Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis

Jianxiang Yu[♡*], Zichen Ding[♡*], Jiaqi Tan[♡], Kangyang Luo[♡], Zhenmin Weng[♡], Chenghua Gong[♡], Long Zeng[♡], Renjing Cui[♡], Chengcheng Han[♡], Qiushi Sun[♢], Zhiyong Wu[♢], Yunshi Lan[♡], Xiang Li[♡†]

[♡] East China Normal University, Shanghai, China

[♢] Shanghai AI Laboratory, Shanghai, China

sea.ecnu@gmail.com

https://ecnu-sea.github.io/

Abstract

In recent years, the rapid increase in scientific papers has overwhelmed traditional review mechanisms, resulting in varying quality of publications. Although existing methods have explored the capabilities of Large Language Models (LLMs) for automated scientific reviewing, their generated contents are often generic or partial. To address the issues above, we introduce an automated paper reviewing framework SEA. It comprises of three modules: Standardization, Evaluation, and Analysis, which are represented by models SEA-S, SEA-E, and SEA-A, respectively. Initially, SEA-S distills data standardization capabilities of GPT-4 for integrating multiple reviews for a paper. Then, SEA-E utilizes standardized data for fine-tuning, enabling it to generate constructive reviews. Finally, SEA-A introduces a new evaluation metric called mismatch score to assess the consistency between paper contents and reviews. Moreover, we design a self-correction strategy to enhance the consistency. Extensive experimental results on datasets collected from eight venues show that SEA can generate valuable insights for authors to improve their papers.

1 Introduction

With the rapid pace of scientific advancement, there has been a significant increase in the volume of research publications (Bornmann and Mutz, 2015; Gao et al., 2024; Lin et al., 2023a). Nevertheless, it poses considerable challenges for traditional scientific feedback mechanisms (Liang et al., 2023). On one hand, it exacerbates the pressure on the peer review process (Lee et al., 2013; Björk and Solomon, 2013); on the other hand, the disparate quality of these numerous publications can negatively affect the scientific research milieu (Kelly et al., 2014; Liu and Shah, 2023). Consequently,

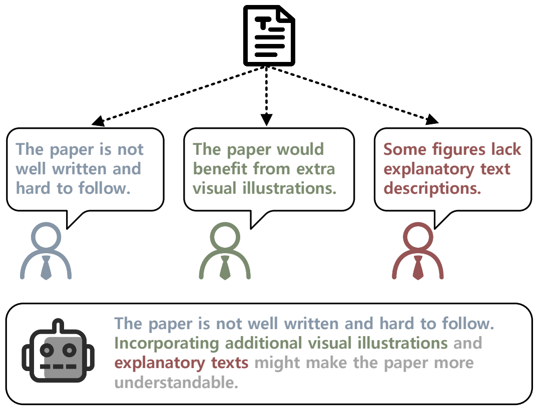

Figure 1: Multiple reviews of a paper often provide helpful but partial opinions on certain aspects. Integrating these reviews can offer more comprehensive feedback on the paper.

there is a need for an automated scientific reviewing framework designed to generate constructive reviews with strong evidence supports to help authors improve the caliber of their works (Yuan et al., 2022).

However, the task of delivering timely, thorough, and perceptive feedback on research papers is inherently intricate and cognitively demanding (Horbach and Halffman, 2018). Traditional language models typically struggle to handle such lengthy texts, let alone provide valuable review insights (Cohan et al., 2020; Wang et al., 2020). Fortunately, Large Language Models (LLMs) have demonstrated emergent capabilities (Wei et al., 2022), which have shown state-of-the-art performance in a wide range of tasks (Brown et al., 2020; Touvron et al., 2023; Tan et al., 2024; Sun et al., 2024). Further, they have also been strengthened to handle increasingly longer contexts (Jiang et al., 2023), facilitating the possibility for automated reviewing (Liang et al., 2023; Gao et al., 2024).

Currently, some efforts have been made to ex-

[*] Equal Contribution
[†] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10164–10184
November 12-16, 2024 ©2024 Association for Computational Linguistics

plore the capabilities of LLMs for automated paper reviewing. For example, Liu and Shah (2023) and Liang et al. (2023) investigate the potential reliability and credibility of paper reviews generated by LLMs with specially designed prompts. Yet most of these LLMs are tailored for broad and general-purpose applications (Wei et al., 2023), so simply prompting LLMs in reviewing papers could output generic comments of less value (Liang et al., 2023). Further, certain studies have developed peer review datasets and fine-tuned LLMs to learn the paradigm of paper reviewing (Wei et al., 2023; Gao et al., 2024). However, in the supervised fine-tuning (SFT) process, these methods simply utilize a review for a paper that can be biased, partial (see Figure 1) and often formalized in various formats and criteria, which could hinder the potential of LLMs for automated paper reviewing (Lin et al., 2023b; Gao et al., 2024). Also, they lack a self-correction mechanism when the generated reviews are less appealing.

To tackle the issues, in this paper, we propose a novel automated paper reviewing framework, namely, **SEA**, which consists of three modules: **Standardization**, **Evaluation**, and **Analysis**, as shown in Fig. 2. We next summarize the details of each module.

In the **Standardization** module, we develop a model SEA-S, which aims to standardize reviews. Specifically, we first utilize GPT-4 to integrate multiple reviews of a paper into one that is in a unified format and criterion with constructive contents, and form an instruction dataset for SFT. After that, we fine-tune an open-source LLM Mistral-7B to distill the knowledge of GPT-4.

In the **Evaluation** module, we fine-tune another Mistral-7B to derive the SEA-E model, which can comprehensively analyze papers and generate high-quality reviews. Given papers that are in PDF format, we parse them into text and LaTeX codes, and input their corresponding multiple reviews into SEA-S to generate standardized reviews. The parsed papers, standardized reviews and human-crafted prompts constitute another instruction dataset for SFT, leading to SEA-E.

In the **Analysis** module, we further introduce a self-correction strategy that promotes SEA to rethink and regenerate more constructive reviews, when the generated reviews are inconsistent with the parsed papers. To measure the inconsistency, we put forward a metric, namely, *mismatch score*. We also train a regression model SEA-A to estimate scores for the generated reviews.

Extensive experiments on eight diverse datasets show that the reviews generated by the SEA framework significantly outperform existing methods in terms of quality, comprehensiveness, and consistency. To sum up, we highlight our contributions as follows:

- We propose a novel framework SEA for automated paper reviewing.
- We present an effective model SEA-S for standardizing reviews from various academic venues in different formats and criteria.
- We devise a self-correction strategy to improve the consistency between papers and reviews.
- We conduct extensive experiments to show the superiority of SEA over other competitors.

Finally, it is important to emphasize that the purpose of this paper is not to directly recommend the acceptance/rejection on papers. We anticipate our framework SEA can facilitate timely feedback for researchers, thereby enhancing the quality of their work and enabling them to transition efficiently to subsequent projects.

2 Related Works

2.1 Long-context Large Language Models

LLMs have recently achieved substantial progress in accommodating lengthy contexts. For example, LongLLaMA (Tworkowski et al., 2024) and LongLoRA (Chen et al., 2023b) support long contexts processing by modifying the attention mechanism. There are also some positional encoding methods proposed, including ALiBi (Press et al., 2021), xPOS (Sun et al., 2022) and RoPE variants (Chen et al., 2023a; Xiong et al., 2023).

Assessing the capability of LLMs in handling long contexts has also attracted significant attention. The needle-in-a-Haystack (NIAH) test (Kamradt, 2023) has been widely adopted to evaluate long-context LLMs. Further, RULER (Hsieh et al., 2024) extends the vanilla NIAH test to provide a more thorough assessment. Based on the RULER evaluation results, we select Mistral-7B (Jiang et al., 2023) as the base model in our paper. Mistral-7B is a compact LLM that has been shown to handle at least 16K tokens, sufficient to meet the input requirements of most academic papers.

