.gao, xylin1028, wenjiewang96, fulifeng93}@gmail.com

Abstract

With the rapid and continuous increase in academic publications, identifying high-quality research has become an increasingly pressing challenge. While recent methods leveraging Large Language Models (LLMs) for automated paper evaluation have shown great promise, they are often constrained by outdated domain knowledge and limited reasoning capabilities. In this work, we present PaperEval, a novel LLM-based framework for automated paper evaluation that addresses these limitations through two key components: 1) a domain-aware paper retrieval module that retrieves relevant concurrent work to support contextualized assessments of novelty and contributions, and 2) a latent reasoning mechanism that enables deep understanding of complex motivations and methodologies, along with comprehensive comparison against concurrently related work, to support more accurate and reliable evaluation. To guide the reasoning process, we introduce a progressive ranking optimization strategy that encourages the LLM to iteratively refine its predictions with an emphasis on relative comparison. Experiments on two datasets demonstrate that PaperEval consistently outperforms existing methods in both academic impact and paper quality evaluation. In addition, we deploy PaperEval in a real-world paper recommendation system for filtering high-quality papers, which has gained strong engagement on social media—amassing over 8,000 subscribers and attracting over 10,000 views for many filtered high-quality papers—demonstrating the practical effectiveness of PaperEval.

Code — https://github.com/ZhengWwwq/PaperEval

Introduction

In recent years, the explosive growth of academic publications has reflected the vitality of the research community, while simultaneously posing a critical challenge: How can researchers efficiently identify high-quality, impactful work to learn effectively and drive innovation? In this context, the task of automated paper evaluation is becoming increasingly crucial. It aims to evaluate paper quality and predict future impact, thereby facilitating the selection of high-quality work, supporting researchers in navigating the expanding scientific landscape, and ultimately promoting more efficient and impactful research progress.

Technically, the paper evaluation task aims to analyze the paper features to assess research quality from various dimensions, such as academic impact (Xia, Li, and Li 2022; Zhao et al. 2024) and overall quality (Lin et al. 2023). Existing studies mainly rely on traditional neural models or Large Language Models (LLMs) for this task:

- **Traditional methods** utilize neural models, such as Multi-Layer Perceptrons (MLPs), or Long-Short Term Memory networks (LSTM), to evaluate research papers based on predefined features, including structural indicators like paper length and reference count (Vergoulis et al. 2020; Ruan et al. 2020), as well as textual patterns (Ma et al. 2021; Yang et al. 2018). However, these methods often overlook the semantic content of papers (*e.g.,* abstract and main text), which is essential for accurate evaluation, ultimately leading to unsatisfactory performance.
- **LLM-based methods** leverage rich textual information (*e.g.,* title, abstract, and main text) to learn informative paper representations and employ a scoring module to produce evaluation scores. Empowered by the advanced semantic understanding capabilities of LLMs, these methods demonstrate strong potential in capturing the technical soundness of research papers, yielding more accurate evaluation results (Lu et al. 2024; Liu et al. 2025a; Zhao et al. 2025; de Winter 2024).

Despite promising progress, LLM-based methods still face notable limitations: 1) Due to the time lag in their training data, LLMs often lack awareness of newly published work, making it difficult to compare and assess the novelty and contribution in fast-evolving areas. 2) Research papers often contain intricate motivations and nuanced methodological designs that require deep reasoning beyond surface-level representation learning.

To address these limitations, we propose ***PaperEval***, a framework that retrieves domain-relevant reference papers, jointly encodes them with the target paper into an LLM, and performs latent reasoning to generate accurate evaluation.

- **Domain-aware paper retrieval.** To mitigate the issue of outdated domain knowledge, PaperEval integrates a retrieval module that identifies concurrent and thematically relevant work as reference papers, which are jointly fed

*Corresponding authors.

into the LLM along with the target paper for evaluation and provide essential context and background for LLMs to better evaluate the novelty and contributions of the target paper within the current research landscape.

- **Reasoning-enhanced paper evaluation.** Evaluating research papers requires a deep understanding of complex motivations and nuanced methodologies. This challenge is further intensified by the need to compare concurrently retrieved work. This motivates us to stimulate the reasoning mechanism of LLMs to support deep comprehension, precise comparison, and fair evaluation. While chain-of-thought reasoning (Wei et al. 2022) offers interpretable intermediate steps, it typically requires annotated reasoning paths for supervision (Weng et al. 2023), which are scarce in paper evaluation scenarios. In contrast, latent reasoning (Hao et al. 2024) enables implicit multi-step reasoning within the hidden representations of LLMs, eliminating the need for explicit annotations. More importantly, LLM-based paper evaluation focuses on learning more informative paper representations to enhance evaluation accuracy, which aligns naturally with the latent reasoning paradigm, seamlessly integrating reasoning directly at the representation level. As such, we incorporate latent reasoning into PaperEval for comprehensive representation learning.

Despite the significant potential of latent reasoning, an effective optimization strategy is essential to guide the reasoning process toward the ultimate ranking goal of paper evaluation. Specifically, the paper evaluation task focuses on comparing and identifying valuable work within a large collection, with a primary emphasis on relative ranking rather than absolute scoring. However, learning accurate rankings is inherently more challenging, as even minor prediction errors may cause substantial shifts in the ranking positions. To address this, we propose a **progressive ranking optimization** strategy, which encourages the latent reasoning to progressively improve relative ranking. In particular, at each reasoning step, we compute the temperature-controlled softmax over the predicted scores of a batch of papers, which is then aligned with the ground-truth order using a ranking loss. To gradually refine the LLM's ranking, we progressively decrease the temperature during latent reasoning, making the predicted distributions increasingly sharper and more sensitive to ranking errors. This progressive refinement encourages the LLM to iteratively produce more confident and discriminative rankings within each training batch. By learning to distinguish fine-grained differences among batch samples, the LLM enhances ranking reasoning capabilities, which can naturally generalize to global ranking across the entire dataset, as theoretically supported by (Lan et al. 2009).

We evaluate the effectiveness of PaperEval on two datasets, covering key evaluation dimensions including academic impact and overall quality. Extensive experimental results demonstrate its superiority over traditional and LLM-based baselines. Furthermore, we deploy PaperEval in a real-world recommendation system to filter high-quality papers from thousands of daily publications. The system powers social media services with over 8,000 subscribers, and several recommended papers have received over 10,000 views on social platforms, demonstrating the practical evaluation effectiveness of PaperEval. Our code and data are available in the Supplementary Materials.

In summary, our key contributions are as follows:

- We propose PaperEval, a novel LLM-based framework for automated paper evaluation that combines a domain-aware paper retrieval module with a latent reasoning mechanism to enable more accurate assessments.
- We develop a progressive ranking optimization strategy that supervises the LLM reasoning process to iteratively refine its ranking predictions, effectively aligning with the relative ranking objective of paper evaluation.
- PaperEval achieves state-of-the-art performance on two datasets in both academic impact and paper quality evaluation, demonstrating the superiority of PaperEval with progressive ranking optimization.
- We deploy PaperEval in a real-world paper recommendation system, which selects the top 10 high-quality papers each day from thousands of new submissions in fast-evolving research areas.

Related Work

Paper Evaluation. Paper evaluation aims to assess a paper's quality or predict its academic impact. From the quality perspective, a central task is paper rating — predicting whether a paper will be accepted by peer review committees (Lin et al. 2023). Existing methods fall into three main categories. The first leverages neural architectures like CNNs and attention-based models to capture local and global textual interactions (Yang et al. 2018; Deng et al. 2020). The second extracts metadata features and employs traditional models such as random forests (Wang et al. 2024). The third directly encodes the textual content using pretrained models like BERT (Devlin et al. 2019), as in recent work (Xue et al. 2023; Liu et al. 2025a). For evaluating academic impact, prior studies focus on predicting citation counts, citation levels, or other derived impact metrics (Zhao et al. 2024), which can similarly be grouped into three strategies. Metadata-based approaches use handcrafted or extracted features with classical models like MLPs or decision trees (Wang, Yu, and Yu 2011; Qiu and Han 2024; Ruan et al. 2020; Zhang and Wu 2024). Graph-based methods model early citation dynamics using citation graphs and apply graph neural networks for future trend prediction (Yan et al. 2024; He et al. 2023; Li et al. 2023; Jiang, Koch, and Sun 2021). LLM-based approaches either prompt models with a paper's title and abstract to generate citation scores (de Winter 2024), or map dense representations to impact-related metrics (Zhao et al. 2025). Despite these advances, accurate and fine-grained paper evaluation remains challenging. In this work, we explore how integrating retrieval techniques with the reasoning capabilities of LLMs can address this challenge.

Latent Reasoning. Unlike Chain-of-Thought (CoT) reasoning (Wei et al. 2022; Su et al. 2025; Carrow et al. 2025), which relies on explicitly generated intermediate steps, latent reasoning performs inference directly within a model's

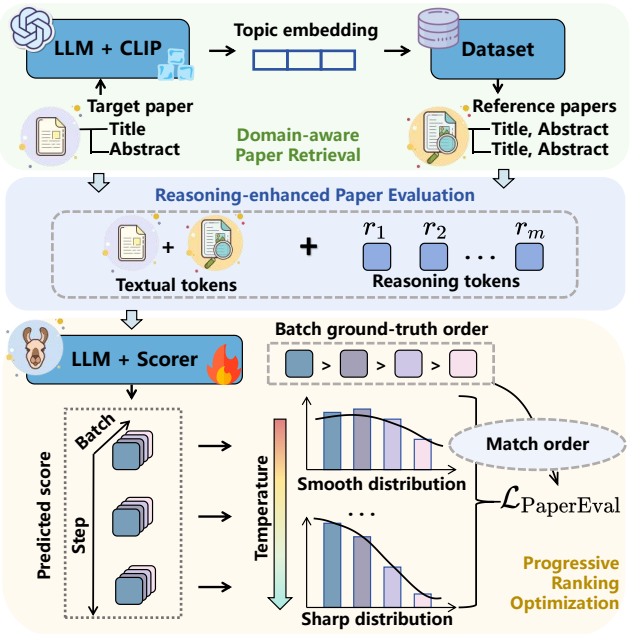

Figure 1: Overview of PaperEval. It comprises two key components: domain-aware paper retrieval and reasoning-enhanced paper evaluation, where the model is fine-tuned using the progressive ranking optimization strategy.