2.2 Automated Scientific Reviewing

Automating scientific reviewing began its investigation in the era of small language models. The early

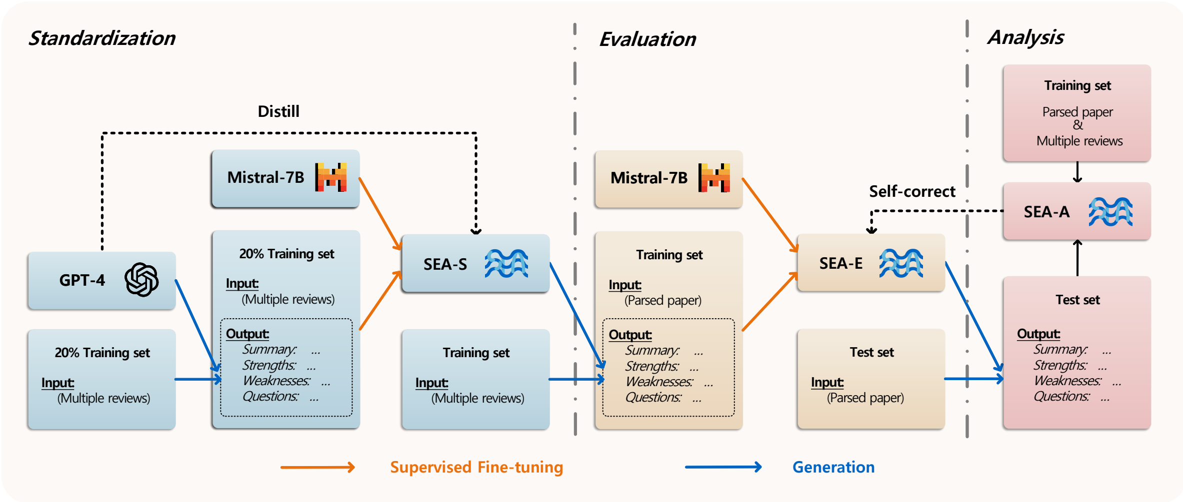

Figure 2: The overall framework of SEA consists of three modules: Standardization, Evaluation and Analysis.

work (Zhang et al., 2022) utilizes RoBERTa (Liu et al., 2019) to assess the textual fluency of papers and fairness disparity in peer review. In peer grading, Morris et al. (2023) fine-tune distilBERT (Sanh et al., 2019) using course grading data from massive open online courses to examine the reliability of peer grading scores. However, due to the restricted capability of language models in handling lengthy contexts, automating scientific reviewing of a full paper has not been studied before the advent of LLMs.

Recently, since LLMs exhibit advancements in various NLP tasks, some studies are exploring the capabilities of LLMs in automated paper reviewing. For example, Liu and Shah (2023) and Liang et al. (2023) customize prompts to guide GPT-4 in generating scientific feedbacks. Wei et al. (2023) conduct continuous training of LLaMA2-70B (Touvron et al., 2023) on academic data, resulting in an academically enhanced model AcademicGPT. Further, Gao et al. (2024) collect a large-scale peer review dataset, and propose a two-stage review generation framework REVIEWER2 with question-guided prompts.

3 SEA

This section details three major modules (i.e., Standardization, Evaluation and Analysis) of SEA, and the overall framework is illustrated in Figure 2.

3.1 SEA-S: Standardization

To explore the potential of LLMs in automated scientific reviewing, a high-quality labeled dataset is generally needed for supervised fine-tuning (SFT). This process feeds LLMs with more peer reviews, thereby enhancing the quality of its generated ones. However, in the peer review datasets, each paper is often associated with multiple peer reviews, with each review offering a limited perspective based on the reviewer's field and expertise. On the other hand, the review formats and criteria could vary across different academic venues, and directly performing SFT on existing peer review datasets can lead to inconsistencies. Therefore, we first have to standardize reviews in a unified format and criterion with comprehensive contents before SFT. For each paper, we integrate all the reviews into one, which can eliminate redundancy and error in multiple reviews. The integrated review is expected to focus on the major advantages and disadvantages of the paper, thereby enhancing its quality.

To perform data standardization, we attempt several representative open-source and closed-source models, such as Mistral-7B, GPT-3.5 and GPT-4. We empirically observe that Mistral-7B and GPT-3.5 tend to simply concatenate the original contents. In contrast, GPT-4 leads them by integrating reviews in an unified format and providing detailed evidence for each argument (The comparative examples are given in Figure 6 of Appendix A.1). However, the API for GPT-4 is costly and inflexible. Inspired by Alpaca (Taori et al., 2023), we distill GPT-4's excellent data standardization capabilities into open-source models.

Specifically, we first randomly select 20% of the papers from the training set along with their reviews $\{[r_{i1}^{\text{origin}}, r_{i2}^{\text{origin}}, \ldots, r_{im}^{\text{origin}}]\}_{i=1}^{n}$, where n is the number of selected papers and m is the number of reviews corresponding to paper p_i. Next, for each paper p_i, we input all its reviews along with the customized instruction $inst_s$ into GPT-4, which in turn yields the standardized review $r_i^{\text{GPT-4}}$. In this way, we can construct the *instruction dataset* for the data standardization model SEA-S that takes Mistral-7B as the base model. Formally, the triplet in the dataset is $<inst_s, [r_{i1}^{\text{origin}}, r_{i2}^{\text{origin}}, \ldots, r_{im}^{\text{origin}}], r_i^{\text{GPT-4}}>$, which is further served for SFT. After fine-tuning SEA-S, we feed all the reviews in the training set into SEA-S for data standardization, which outputs the integrated reviews $\{r_i^{\text{SEA-S}}\}_{i=1}^{N}$. Here, N denotes the number of papers in the training set. In summary, SEA-S provides a novel paradigm for integrating peer review data in an unified format across various conferences.

3.2 SEA-E: Evaluation

In the Evaluation module, we aim to construct a talented LLM that can deeply understand papers and generate constructive reviews. Notably, since raw crawled papers are in PDF format, we first apply Nougat (Blecher et al., 2023) as the parser, which is a model based on Visual Transformer and is specially designed for parsing academic documents. In particular, Nougat can parse formulas into LaTeX codes instead of corrupted text encoding, enabling LLMs to gain a deeper understanding of papers' contents. Further, due to the long-text characteristic of papers, we choose the open-source model Mistral-7B as the backbone model, which has demonstrated its ability in effectively handling up to 16K tokens for the long-context benchmark RULER (Hsieh et al., 2024).

Based on the outputs of the SEA-S model, we next construct the *instruction dataset* for the evaluation model SEA-E. Each triplet in the dataset is denoted as $<inst_e, \hat{p}_i, r_i^{\text{SEA-S}}>$, where $inst_e$ is the specially designed instruction for evaluation, $\hat{p}_i$ is the parsed paper, and $r_i^{\text{SEA-S}}$ is the standardized review. Note that $r_i^{\text{SEA-S}}$ contains solid evidence for each argument in the review. This endows SEA-E with the capability to generate comprehensive and constructive reviews after SFT.