hidden representations (Hao et al. 2024; Biran et al. 2024). This implicit approach has gained momentum in LLM-based recommendation (Tang et al. 2025; Shen et al. 2025a,b; Liu et al. 2025b) and retrieval tasks (Ji et al. 2025), as it avoids the need for annotated reasoning traces while enabling richer, more informative representations. However, a key challenge remains: how to effectively supervise the latent reasoning process. On the one hand, it is essential to ensure that the reasoning process unfolds progressively toward the correct output, allowing the model to refine its prediction step by step (Tang et al. 2025; Liu et al. 2025b; Ji et al. 2025). On the other hand, models must avoid degenerate reasoning, where the hidden states prematurely converge and hinder iterative refinement. For instance, Tang et al. (Tang et al. 2025) introduce a loss that encourages representational diversity across reasoning steps to mitigate this issue. Despite these efforts, ensuring that the model consistently refines its predictions toward accurate outcomes remains challenging. In this work, we propose a progressive optimization strategy that progressively adjusts the softmax temperature and incorporates a ranking-based loss. This design explicitly guides the latent reasoning process toward better alignment with ground-truth evaluation targets.

PaperEval

As shown in Figure 1, **Domain-Aware Paper Retrieval** module selects relevant reference papers, which are jointly encoded with the target paper by **Reasoning-Enhanced Paper Evaluation** through multi-step latent reasoning to produce a quality prediction. To guide the reasoning toward more accurate ranking, we apply a **Progressive Ranking Optimization** strategy.

LLM-based Paper Evaluation

Given a set of N research papers $\mathcal{P} = \{p_i\}_{i=1}^{N}$, where each paper p_i is associated with a ground-truth score s_i reflecting various evaluation aspects (*e.g.,* academic impact and overal quality), LLM-based paper evaluation aims to learn informative paper representations that enable accurate prediction of these scores. Due to the high computational cost of processing full paper bodies, recent methods typically utilize the most representative textual elements (e.g., title and abstract) to construct paper representation w_i, then integrate a lightweight scorer to produce a predicted score $\hat{s}_i$. Formally,

$$w_i = \text{LLM}(p_i)[-1], \quad \hat{s}_i = \text{Scorer}(w_i), \tag{1}$$

where $\text{LLM}(p_i)[-1]$ denotes the final hidden state output by the LLM for paper p_i, and $\text{Scorer}(\cdot)$ is usually implemented as a lightweight MLP.

Domain-aware Paper Retrieval

To equip LLMs with up-to-date domain knowledge for more accurate assessment of the novelty and contributions of each target paper, we introduce a domain-aware retrieval module (depicted at the top of Figure 1). Specifically, given the title and abstract of research papers, we first employ ChatGPT (Achiam et al. 2023) to generate representative topic keyphrases, which are then encoded into topic embeddings using the CLIP text encoder (Radford et al. 2021). To identify relevant work for each target paper p_i, we first compute the cosine similarity between the target topic embedding and those of all other papers in the corpus, where papers with similarity exceeding a predefined threshold γ are retained as candidates. Considering the rapidly evolving nature of many research fields, we further filter these candidates by selecting only concurrent relevant papers, whose publication dates are closest to the target paper p_i. The resulting set, denoted as $\mathcal{R}_i$, contains at most k papers and serves as the domain-aware reference set, providing essential contextual background to help the LLM more accurately assess the target paper's position within the current research landscape.

Reasoning-enhanced Paper Evaluation

To achieve a comprehensive understanding of the motivation and methodological designs of the target paper and effectively incorporate the retrieved domain-aware reference set for contextualized evaluation, PaperEval adopts a latent reasoning mechanism that performs implicit multi-step reasoning for more accurate and reliable evaluation.

Formally, given the target paper p_i and the corresponding domain-aware reference set $\mathcal{R}_i$, we first construct a textual prompt based on the title and abstract of both target and reference papers, and then tokenize it into a sequence of tokens T_i. To stimulate the reasoning capabilities of LLMs (as shown in the middle of Figure 1), PaperEval introduces m reasoning tokens $\{r_1, r_2, \cdots, r_m\}$, which represent intermediate reasoning steps. These tokens guide LLMs to progressively refine the latent states, yielding increasingly in-

formative and discriminative paper representations. The process is formulated as follows:

$$w_i^{(1)}, w_i^{(2)}, \cdots, w_i^{(m)} = \text{LLM}(T_i, r_1, r_2, \cdots, r_m)[-m:], \quad (2)$$

where $w_i^{(j)}$ denotes the intermediate paper representation at the j-th reasoning step. Each representation is then passed through a scorer to obtain a predicted score: $\hat{s}_i^{(j)} = \text{Scorer}(w_i^{(j)})$, resulting in m predictions, all of which are supervised during training, as detailed in the next section.

Progressive Ranking Optimization

Since paper evaluation focuses on identifying the most valuable papers, relative ranking is prioritized over predicting absolute scores. However, learning accurate rankings is challenging, as even small mistakes can lead to substantial shifts in order. To address this, we propose a **progressive ranking optimization** strategy (illustrated at the bottom of Figure 1) that encourages the model to iteratively refine its predictions during multi-step reasoning, gradually improving its ranking accuracy.

Training. Inspired by ListMLE (Xia et al. 2008), which learns to predict rankings by maximizing the likelihood of the ground-truth order, we adapt it to the paper evaluation scenario, encouraging the predicted scores to yield a ranking consistent with the ground-truth.

Given a training batch of B target papers, we first sort the papers in descending order of their ground-truth scores, which serves as the supervision signal to guide the model to focus on relative rankings. At each reasoning step j, the model predicts a batch of scores $\{\hat{s}_i^{(j)}\}_{i=1}^{B}$, which is converted into a score distribution via a softmax function. As the reasoning process deepens, we hope the model predictions are progressively refined, gradually converging toward the optimal relative rankings with increasing confidence. Motivated by this, we introduce progressive temperature annealing into the softmax function to progressively sharpen the score distribution, which increases confidence in the prediction and amplifies the penalty for incorrect rankings, providing stronger supervision. Specifically, we apply a linearly decreasing temperature schedule:

$$\tau^{(j)} = \tau_{\max} + \frac{j}{m}(\tau_{\min} - \tau_{\max}), \quad (3)$$

where $\tau_{\min} < \tau_{\max} \in \mathbb{R}^+$ indicates the upper and lower bounds of temperature, and $\tau^{(j)}$ refers to the annealed temperature for reasoning step j. Therefore, the score distribution at step j is defined as:

$$\hat{f}_i^{(j)} = \frac{\exp(\hat{s}_i^{(j)}/\tau^{(j)})}{\sum_{t=1}^{B} \exp(\hat{s}_t^{(j)}/\tau^{(j)})}. \quad (4)$$

Based on the score distribution at each reasoning step, we adapt the ListMLE loss into the paper evaluation setting, encouraging the model to generate correct relative rankings:

$$\mathcal{L} = -\sum_{j=1}^{m} \log \prod_{i=1}^{B} \frac{\hat{f}_{r(i)}^{(j)}}{\sum_{k=i}^{B} \hat{f}_{r(k)}^{(j)}}, \quad (5)$$

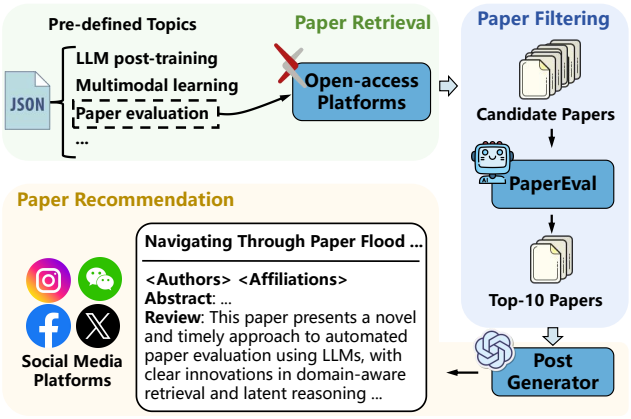

Figure 2: Workflow of the paper recommendation system, which consists of three main phases: paper retrieval, paper filtering, and paper recommendation.

where $r(i)$ denotes the paper ranked at the i-th position in the ground-truth permutation. This loss can be interpreted as the log-probability of sequentially sampling papers according to the ground-truth order without replacement, based on the model's predicted scores. It effectively guides the model to progressively refine the score predictions that better reflect the desired relative rankings.

Inference. After the multi-step reasoning process, the prediction at the final reasoning step $\hat{s}_i^{(m)}$ represents the most informed and discriminative evaluation of the target paper p_i, as it integrates progressively refined judgments across all reasoning steps. Therefore, during inference, we directly adopt $\hat{s}_i^{(m)}$ as the final evaluation score for p_i.

Practical Application of PaperEval

We deploy **PaperEval** as the core filtering module of an automated paper recommendation system, which has been successfully launched on social media. As shown in Figure 2, the system operates through three main phases:

- **Paper retrieval.** The paper recommendation system first selects key topics from pre-defined topics based on certain rules, and retrieves thousands of candidate papers from open-access paper platforms such as arXiv.
- **Paper filtering.** Built upon PaperEval, the system evaluates candidate papers and selects the top 10 most valuable ones based on quality and impact.
- **Paper recommendation.** For the selected papers, the system employs a post generator to synthesize concise reviews and organize key information into easily digestible posts, which are then published on social media to provide followers with timely paper recommendations.

Experiments

In this section, we evaluate our proposed PaperEval on two datasets targeting future impact and overall paper quality. We aim to answer the following research questions:

- **RQ1:** How does PaperEval compare to both LLM-based models and traditional neural baselines?
- **RQ2:** What is the contribution of different components (e.g., paper retrieval, latent reasoning) to the overall performance of PaperEval?
- **RQ3:** How does PaperEval improve its ranking predictions through progressive latent reasoning?

Experimental Settings

Datasets. To evaluate the performance of PaperEval, we conduct experiments on two datasets. The NAID dataset, which is publicly available, provides scores reflecting scientific impact. Furthermore, we construct a new dataset, the ICLR-based dataset, to assess research quality through peer review scores. This dual perspective allows for a comprehensive evaluation of our model's ability to predict both long-term scholarly influence and immediate research quality.