3.3 SEA-A: Analysis

Now, we step into the Analysis module, where a *mismatch score* is proposed to measure the consistency between papers and their generated reviews. Given a paper p with m raw reviews, let us denote its ground-truth *paper ratings* as $S_p = \{s_{pr_1}, s_{pr_2}, \ldots, s_{pr_m}\}$ and *confidence scores* as $C_p = \{c_{pr_1}, c_{pr_2}, \ldots, c_{pr_m}\}$, where each s_{pr_i} and c_{pr_i} indicate the rating and confidence score given by the i-th reviewer. We next use the confidence scores as weights and calculate the weighted average rating of paper p, which is further subtracted from the reviewer's rating to serve as the ground truth mismatch score. Formally, we have:

$$y_{true}^{pr_i} = s_{pr_i} - \frac{\sum_{j=1}^{m} c_{pr_j} * s_{pr_j}}{\sum_{j=1}^{m} c_{pr_j}}. \tag{1}$$

From the equation, we see that, when a reviewer's rating is greater than the weighted average, the review may tend to emphasize the paper's strengths; otherwise, the review may be preferably critical of the paper. Generally, the greater the difference, the lower the review quality. When $y_{true}^{pr_i} = 0$, we consider the review to be relatively neutral and consistent with the paper content. For example, when the review ratings of a paper are {2, 6, 6, 6} and all are given with full confidence, the quality of the review rated 2 is considered to be lower because it deviates significantly from the weighted average rating of 5.

To estimate the mismatch score, we train a lightweight regression model SEA-A. Specifically, each parsed paper $\hat{p}$ and its corresponding review r generated from SEA-E form a pair $<\hat{p}, r>$, which serves as the input. We first utilize the pre-trained sentence representation model SFR-Embedding-Mistral (Rui Meng, 2024) that is designed for long contexts to transform the texts of papers and reviews into representations $h_{\hat{p}}$ and h_r, respectively. Then, we compute the *query* and *key* vectors for both the paper and the review separately:

$$\begin{aligned} q_{\hat{p}} &= W^q h_{\hat{p}}, \quad q_r = W^q h_r, \\ k_{\hat{p}} &= W^k h_{\hat{p}}, \quad k_r = W^k h_r. \end{aligned} \tag{2}$$

Here, W^q and W^k are learnable weight matrices. Based on the query and key vectors, we calculate the estimated mismatch score y_{pred}^{pr} by:

$$y_{pred}^{pr} = w(q_{\hat{p}} {k_r}^T + q_r {k_{\hat{p}}}^T) + b. \tag{3}$$

Finally, we use the mismatch score y_{true}^{pr} as the ground truth and the Mean Squared Error (MSE) loss as the objective to train the regression model SEA-A. The smaller the absolute value of the mismatch score, the higher the consistency between the review and the paper.

Table 1: Dataset Statistics

	CONLL-16	ACL-17	COLING-20	ARR-22	NeurIPS-16-22	ICLR-17-23	NeurIPS-23	ICLR-24	Total
# papers	22	136	88	364	1,048	1,617	3,368	5,653	12,296
# tokens per paper	8,163	8,400	7,571	8,229	10,499	9,586	11,205	9,815	10,142
# reviews	39	272	112	684	3,847	5,779	15,027	21,839	47,602
# tokens per review	532	558	539	539	527	602	642	594	603
% accepted	50%	67%	93%	100%	97%	30%	95%	37%	60%
domain	NLP/CL	NLP/CL	NLP/CL	NLP/CL	ML	ML	ML	ML	multi

Table 2: The overall performance (%) on four **cross-domain** datasets: CONLL-16, ACL-17, COLING-20, ARR-22, and four **in-domain** datasets: NeurIPS-16-22, ICLR-17-22, NeurIPS-23, ICLR-24. We highlight the best score on each dataset in **bold** and the runner-up score with an underline.

Method	BLEU	ROUGE (Recall)			ROUGE (F1-score)			BERTScore	Tokens
		R-1	R-2	R-L	R-1	R-2	R-L		
CONLL-16									
M-7B	18.92	20.81	4.81	10.30	28.66	6.81	14.18	82.49	554
M-7B-R	18.16	21.96	5.17	10.62	29.56	7.18	14.31	82.57	357
M-7B-3.5	19.70	26.51	5.58	13.96	30.19	6.45	15.37	82.01	627
SEA-E	29.07	34.91	7.79	15.29	38.64	8.67	16.73	82.85	793
SEA-EA	**31.01**	**36.96**	**8.91**	**16.34**	**40.49**	**9.68**	**17.57**	**82.94**	798
ACL-17									
M-7B	18.92	21.53	5.23	10.50	27.99	6.93	13.54	82.75	569
M-7B-R	18.15	21.84	5.19	10.76	27.71	6.87	13.55	82.56	357
M-7B-3.5	16.73	27.27	6.26	14.47	26.09	6.19	13.19	82.37	636
SEA-E	25.67	33.13	7.71	14.94	35.52	8.45	15.62	83.08	772
SEA-EA	**27.90**	**35.83**	**8.84**	**15.83**	**38.03**	**9.48**	**16.36**	**83.19**	806
COLING-20									
M-7B	21.97	29.11	6.42	14.80	31.91	7.01	15.83	82.76	579
M-7B-R	19.49	29.21	6.69	15.20	30.23	6.80	15.25	82.27	361
M-7B-3.5	18.13	34.03	7.56	18.43	28.49	6.10	14.77	82.12	617
SEA-E	22.93	40.62	9.23	20.05	34.37	7.65	16.15	82.85	774
SEA-EA	**24.85**	**42.97**	**10.57**	**20.89**	**36.67**	**8.76**	**16.96**	**83.09**	782
ARR-22									
M-7B	22.07	25.28	6.96	12.46	32.60	9.16	15.99	83.25	575
M-7B-R	20.27	24.89	6.70	12.60	31.22	8.66	15.71	82.70	357
M-7B-3.5	20.18	31.70	7.90	16.38	30.82	7.86	15.33	82.65	650
SEA-E	27.92	37.64	9.37	17.18	38.94	9.84	17.35	83.38	787
SEA-EA	**30.05**	**40.34**	**10.82**	**18.17**	**41.37**	**11.19**	**18.20**	**83.59**	818

Method	BLEU	ROUGE (Recall)			ROUGE (F1-score)			BERTScore	Tokens
		R-1	R-2	R-L	R-1	R-2	R-L		
NeurIPS-16-22									
M-7B	14.91	14.47	4.89	7.15	23.31	7.94	11.56	83.10	612
M-7B-R	13.94	14.47	4.79	7.29	22.70	7.67	11.44	82.73	362
M-7B-3.5	16.95	20.41	6.02	10.72	26.45	8.13	13.45	82.56	629
SEA-E	24.83	24.12	7.31	10.66	34.06	10.44	15.11	83.35	782
SEA-EA	**27.08**	**26.76**	**8.38**	**11.55**	**36.91**	**11.69**	**15.99**	**83.52**	838
ICLR-17-23									
M-7B	13.75	13.10	4.42	6.51	21.65	7.36	10.80	83.26	607
M-7B-R	12.98	13.38	4.45	6.85	21.36	7.26	10.91	82.80	359
M-7B-3.5	17.85	18.26	5.70	9.27	27.37	8.69	13.94	82.87	637
SEA-E	23.34	22.38	6.84	9.93	32.50	10.07	14.49	83.58	783
SEA-EA	**25.47**	**24.80**	**7.87**	**10.81**	**35.23**	**11.32**	**15.43**	**83.73**	841
NeurIPS-23									
M-7B	12.42	11.96	4.96	6.13	20.55	8.55	10.55	83.86	617
M-7B-R	11.92	11.88	4.87	6.16	20.14	8.31	10.49	83.44	366
M-7B-3.5	16.71	16.80	6.12	8.53	26.51	9.74	13.50	83.20	650
SEA-E	21.34	20.32	7.27	9.14	31.34	11.26	14.14	84.02	794
SEA-EA	**23.32**	**22.49**	**8.38**	**9.91**	**34.03**	**12.73**	**15.03**	**84.20**	844
ICLR-24									
M-7B	13.93	13.48	5.29	6.73	22.55	8.89	11.28	83.79	614
M-7B-R	13.91	14.17	5.41	7.21	22.94	8.85	11.69	83.81	380
M-7B-3.5	18.72	19.40	6.52	9.64	29.26	9.93	14.58	83.29	649
SEA-E	23.88	23.28	7.90	10.13	34.29	11.71	14.98	84.04	793
SEA-EA	**25.96**	**25.62**	**8.97**	**10.97**	**36.97**	**13.02**	**15.88**	**84.15**	852

After SEA-A is trained, we further introduce a *self-correction strategy* to analyze each review generated by SEA-E. When the estimated mismatch score y^{pr}_{pred} is larger than a pre-set threshold θ, we regenerate the review by adding the current mismatch score as additional prompt to ensure the consistency between the paper and the review.

4 Experiments

4.1 Experimental Details

Datasets. We crawl the latest papers and their corresponding reviews from OpenReview [1], including NeurIPS-2023 and ICLR-2024. We randomly sample 90% of the data according to the distribution of "Rating" to serve as the training set, with the remaining 10% used as the test set for evaluation. Our test set also includes subsets from REVIEWER2 (Gao et al., 2024) for NeurIPS (2016-2022) and ICLR (2017-2023). Additionally, we conduct evaluations on cross-domain datasets from Natural Language Processing (NLP) and Computational Linguistics (CL) fields, incorporating data from PeerRead (Kang et al., 2018) for CONLL-2016 and ACL-2017, and from NLPeer (Dycke et al., 2022) for COLING-2020 and ARR-2022. All the datasets include the original PDF files of the papers and structurally formatted reviews. Different review data exhibits format difference across various conferences and years. The statistics of our datasets are summarized in Table 1.

Setup. We use *Mistral-7B-Instruct-v0.2* (Jiang et al., 2023) with a context length of 32k as our backbone model. In the Evaluation module, the reviews that our methods generate consists of three parts: a textual part with "Summary", "Strengths", "Weaknesses", and "Questions"; a quantitative part that includes "Soundness", "Presentation", "Contribution", and "Rating"; and finally, the paper decision (Accept/Reject) with corresponding reasons. In the Analysis module, we utilize 80% of the entire training set for training and the remaining 20% for validation. We set the threshold θ to the average mismatch score in the validation set. In our framework, there are two methods for generating reviews: **SEA-E** and **SEA-EA**, where SEA-EA is an en-

[1] https://openreview.net/

hanced model that combines the Analysis module with SEA-E. For SEA-EA, if the mismatch score between generated reviews and papers surpasses θ, this score will be incorporated into the prompts to improve the quality of generated reviews. Moreover, if the mismatch score consistently exceeds θ across 10 successive trials, the generation process will be terminated. The review with the smallest score will be selected as the final output.

Baselines. We compare the following baseline methods, which are divided into two categories: (1) Direct inference with LLMs: We directly use **M**istral-**7B** (**M-7B**) for inference, guided by $inst_e$ to generate reviews in the specified format. (2) SFT methods: From all reviews for each paper in the training set, we randomly select one review as the output for SFT, referred to as **M**istral-**7B**-**R**andom (**M-7B-R**). **M**istral-**7B**-GPT-**3.5** (**M-7B-3.5**) refers to the method where reviews for each paper are standardized using *gpt-3.5-turbo*, and these standardized outputs are then applied in the SFT stage. Moreover, REVIEWER2 (Gao et al., 2024) is a two-stage review generation framework. Due to time-consuming, we use a smaller test set and compare it with REVIEWER2. The detailed experimental results are provided in the Appendix B.

We unify the instruction $inst_e$ and input $\hat{p}$ across all the baseline methods and our framework. Here, $inst_e$ is the instruction for SEA-E, and $\hat{p}$ represents the parsed paper. Detailed information about $inst_e$ can be found in Table 10 in Appendix A.1.

4.2 Main Results

We use BLEU (Papineni et al., 2002), ROUGE (Recall), ROUGE (F1-score) (Lin, 2004), and BERTScore (Zhang et al., 2019) as metrics to evaluate the quality of generated reviews across eight datasets. Specifically, BLEU and ROUGE measure the similarity between papers and reviews based on n-grams, while BERTScore focuses on semantic similarity in the embedding space. For the ROUGE metric, recall measures how comprehensively the generated reviews capture the key information from raw papers, while the F1 score assesses the balance between precision and recall in the generated contents. To measure the completeness and comprehensiveness of the generated reviews, we simply concatenate all the reviews of each paper to serve as a benchmark for evaluation. Moreover, we have also counted the average number of tokens in the generated reviews.

The results in Table 2 show that SEA outperforms other baseline models across all the testing scenarios, with particularly notable gains on the ROUGE (Recall) metric. This confirms that our proposed framework SEA is capable of generating comprehensive and constructive reviews. Further, SEA not only performs excellently on in-domain tasks but also shows strong performance on cross-domain datasets, demonstrating its robust generalizability. It is also worth noting that SEA-EA surpasses SEA-E in all cases, underscoring the effectiveness of the self-correction strategy in generating well-grounded reviews consistent with raw papers. However, for M-7B-R, we notice that randomly selecting a review as the output of SFT often leads to shorter texts. To some extent, the quality of a review is positively correlated with its length, which explains its poor performance. Although directly inferring with M-7B can generate longer text, it fails to align with human reviews, resulting in lower evaluation scores. For M-7B-3.5, its performance is poorer than SEA-E, which further indicates the effectiveness of SEA-S. Consequently, using high-quality standardized data generated by SEA-S can effectively improve the performance of SFT. In Appendix A.2 we give concrete examples of reviews generated by different models.

4.3 Comparison of Standardized Results

We show the standardized results on papers in the training set of NeurIPS-2023 and ICLR-2024 that have different rating criteria. In addition, reviews are organized in various formats.

Content analysis. We first compare SEA-S with Mistral-7B, GPT-3.5, and GPT-4 to evaluate their review standardization performance. All the models are fed with the same inputs, including the instruction $inst_s$ and multiple reviews. Since there is no ground-truth text for this standardized task, we utilize reviews generated by SEA-S as *references*, while reviews generated by other models serve as *candidates*. Next, we calculate *recall* and *precision* values of ROUGE for candidates compared to references. Based on the content intersection of reference and candidate, recall and precision refer to the percentage of intersection in reference and candidate, respectively. From the two metrics, we can deduce the percentages of overlapping and exclusive semantic information in both reviews, whose results are shown in Figure 3. We compare the model performance w.r.t. different ROUGE metrics, including ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL). The light blue area in the figure indicates the overlapping contents, while the dark blue and light grey areas represent the exclusive contents by SEA-S (reference) and other models (candidate), respectively.

From the figure, we see that, SEA-S can generate a significantly larger percentage of exclusive contents than both Mistral-7B and GPT-3.5. This further verifies that SEA-S can better standardize reviews with richer information. We also surprisingly observe that SEA-S can output slightly more exclusive contents in standardized reviews than GPT-4. The reason could be that the instruction dataset for SFT in SEA-S is derived from GPT-4. Considering the high cost of GPT-4, this demonstrates the effectiveness of small models for review standardization. On the other hand, recap that the difference between M-7B-3.5 and SEA-E only lies in the data standardization step. The advantage of SEA-E over M-7B-3.5 in Table 2 shows that SEA-S has better data standardization capability.