(1) **NAID** (Zhao et al. 2025): This dataset is derived from arXiv. Each paper includes its title, abstract, some metadata (e.g., paper length, number of references), and an impact score that quantifies its relative citation rank within the same domain and publication period, serving as an indicator of the paper's future impact in its field. The dataset contains 11,118 papers in the training set and 1,237 papers in the test set.

(2) **ICLR**: This dataset contains peer review data from ICLR 2021 to 2024 via the OpenReview platform. Each paper includes the title and abstract. To compute paper ratings, we first calculate the average of all review scores, remove scores that deviate by more than 3 points from this average, and then compute the average of the remaining scores as the final scores. This score is normalized to the range $[0, 1]$ and used as the overall quality score. The training set consists of 14,914 papers, and the test set consists of 1657 papers.

The training and test sets are randomly split in a 9:1 ratio. For both datasets, we randomly split 10% training data as the validation dataset during training.

Baselines. We compare PaperEval against various baselines, including both traditional and LLM-based methods:

- **Traditional methods:** 1) **MLP-based** (Ruan et al. 2020) method uses metadata of the target paper as input to an MLP to predict the evaluation score. We do not evaluate this method on the ICLR dataset due to the lack of metadata. 2) **LSTM-based** (Ma et al. 2021) method encodes the abstract of each paper into a sequence representation and applies an LSTM network to predict the target score.
- **LLM-based methods:** 3) **GPT-part** (de Winter 2024) method prompts ChatGPT to predict a paper's score based solely on its title and abstract. We use GPT-4o (Hurst et al. 2024) as the underlying model. 4) **GPT-all** (Lu et al. 2024) approach treats the LLM as a reviewer, prompting it to read the entire paper and generate a full review, including an overall quality score. For cost considerations, we choose GPT-4o-mini as our base model. 5) **SciBERT** (Beltagy, Lo, and Cohan 2019) method is a BERT-based model pretrained on scientific text. We fine-tune it with a simple regression module to perform paper evaluation. 6) **NAIP** (Zhao et al. 2025) method uses LLaMA3-Smaug (Pal et al. 2024) as the backbone LLM. An additional regression module is applied to the output embedding to generate the final score.

Evaluation Metrics. We adopt a variety of evaluation metrics targeting two key aspects: top-K ranking quality and overall ranking consistency. Rankings are computed over all test samples based on their predicted scores.

- **Top-K ranking quality**: Following (Zhao et al. 2025), we use Normalized Discounted Cumulative Gain (NDCG)@$\{10,20\}$ to measure the ranking quality of the top-K recommended papers.
- **Overall ranking consistency**: Following (Ng and Abrecht 2015), we employ Spearman's rho and Kendall's tau to assess how well the predicted rankings align with the ground-truth rankings.

Overall Performance (RQ1)

Table 1 presents a comprehensive comparison between PaperEval and all baseline methods. We summarize our key observations as follows:

- MLP-based models that leverage metadata outperform LSTM baselines on the NAID dataset, highlighting the value of metadata features. However, their lack of semantic understanding limits their overall evaluation capability.
- Pretrained LLMs significantly outperform MLP-based, LSTM-based baselines. This highlights the effectiveness of language model–based semantic understanding in evaluating research papers. Moreover, fine-tuned models (SciBERT, NAIP) achieve better performance than prompting approaches (GPT-part, GPT-all), indicating that task-specific fine-tuning can more effectively enhance a model's capability in evaluating scientific papers. Furthermore, NAIP outperforms the SciBERT-based method, suggesting that large-scale language models possess stronger semantic capabilities for understanding scientific papers, thereby achieving better performance.
- PaperEval consistently achieves **state-of-the-art** performance across all evaluation metrics and datasets. By retrieving domain-relevant references and employing latent reasoning to model complex academic semantics, PaperEval more accurately captures the novelty and quality of target papers, resulting in superior performance in both top-K ranking quality and overall ranking consistency. With the same experimental setup as NAIP, the computational cost remains fully comparable.

In-depth Analysis (RQ2)

In this section, we conduct experiments to further investigate how the designs in PaperEval affect the performance.

Ablation Study. To assess the contribution of each design in PaperEval, we conduct ablations on the NAID dataset: 1) "w/o Ret." removes the paper retrieval module, which means we do not leverage reference papers anymore. 2) "w/o Rea" skips reasoning and directly outputs predictions. 3) "w/o Opt." replaces progressive ranking optimization with Mean Squared Error (MSE) on final scores.

Metrics	NAID				ICLR			
	N@10 ↑	N@20 ↑	Spearman ↑	Kendall ↑	N@10 ↑	N@20 ↑	Spearman ↑	Kendall ↑
MLP-based	0.5109	0.5605	0.0505	0.2868	-	-	-	-
LSTM-based	0.4506	0.4512	-0.0009	0.0904	0.5119	0.5515	0.1355	0.0929
GPT-part	0.5332	0.5258	0.0748	0.0527	0.6600	0.6428	0.0572	0.0417
GPT-all	-	-	-	-	0.6268	0.6365	0.0579	0.0481
SciBERT	0.5784	0.5615	0.0365	0.2709	0.6491	0.6877	0.2114	0.1457
NAIP	0.9274	0.9079	0.4514	0.3163	0.7510	0.7306	0.3188	0.2236
PaperEval (Ours)	**0.9589**	**0.9521**	**0.4953**	**0.3438**	**0.7784**	**0.7386**	**0.3276**	**0.2285**

Table 1: Performance comparison of PaperEval and baseline models on the NAID and ICLR datasets. N@{10,20} denotes NDCG@{10,20}, while Spearman and Kendall represent Spearman's rho and Kendall's tau, respectively. The best performance for each metric is shown in **bold**.