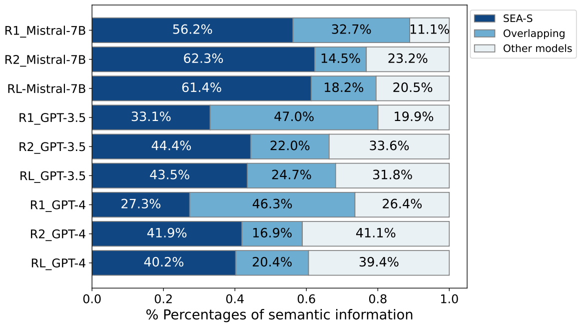

Figure 3: Content analysis results.

Format analysis. Standardized data formats can help LLMs better understand the correspondence between the instruction and generated content during SFT. To perform format analysis, we utilize regular expression matching based on instruction formats to calculate the proportion of correctly formatted reviews integrated by different models. The results given in Figure 4 demonstrate that SEA-S is capable of generating 100% correctly formatted data. In contrast, Mistral-7B and GPT-3.5 show poor performance, particularly the former, which generates a large amount of data that does not meet the format requirements. Also, we observe that around 10% of the data integrated by GPT-4 does not fully comply with the instruction. Compared to GPT-4, SEA-S benefits from SFT and thus shows superior instruction adherence. Overall, SEA-S demonstrates excellent effectiveness in handling reviews of various formats and criteria. Details of the instruction for standardizing reviews and specific examples are given in Appendix A.1.

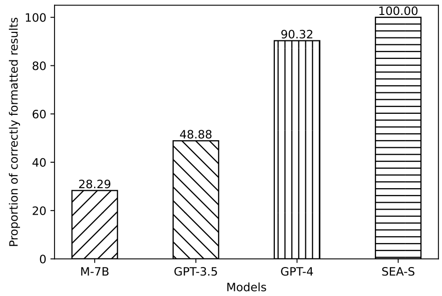

Figure 4: Format analysis of different models.

4.4 Mismatch Score in SEA-A

To analyse the consistency between the reviews generated by different models and the corresponding papers, we input the reviews and their respective papers in the test set into the trained SEA-A model to calculate the average mismatch score for each model across different datasets. As illustrated in Figure 5, SEA-EA, due to its self-correction strategy, consistently outperforms others across all the datasets. Further, SEA-E is the runner-up method. This verifies that the reviews generated by both methods have a higher consistency with their corresponding papers. Mistral-7B, which has not undergone fine-tuning, fails to learn the correspondence between papers and reviews, resulting in higher mismatch scores. Although M-7B-R and M-7B-3.5 are fine-tuned, they are still worse than our methods. This can be explained by the insufficient model standardization capability.

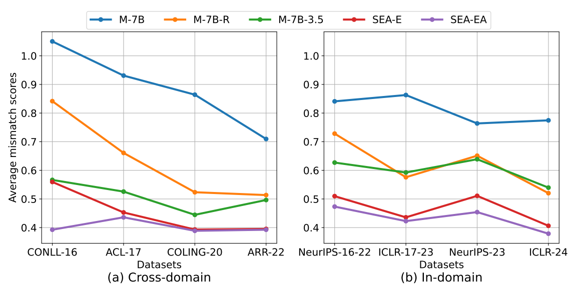

Figure 5: The performance of different models on mismatch scores across various datasets.

To further study mismatch score, for each paper, we randomly select a review from other papers in the test set as the "negative" review. The negative review is expected to derive a larger mismatch score than the generated review, which is empirically observed from our results given in Appendix A.3. This again confirms that our regression model is capable of quantitatively assessing the consistency across reviews and papers.

4.5 Quantitative Score Analysis

We conduct a further quantitative analysis on the four scores in the generated reviews on two datasets with actual scores, NeurlPS-2023 and ICLR-2024. The four scores include "Soundness", "Presentation", and "Contribution", which are integers from $[1, 4]$, and "Rating", which is an integer from $[1, 10]$. The rating criterion is given in the instruction of SEA-E in Table 11. In practice, each paper has multiple reviews and each review has the above four scores. Therefore, given a paper, for each score, we use the "Confidence" score in each review as the weight and calculate the weighted average as the reference score.

Table 3: Quantitative Score Analysis.

	Method	Soundness	Presentation	Contribution	Rating
NeurIPS-23	M-7B	N/A	N/A	N/A	8.51 (10%)
	M-7B-R	0.20 (99%)	0.26 (99%)	0.32 (99%)	1.44 (99%)
	M-7B-3.5	0.15 (99%)	0.16 (99%)	0.27 (99%)	1.14 (99%)
	SEA-E	0.12 (100%)	**0.14 (100%)**	0.18 (100%)	0.80 (100%)
	SEA-EA	**0.11 (100%)**	0.15 (100%)	**0.17 (100%)**	**0.73 (100%)**
ICLR-24	M-7B	N/A	N/A	N/A	12.96 (13%)
	M-7B-R	0.32 (99%)	0.39 (99%)	0.42 (99%)	2.12 (99%)
	M-7B-3.5	0.32 (86%)	0.28 (86%)	0.45 (86%)	2.50 (86%)
	SEA-E	0.28 (100%)	0.30 (100%)	0.38 (100%)	2.11 (100%)
	SEA-EA	**0.27 (100%)**	**0.24 (100%)**	**0.34 (100%)**	**1.72 (100%)**

To assess the discrepancy between the generated scores and the reference scores, we use the Mean Squared Error (MSE) metric. The lower the MSE value, the more accurate the generated results. In Table 3, the percentages in parentheses indicate the proportions of generated reviews with valid scores, while "N/A" denotes those with unsuccessful generations (e.g. text is generated instead of scores). It can be seen that our proposed method ensures the validity of the output format, whereas other models tend to generate content that does not comply with the instruction to varying degrees, especially M-7B that has not undergone SFT. The MSE metric shows that our proposed methods outperform the baseline models in practically all cases. Although SEA-E scores larger than M-7B-3.5 by 0.02 in the "Presentation" on ICLR-2024, SEA-E achieves 100% valid scores in generation, whereas M-7B-3.5 only reaches 86%. Additionally, SEA-EA demonstrates improvements over SEA-E in most cases, further validating that a self-correcting strategy allows for high consistency between generated results and human feedback on quantitative evaluation results.

4.6 Qualitative Decision Analysis

In this part, we analyze "Decision" and "Reason" of the generated review, i.e., the final decision (accept or reject) of the paper and the corresponding reasons. Typically, the Area Chair (AC) gives the final decision and meta-reviews. We calculate the accuracy, precision, recall, and F1-score of the generated results compared to the final decisions, and use BERTScore to measure the semantic similarity between the reasons and meta-reviews. The model M-7B-R randomly selects a review for SFT that does not include the decision or the meta reviews, hence we do not take it as baseline.

Table 4: Qualitative Decision Analysis. The symbol (*) indicates that there are incompleteness or errors in the generated content; only valid generations are counted.

	Method	Accuracy	Precision	Recall	F1-score	BERTScore
NeurIPS-23	M-7B*	93.18	94.01	99.05	96.47	84.27
	M-7B-3.5*	81.01	95.34	83.91	89.26	84.04
	SEA-E	99.41	99.37	**100.0**	99.69	84.21
	SEA-EA	**99.70**	**99.69**	**100.0**	**99.84**	**85.22**
ICLR-24	M-7B*	36.81	37.14	**97.65**	53.82	84.19
	M-7B-3.5*	50.27	39.63	61.03	48.06	84.61
	SEA-E	54.16	43.31	69.95	53.50	85.07
	SEA-EA	**58.23**	**46.48**	71.36	**56.30**	**86.08**

From Table 4, it can be seen that SEA-EA leads to the largest accuracy and BERTScore values, where the latter shows the model's effectiveness in generating reasons semantically aligned with meta-reviews. Due to the acceptance rate of 95% in the NeurIPS-2023 test set (see Table 1), the overall results are large. For ICLR-2024, the accuracy of SEA-EA surpasses that of SEA-E over 4%, further indicating the effectiveness of the self-correction strategy. Additionally, we note that M-7B exhibits high recall about 97%, but poor precision, suggesting a tendency to cater to human preferences by accepting most papers. In contrast, our method performs better in both Precision and F1-score, which indicates that ours can identify papers of different quality more effectively. Overall, SEA aligns more closely with actual AC decisions and refrains from favoring decisions that lean towards acceptance.