Method	NAID		ICLR	
	N@10	Spearman	N@10	Spearman
PaperEval	0.9589	0.4953	0.7784	0.3276
- w/o Ret.	0.9432	0.4702	0.7235	0.3545
- w/o Rea.	0.9449	0.4968	0.7559	0.3479
- w/o Opt.	0.9585	0.4808	0.7166	0.3198

Table 2: Effect of designs in PaperEval. "Ret." denotes domain-aware paper retrieval method, "Rea." denotes the latent reasoning progress, "Opt." denotes our proposed progressive ranking optimization.

Method	NAID		ICLR	
	N@10	Spearman	N@10	Spearman
List-wise	0.9589	0.4953	0.7784	0.3276
Pair-wise	0.9296	0.4738	0.7350	0.3139
Distribution	0.9265	0.4772	0.7559	0.3071
Regression	0.9585	0.4808	0.7166	0.3198

Table 3: Performance comparison of various loss designs.

From the experimental results shown in Table 2, we observe: 1) The performance decline without domain-aware retrieval underscores the importance of contextual references in enhancing evaluation quality. 2) Latent reasoning brings clear gains in top-k performance by better aligning papers with retrieved papers. However, its tendency to converge quickly (*e.g.,* $m = 8$ converges in 4 epochs, while $m = 12$ takes only 2) combined with the difficulty of supervising the reasoning process, may slightly hurt overall ranking consistency. 3) Removing the optimization strategy significantly reduces performance, since directly regressing dense relevance scores makes it difficult for the model to distinguish subtle differences in paper quality and relative order.

Loss Variants Comparison. To assess the effectiveness of our list-wise ranking loss design, we conduct a series of experiments comparing it with alternative loss designs, all following the same progressive temperature-controlled setting. Specifically, we evaluate the following variants:

- **Pair-wise ranking**: Inspired by RankNet (Burges et al. 2005), we design a temperature-controlled pair-wise ranking loss to examine whether pair-wise supervision is more suitable for PaperEval than list-wise ranking.
- **Distribution similarity**: To investigate whether aligning the predicted and target distributions is the key factor, we generate the ground-truth distribution via a temperature-controlled softmax and compute the KL-divergence between it and the predicted distribution as the loss.
- **Score regression**: As a control setting, similar to (Zhao et al. 2025), we remove the ranking-based objective and directly apply an MSE loss between the final predicted score and the ground-truth label. This setup allows us to examine whether ranking is a more effective learning signal than direct score supervision.

To ensure a fair comparison, we perform hyperparameter tuning (learning rate, number of epochs) for each variant.

From the results summarized in Table 3, we observe the following: 1) List-wise ranking loss outperforms pair-wise ranking loss. This suggests that optimizing over the full ranking list provides a more informative and fine-grained training signal than relying solely on pairwise comparisons. The list-wise objective encourages the model to consider global ranking consistency, which is particularly beneficial in complex evaluation tasks like ours. 2) The distribution similarity loss fails to capture fine-grained relative order. While it encourages the overall score distribution to match the target distribution, it lacks explicit supervision over the relative ranking between individual papers. As a result, its performance falls short, indicating that alignment at the distribution level is insufficient for our goal of precise paper ranking. 3) Unlike ranking-based objectives, score regression focuses on predicting exact values, which often misaligns with identifying the relative ranking of the papers. As our results show, regression consistently underperforms on all metrics, confirming that optimizing for relative order is more suitable and effective.

Hyperparameter Analysis. In this section, we examine the effectiveness of two key hyperparameters on the NAID dataset: the number of retrieved reference papers k and reasoning steps m. The results are shown in Figure 3.

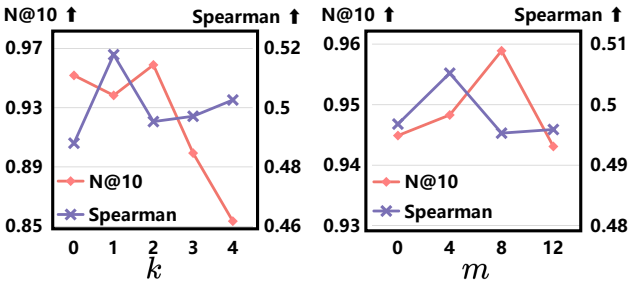

Figure 3: Performance comparison with different hyperparameter settings. We vary the number of reference papers k and the number of reasoning steps m.

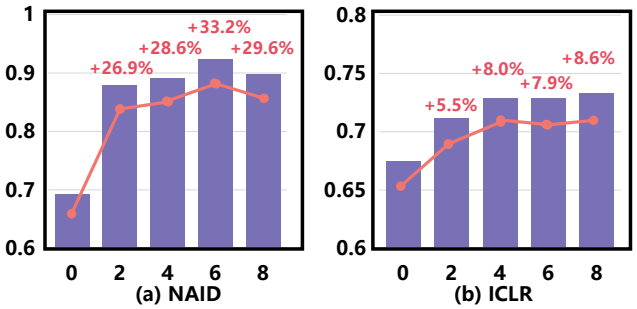

Figure 4: N@10 performance across reasoning steps to evaluate progressive refinement.