4.7 Human and GPT evaluation

To further validate the performance of SEA, we supplemented the study with a questionnaire experiment for SEA-E and SEA-S, evaluated by both humans and GPT. The specific content of the questionnaire can be found in Appendix C.

4.7.1 Review Quality (SEA-E)

Referring to previous review quality instruments (Van Rooyen et al., 1999), we design 10 questions across different aspects, with each question rated on a scale of 1 to 10, where a higher score indicates better review quality. We then randomly sample 20 papers and evaluate the review quality generated by different models.

Human evaluation. We invite 20 qualified experts to evaluate reviews generated by various models, with each expert randomly assigned to assess 5 papers. Additionally, we include an extra question in the evaluation process: after anonymizing the model names, experts are asked to select the best review for each paper. The evaluation results from each expert were then aggregated, and the mean scores were calculated.

GPT evaluation. Given the inherent subjectivity in human evaluations and inspired by the work of Zheng et al. (2023), we employ *gpt-4o-2024-05-13* to score the reviews generated by different models using the same set of 10 questions. Due to the limitation of context length, GPT-4o did not perform the top-1 ranking evaluation.

Table 5: Review quality evaluation by humans and GPT. ('R2' refers to REVIEWER2.)

	Human Evaluation				GPT Evaluation			
	M-7B-R	M-7B-3.5	R2	SEA-E	M-7B-R	M-7B-3.5	R2	SEA-E
Q1	5.1	6.2	3.8	**7.9**	5.0	6.5	4.5	**7.2**
Q2	4.6	5.6	3.9	**7.8**	4.5	6.3	3.9	**7.2**
Q3	4.5	5.3	3.6	**8.0**	5.6	6.6	5.1	**7.9**
Q4	4.3	5.5	3.5	**7.8**	2.8	3.8	2.7	**5.5**
Q5	4.4	5.6	4.0	**8.0**	4.5	5.4	4.2	**7.2**
Q6	4.0	5.2	3.5	**8.2**	3.3	4.0	3.1	**6.2**
Q7	4.4	5.3	3.8	**7.4**	3.2	4.4	3.1	**5.6**
Q8	4.5	5.6	3.8	**7.7**	6.9	7.7	6.3	**8.3**
Q9	4.6	5.6	3.6	**7.8**	8.2	8.5	7.0	**9.2**
Q10	4.5	5.4	3.6	**7.9**	4.9	5.9	4.4	**7.3**
Top-1	0.0	0.1	0.0	**0.9**	-			

From the Table 5, it can be seen that SEA-E performs exceptionally well in both GPT-4o and human evaluations, significantly outperforming other models on ten questions. Additionally, in 90% of the cases, SEA-E was preferred by the experts.

4.7.2 Standardized Content Quality (SEA-S)

We design seven questions to assess the standardized content across different models, using the same scoring criteria for evaluating review quality. Then, we randomly select 10 papers for evaluation.

Human evaluation. We invite 10 experts to evaluate the standardized content generated by different models, with each expert being randomly assigned 5 papers for assessment.

GPT evaluation. We adopt the *GPT-4o* to score the standardised content generated by the different models based on the same questionnaire.

Table 6: Standardized content evaluation by humans and GPT.

	Human Evaluation				GPT Evaluation			
	M-7B	GPT-3.5	GPT-4	SEA-S	M-7B	GPT-3.5	GPT-4	SEA-S
Q1	6.0	5.6	8.2	**9.2**	8.4	9.0	9.0	**9.1**
Q2	6.7	7.5	7.6	**8.3**	8.9	8.8	8.7	**9.3**
Q3	5.3	5.9	7.8	**9.1**	8.5	8.8	9.1	**9.3**
Q4	7.3	3.7	7.3	**9.0**	8.7	7.1	8.0	**8.9**
Q5	7.5	5.0	7.7	**9.1**	9.6	9.7	9.4	**9.9**
Q6	6.7	6.8	6.9	**8.1**	9.3	9.6	9.4	**9.9**
Q7	5.1	5.2	8.1	**8.8**	8.5	9.0	9.0	**9.2**

Table 6 presents the results from two distinct evaluation methods. The human evaluation emphasizes SEA-S's strengths, especially in Q4, Q5, and Q6, where it excels in content relevance, logical coherence, and conciseness. While GPT-4o generally assigns higher scores across the board, SEA-S consistently stands out compared to other models.

Based on the results of the questionnaire, the effectiveness of the SEA framework is further confirmed. It demonstrates the ability to generate standardized content, leading to comprehensive and high-quality review feedback.

5 Conclusion

In this paper, we present SEA, a novel framework for automated paper reviewing. Specifically, we propose a new paradigm for constructing a standardized review dataset. Based on this dataset, we fine-tune an LLM to generate high-quality reviews. Moreover, we propose a new evaluation metric to measure the consistency between papers and generated reviews. Comprehensive experimental results demonstrate that the SEA framework can generate feedback that aligns with human reviews. In summary, we emphasize that the initial motivation for the field of automated paper review stems from the time-consuming and labor-intensive nature of traditional peer reviewing. Automated peer reviewing can provide timely feedback, enhancing research quality and accelerate the progress of scientific development. Therefore, we anticipate that the SEA framework will help researchers improve the quality of their work and shed light on the field of automated scientific reviewing.

Limitations

Despite these notable achievements, it is crucial to acknowledge the limitations of SEA, particularly its limited expansion into various academic disciplines and insufficient alignment with human standards. Here we elaborate on some of these constraints, along with intriguing future explorations.

Domain Expansion. Although the SEA framework has been successful in automating paper review generation within the machine learning field, it has not yet been expanded to other academic disciplines, such as physics and mathematics. As a universal automated paper review framework, SEA is able to generalize across any field. Thus, it would be exhilarating to investigate whether SEA can yield high-quality review feedback when applied to other academic disciplines.

Enhanced Consistency-Guided Training. Although optimizing the output of SEA-E by calculating mismatch scores between review and the original paper can generate review that are more consistent in content, we did not enhance SEA-E using natural language guidance based on scores during the training phase. To improve SEA-E in following instructions during the self-correction phase, we plan to collect relevant natural language guided self-correction dataset. By training on this dataset, we will further enhance SEA-E in content preference, enabling it to generate review feedback that aligns more accurately with the original paper.

Rebuttal Exploration. In the academic peer review process, the rebuttal stage is a critical component. During this stage, authors have the opportunity to correct potential misunderstandings by reviewers, clarify specific parts of their paper, or provide additional data and information to enhance the support for their research findings. Therefore, in our future research, we will explore methods to assist authors in making effective rebuttals.

Ethical Considerations

This paper proposes an automated paper reviewing framework that utilizes advanced long-context LLMs and supervised fine-tuning to align with human reviews and generate comprehensive reviews. This assists authors in improving the quality of their papers. As we explore the extensive potential of automated paper reviewing, it is essential to consider potential consequences associated with this technology. A significant concern is the misuse of the model. In the formal review processes of academic conferences, authors may receive reviews generated by the model without their knowledge. This situation could not only impact the fairness and transparency of the review process but also raise issues of trust and authenticity. To mitigate these risks, we will incorporate specific clauses in our usage license that strictly prohibit any misuse of the system, thereby ensuring it serves as a beneficial tool in academia.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (2023YFC3341200) and National Science Foundation of China (NSFC) (62137001).

References

Bo-Christer Björk and David Solomon. 2013. The publishing delay in scholarly peer-reviewed journals. *Journal of informetrics*, 7(4):914–923.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023a. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

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.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *arXiv preprint arXiv:2211.06651*.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

SPJM (Serge) Horbach and W (Willem) Halffman. 2018. The changing forms and expectations of peer review. *Research integrity and peer review*, 3:1–15.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

G Kamradt. 2023. Needle in a haystack–pressure testing llms.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. Peer review in scientific publications: benefits, critiques, & a survival guide. *Ejifcc*, 25(3):227.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for information Science and Technology*, 64(1):2–17.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023a. Automated scholarly paper review: concepts, technologies, and challenges. *Information fusion*, 98:101830.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023b. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wesley Morris, Scott Crossley, Langdon Holmes, and Anne Trumbore. 2023. Using transformer language models to validate peer-assigned essay scores in massive open online courses (moocs). In *LAK23: 13th international learning analytics and knowledge conference*, pages 315–323.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ofir Press, Noah Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng, Ye Liu. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Qiushi Sun, Zhirui Chen, Fangzhi Xu, Kanzhi Cheng, Chang Ma, Zhangyue Yin, Jianing Wang, Chengcheng Han, Renyu Zhu, Shuai Yuan, et al. 2024. A survey of neural code intelligence: Paradigms, advances and beyond. *arXiv preprint arXiv:2403.14734*.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.

Keren Tan, Kangyang Luo, Yunshi Lan, Zheng Yuan, and Jinlong Shu. 2024. An llm-enhanced adversarial editing system for lexical simplification. *arXiv preprint arXiv:2402.14704*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.

Susan Van Rooyen, Nick Black, and Fiona Godlee. 1999. Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *Journal of clinical epidemiology*, 52(7):625–629.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Shufa Wei, Xiaolong Xu, Xianbiao Qi, Xi Yin, Jun Xia, Jingyi Ren, Peijun Tang, Yuxiang Zhong, Yihao Chen, Xiaoqin Ren, et al. 2023. Academicgpt: Empowering academic research. *arXiv preprint arXiv:2311.12315*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv preprint arXiv:2211.06398*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A More Detailed Description of the Framework SEA

A.1 SEA-S

We further analyse the performance of SEA-S, the open-source model Mistral-7B, and the closed-source models GPT-3.5 and GPT-4 in standardised review experiments.

Instruction. In Table 10, we demonstrate our instructions for generating standardized review based on multiple reviews for each paper. We specify in the instruction that the model should integrate multiple reviews into three parts: textual descriptions, quantitative scores, and review results. The textual descriptions include "Summary", "Strengths", "Weaknesses", and "Questions", while the quantitative scores cover "Soundness", "Presentation", "Contribution", and "Rating". These elements are formatted in alignment with the original review template. Additionally, we incorporate the Area Chair's (AC) decision into the generated content, and instruct the model to generate corresponding acceptance or rejection reasons.

Standardization Examples. Figure 6 shows standardization examples from Mistral-7B, GPT-3.5, and SEA-S, which incorporate multiple reviews for the same paper. We can observe from the figure that the output of SEA-S is both rich and concise without redundant information. In contrast, the output from Mistral-7B not only lacks complete format but also has sparse content, with the missing parts highlighted in orange in the figure. As for the review generated by GPT-3.5, a significant portion consists merely of straightforwardly extracting original review content, failing to eliminate redundant information as instructed, such as the overuse of the phrase "Lack of", which is indicated in red to show the excessive repetition.

A.2 SEA-E

In Table 11, we present the instruction designed to generate reviews that conform to the specified format based on the content of the paper. In Figures 7 and 8, we display the reviews generated by different models for a particular paper, including Mistral-7B (M-7B), Mistral-7B-Random (M-7B-R), Mistral-7B-GPT-3.5 (M-7B-3.5), SEA-E, and SEA-EA. We can observe the following points: (1) Mistral-7B raises broad and general issues, tending to please humans. In the "Strengths" part, it splits the complexity issue into two points, which is not concise, and the content of the "Weaknesses" part does not match the paper decision. (2) Mistral-7B-Random visibly generates shorter texts with reduced detail. (3) Mistral-7B-GPT-3.5 generates duplicates due to insufficient standardization of the instruction dataset at the SFT stage, resulting in lower-quality reviews. (4) SEA-E and SEA-EA generate clearer viewpoints and ensure extensive coverage of content. (5) SEA-EA focuses more on the details within the paper. These comparisons demonstrate the superiority of SEA-E and SEA-EA in generating reviews.

A.3 SEA-A

To demonstrate the effectiveness of the regression model SEA-A, we randomly select a review for each paper from each dataset to form a paper-review pair. Then, we use SEA-A to calculate mismatch score, which is displayed in Table 7. Since SEA-A is trained with a majority of low-scoring samples, the values of the mismatch scores are not substantial. To enhance the intuitiveness of the main text, we present the results as Figure 5, and here Table 8 demonstrates the specific values instead. By comparing Table 7 with Table 8, each element of the former is larger than the corresponding item of the latter. Therefore, our regression model has the ability to discern the consistency between different reviewers and papers.

Table 7: Performance of mismatch scores in random pairs of papers and reviews.

Datasets	M-7B	M-7B-3.5	M-7B-R	SEA-E	SEA-EA
CONLL-16	1.1974	1.0118	0.5904	0.5832	0.5057
ACL-17	1.0146	0.6658	0.5784	0.4855	0.5006
COLING-20	0.9731	0.5553	0.4699	0.4733	0.4420
ARR-22	0.8285	0.5656	0.5262	0.4452	0.4043
NeurIPS-16-22	0.9640	0.8343	0.6974	0.5792	0.5536
ICLR-17-23	0.9850	0.6169	0.6551	0.4755	0.4474
NeurIPS-23	0.9451	0.7252	0.7022	0.5964	0.5513
ICLR-24	0.9348	0.6037	0.5935	0.4256	0.3999

Table 8: Performance of mismatch score in pairs of papers and corresponding reviews.

	Datasets	M-7B	M-7B-R	M-7B-3.5	SEA-E	SEA-EA
Cross-domain	CONLL-16	1.0503	0.8416	0.5665	0.5595	0.3926
	ACL-17	0.9309	0.6608	0.5257	0.4529	0.4359
	COLING-20	0.8642	0.5235	0.4446	0.3931	0.3888
	ARR-22	0.7095	0.5136	0.4964	0.3953	0.3926
In-domain	NeurIPS-16-22	0.8409	0.7282	0.6271	0.5098	0.4733
	ICLR-17-23	0.8630	0.5759	0.5924	0.4358	0.4227
	NeurIPS-23	0.7638	0.6511	0.6388	0.5109	0.4541
	ICLR-24	0.7746	0.5203	0.5396	0.4063	0.3788

Table 9: The overall performance (%) on the smaller test set.