- **The number of retrieved reference papers k.** When the number of retrieved reference papers is too small, the LLM lacks sufficient reference context to support accurate evaluation, leading to limited knowledge grounding and poorer ranking performance. Conversely, when too many references are included, the model struggles to effectively capture their relevance to the target paper. This often causes it to lose focus on the target paper itself, impairing top-k identification accuracy. Notably, we observe that NDCG drops more significantly than Spearman in this case, indicating that excessive references particularly affect top-ranked results. Furthermore, a larger k increases input length and computational cost. These findings suggest that selecting a moderate k is essential to balance contextual richness, ranking focus, and efficiency.
- **The number of reasoning steps m.** From the right part of Figure 3, we observe a clear trend where performance first increases and then decreases as the number of reasoning steps m grows. On the one hand, using too few reasoning steps fails to fully leverage the reasoning process: the LLM produces an output without sufficiently analyzing or inferring from the input, leading to suboptimal results. On the other hand, using too many reasoning steps also degrades performance. This is primarily due to the increased risk of overfitting. We observe during training that models with large m tend to converge quickly but subsequently overfit severely. This overfitting can be attributed to the repeated refinement loop in the reasoning process: when the number of steps is excessive, the model repeatedly reprocesses the same input, leading to a kind of memorization or confirmation bias instead of genuine reasoning. Consequently, the model may lose generalizability and begin to reinforce incorrect intermediate conclusions. Similar to the choice of k, it is crucial to select a moderate number of reasoning steps. A well-balanced m encourages the model to reason carefully and refine its predictions while avoiding the pitfalls of excessive iteration.

Effect of Latent Reasoning (RQ3)

To assess whether our progressive ranking optimization effectively encourages step-by-step refinement, we analyze model outputs at different reasoning steps.

We analyze a partially trained model (before full convergence), where step-wise refinement is more visible. After convergence, due to supervision at each step, predictions tend to stabilize, making refinement less apparent. We report NDCG@10 for its sensitivity to ranking quality. "Step-0" denotes the baseline prediction without reasoning, using the final token output directly.

As shown in Figure 4, the model progressively improves its predictions across reasoning steps, validating the effectiveness of our proposed strategy. The NDCG score consistently rises, reflecting improved alignment with the ground-truth ranking. However, in Figure 4(a), a slight drop at the final step suggests that prolonged latent reasoning may lead to diminishing returns or repetitive thinking, pointing to a potential limitation and direction for future work.

Conclusion

In this work, we focus on automatic paper evaluation, which involves assessing specific aspects of research papers to help researchers navigate the growing volume of academic publications. We propose PaperEval, a novel LLM-based framework that combines domain-aware retrieval with latent reasoning to enable more accurate and reliable evaluations. Furthermore, we design a progressive ranking optimization strategy that guides the reasoning process by progressively refining predictions toward more accurate relative rankings. Experimental results demonstrate that our framework achieves state-of-the-art performance in both academic impact and overall quality assessment. Besides, we deploy PaperEval in a real-world paper recommendation system, which has gained notable traction on social media, attracting over 8,000 subscribers and generating more than 10,000 views for several recommended papers.

Despite strong performance, our framework still has limitations, particularly for the latent reasoning module, which opens promising directions for future research. Enhancing latent reasoning for more accurate and insightful paper evaluation requires more effective supervision strategies that are robust to hyperparameter choices, computationally efficient, and resilient to overfitting. In addition, incorporating multimodal information (*e.g.,* figures and tables) from research papers presents another valuable avenue for enhancing the accuracy and depth of paper evaluation.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62572451).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620.

Biran, E.; Gottesman, D.; Yang, S.; Geva, M.; and Globerson, A. 2024. Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14113–14130.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, 89–96.

Carrow, S.; Erwin, K.; Vilenskaia, O.; Ram, P.; Klinger, T.; Khan, N.; Makondo, N.; and Gray, A. G. 2025. Neural reasoning networks: Efficient interpretable neural networks with automatic textual explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15669–15677.

de Winter, J. 2024. Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, 129(4): 2469–2487.

Deng, Z.; Peng, H.; Xia, C.; Li, J.; He, L.; and Yu, P. S. 2020. Hierarchical Bi-Directional Self-Attention Networks for Paper Review Rating Recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6302–6314.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Hao, S.; Sukhbaatar, S.; Su, D.; Li, X.; Hu, Z.; Weston, J.; and Tian, Y. 2024. Training large language models to reason in a continuous latent space. *arXiv:2412.06769*.

He, G.; Xue, Z.; Jiang, Z.; Kang, Y.; Zhao, S.; and Lu, W. 2023. H2CGL: Modeling dynamics of citation network for impact prediction. *Information Processing & Management*, 60(6): 103512.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv:2410.21276*.

Ji, Y.; Xu, Z.; Liu, Z.; Yan, Y.; Yu, S.; Li, Y.; Liu, Z.; Gu, Y.; Yu, G.; and Sun, M. 2025. Learning more effective representations for dense retrieval through deliberate thinking before search. *arXiv:2502.12974*.

Jiang, S.; Koch, B.; and Sun, Y. 2021. HINTS: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. In *Proceedings of the web conference 2021*, 3158–3167.

Lan, Y.; Liu, T.-Y.; Ma, Z.; and Li, H. 2009. Generalization analysis of listwise learning-to-rank algorithms. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 577–584. Association for Computing Machinery. ISBN 9781605585161.

Li, C.; Hong, R.; Xu, X.; Trajcevski, G.; and Zhou, F. 2023. Simplifying temporal heterogeneous network for continuous-time link prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1288–1297.

Lin, J.; Song, J.; Zhou, Z.; Chen, Y.; and Shi, X. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Information fusion*, 98: 101830.

Liu, C.; Zhang, X.; Zhao, H.; Liu, Z.; Xi, X.; and Yu, L. 2025a. LMCBert: An Automatic Academic Paper Rating Model Based on Large Language Models and Contrastive Learning. *IEEE Transactions on Cybernetics*.

Liu, E.; Zheng, B.; Wang, X.; Zhao, W. X.; Wang, J.; Chen, S.; and Wen, J.-R. 2025b. LARES: Latent Reasoning for Sequential Recommendation. *arXiv:2505.16865*.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv:2408.06292*.

Ma, A.; Liu, Y.; Xu, X.; and Dong, T. 2021. A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics*, 126(8): 6803–6823.

Ng, J.-P.; and Abrecht, V. 2015. Better Summarization Evaluation with Word Embeddings for ROUGE. *ArXiv*, abs/1508.06034.

Pal, A.; Karkhanis, D.; Dooley, S.; Roberts, M.; Naidu, S.; and White, C. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv:2402.13228*.

Qiu, J.; and Han, X. 2024. An early evaluation of the long-term influence of academic papers based on machine learning algorithms. *IEEE Access*, 12: 41773–41786.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ruan, X.; Zhu, Y.; Li, J.; and Cheng, Y. 2020. Predicting the citation counts of individual papers via a BP neural network. *J. Informetrics*, 14: 101039.

Shen, X.; Wang, Y.; Shi, X.; Wang, Y.; Zhao, P.; and Gu, J. 2025a. Efficient reasoning with hidden thinking. *arXiv:2501.19201*.

Shen, Z.; Yan, H.; Zhang, L.; Hu, Z.; Du, Y.; and He, Y. 2025b. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv:2502.21074*.

Su, Y.; Chen, Z.; Du, Y.; Ji, Z.; Hu, K.; Bai, J.; and Gao, X. 2025. Explicit Relational Reasoning Network for Scene Text Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7069–7077.

Tang, J.; Dai, S.; Shi, T.; Xu, J.; Chen, X.; Chen, W.; Jian, W.; and Jiang, Y. 2025. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. *arXiv:2503.22675*.

Vergoulis, T.; Kanellos, I.; Giannopoulos, G.; and Dalamagas, T. 2020. Simplifying impact prediction for scientific articles. *arXiv:2012.15192*.

Wang, M.; Yu, G.; and Yu, D. 2011. Mining typical features for highly cited papers. *Scientometrics*, 87(3): 695–706.

Wang, Z.; Zhang, H.; Chen, H.; Feng, Y.; and Ding, J. 2024. Content-based quality evaluation of scientific papers using coarse feature and knowledge entity network. *Journal of King Saud University-Computer and Information Sciences*, 36(6): 102119.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Weng, Y.; Zhu, M.; Xia, F.; Li, B.; He, S.; Liu, S.; Sun, B.; Liu, K.; and Zhao, J. 2023. Large Language Models are Better Reasoners with Self-Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2550–2575.

Xia, F.; Liu, T.-Y.; Wang, J.; Zhang, W.; and Li, H. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, 1192–1199.

Xia, W.; Li, T.; and Li, C. 2022. A review of scientific impact prediction: tasks, features and methods. *Scientometrics*, 128(1): 543–585.

Xue, Z.; He, G.; Liu, J.; Jiang, Z.; Zhao, S.; and Lu, W. 2023. Re-examining lexical and semantic attention: Dual-view graph convolutions enhanced BERT for academic paper rating. *Information Processing & Management*, 60(2): 103216.

Yan, P.; Kang, Y.; Jiang, Z.; Song, K.; Lin, T.; Sun, C.; and Liu, X. 2024. Modeling scholarly collaboration and temporal dynamics in citation networks for impact prediction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2522–2526.

Yang, P.; Sun, X.; Li, W.; and Ma, S. 2018. Automatic academic paper rating based on modularized hierarchical convolutional neural network. *arXiv:1805.03977*.

Zhang, F.; and Wu, S. 2024. Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics*, 129(7): 4137–4166.

Zhao, P.; Xing, Q.; Dou, K.; Tian, J.; Tai, Y.; Yang, J.; Cheng, M.-M.; and Li, X. 2025. From Words to Worth: Newborn Article Impact Prediction with LLM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1183–1191.

Zhao, P.; Zhang, X.; Cao, J.; Cheng, M.-M.; Yang, J.; and Li, X. 2024. A literature review of literature reviews in pattern analysis and machine intelligence. *arXiv:2402.12928*.