Method	BLEU	ROUGE (Recall) R-1	R-2	R-L	ROUGE (F1-score) R-1	R-2	R-L	BERTScore
CONLL-16								
R2	15.21	17.15	4.27	8.63	25.24	6.40	12.67	**83.00**
M-7B	18.92	20.81	4.81	10.30	28.66	6.81	14.18	82.49
M-7B-R	18.16	21.96	5.17	10.62	29.56	7.18	14.31	82.57
M-7B-3.5	19.70	26.51	5.58	13.96	30.19	6.45	15.37	82.01
SEA-E	29.07	34.91	7.79	15.29	38.64	8.67	16.73	82.91
SEA-EA	**31.01**	**36.96**	**8.91**	**16.34**	**40.49**	**9.68**	**17.57**	82.94
ACL-17								
R2	14.20	17.66	4.42	8.89	23.86	6.25	12.07	82.26
M-7B	18.37	21.32	4.92	10.50	27.39	6.47	13.38	82.56
M-7B-R	17.93	22.14	5.15	10.84	27.50	6.72	13.34	82.47
M-7B-3.5	16.23	27.35	6.13	14.68	25.87	5.99	13.15	82.23
SEA-E	24.86	33.02	7.51	14.97	34.97	8.15	15.38	82.87
SEA-EA	**27.02**	**35.66**	**8.61**	**15.85**	**37.48**	**9.16**	**16.11**	**83.05**
COLING-20								
R2	18.08	23.71	5.49	12.14	28.57	6.75	14.60	82.04
M-7B	21.97	29.11	6.42	14.80	31.91	7.01	15.83	82.76
M-7B-R	19.49	29.21	6.69	15.20	30.23	6.80	15.25	82.27
M-7B-3.5	18.13	34.03	7.56	18.43	28.49	6.10	14.77	82.12
SEA-E	22.93	40.62	9.23	20.05	34.37	7.65	16.15	82.84
SEA-EA	**24.85**	**42.97**	**10.57**	**20.89**	**36.67**	**8.76**	**16.96**	**83.09**
ARR-22								
R2	17.87	22.62	6.20	11.83	28.62	8.13	15.03	79.29
M-7B	23.74	28.81	7.99	14.56	34.31	9.71	17.26	83.41
M-7B-R	21.77	28.49	7.66	14.86	32.60	8.98	16.84	82.72
M-7B-3.5	18.55	34.27	8.55	18.47	29.47	7.64	15.20	82.65
SEA-E	25.27	40.40	10.24	19.40	37.68	9.70	17.50	83.46
SEA-EA	**27.16**	**43.02**	**11.93**	**20.27**	**39.94**	**11.21**	**18.30**	**83.66**

Method	BLEU	ROUGE (Recall) R-1	R-2	R-L	ROUGE (F1-score) R-1	R-2	R-L	BERTScore
NeurIPS-16-22								
R2	10.41	11.00	3.94	5.95	18.23	6.64	9.88	**83.30**
M-7B	14.94	14.85	5.08	7.44	23.47	8.05	11.73	82.91
M-7B-R	12.86	14.14	4.78	7.46	21.65	7.52	11.22	82.56
M-7B-3.5	16.48	21.43	6.33	11.34	26.36	8.12	13.49	82.34
SEA-E	25.03	24.82	7.38	10.98	34.59	10.41	15.30	83.14
SEA-EA	**27.16**	**27.43**	**8.60**	**11.98**	**37.32**	**11.77**	**16.26**	83.28
ICLR-17-23								
R2	9.19	9.25	3.51	5.06	15.94	6.09	8.78	83.39
M-7B	13.53	12.93	4.54	6.46	21.50	7.59	10.78	83.22
M-7B-R	12.83	12.95	4.32	6.60	21.02	7.13	10.74	82.66
M-7B-3.5	16.22	19.10	5.75	10.14	25.71	7.98	13.16	82.73
SEA-E	23.21	22.17	6.88	9.89	32.31	10.14	14.47	83.48
SEA-EA	**25.29**	**24.70**	**7.95**	**10.75**	**35.17**	**11.45**	**15.37**	**83.62**
NeurIPS-23								
R2	7.84	8.29	3.33	4.63	14.68	5.91	8.23	83.19
M-7B	12.84	12.35	5.13	6.36	21.17	8.81	10.92	84.00
M-7B-R	12.34	12.18	4.93	6.27	20.57	8.36	10.65	83.68
M-7B-3.5	16.33	17.39	6.33	8.89	26.29	9.73	13.29	83.28
SEA-E	21.86	20.81	7.46	9.38	31.98	11.49	14.45	84.13
SEA-EA	**23.78**	**22.91**	**8.60**	**10.12**	**34.59**	**13.02**	**15.31**	**84.31**
ICLR-24								
R2	8.91	9.26	3.61	5.06	16.16	6.34	8.88	83.30
M-7B	13.25	12.74	4.90	6.37	21.50	8.30	10.78	83.98
M-7B-R	13.47	13.69	5.23	6.96	22.16	8.57	11.28	83.89
M-7B-3.5	16.88	20.21	6.68	10.56	27.32	9.34	13.79	83.44
SEA-E	23.06	22.58	7.62	9.84	33.57	11.38	14.68	84.05
SEA-EA	**25.44**	**25.19**	**8.81**	**10.70**	**36.62**	**12.88**	**15.61**	**84.23**

Interestingly, we apply the SEA framework to this paper and compare its generated review with the official feedback we receive. We find that the SEA framework aligns with some aspects of the actual review in terms of "Strengths," "Weaknesses," and "Questions." In addition, SEA provides further constructive suggestions for improvement, demonstrating its ability to generate comprehensive and high-quality review comments. The specific generated reviews are shown in Figure 9.

B Compare with REVIEWER2

To further validate the effectiveness of our framework SEA, we compare its performance with the open-source model of REVIEWER2 (Gao et al., 2024). Given that using two LLMs for inference process of REVIEWER2 is more time-consuming, we sample a smaller test set which is a subset of the test set used in this paper. Specifically, we randomly choose 100 samples from each dataset (or use all samples if the dataset contains fewer than 100). When inferring the model of REVIEWER2[2], we follow the settings described in the original paper. Table 9 lists the results for REVIEWER2 (abbreviated as R2), other baseline models, and our proposed framework. The results show that both SEA-EA and SEA-E exhibit excellent performance. In contrast, the results for REVIEWER2 are not ideal in the ROUGE metric and are unstable in the BERTScore metric. This is because REVIEWER2 often generates contents that are relatively short and lack valuable information. In contrast, our methods which fine-tune on a high-quality instruction dataset can generate more comprehensive reviews, demonstrating the superiority of our framework.

[2]https://github.com/ZhaolinGao/Reviewer2

C Human and GPT evaluation

In order to evaluate the review quality generated by SEA-E and the standardized content quality generated by SEA-S, we design a questionnaire consisting of 10 questions and 7 questions, respectively. For the review quality, we add an additional question during the human evaluation phase: selecting the best review from those generated by different models. The specific questionnaire contents are as follows:

SEA-E.

1. Did the reviewer discuss the importance of the research question?
2. Did the reviewer discuss the originality of the paper?
3. Did the reviewer clearly identify the strengths and weaknesses of the method (study design, data collection and data analysis)?
4. Did the reviewer make specific useful comments on the writing, organization, tables and figures of the manuscript?
5. Were the reviewer's comments constructive?
6. Did the reviewer supply appropriate evidence using examples from the paper to substantiate their comments?
7. Did the reviewer comment on the author's interpretation of the results?
8. Did the reviewer maintain objectivity and avoid bias in the review?
9. Was the language used by the reviewer professional and appropriate?
10. How would you rate the quality of this review overall?
11. Please select what you consider to be the best review. (The question only in the human evaluation)

SEA-S.

1. Did this review effectively consolidate similar points from the original reviews?
2. Did the review maintain objectivity and avoid bias?
3. Did this review contain a lot of useful information?
4. Did this review contain a lot of repetitive content?
5. Is there any contradiction in this review?
6. Was the language used by the review professional and appropriate?
7. How would you rate the quality of this review overall?

Mistral-7B

Summary:
This paper proposes XXX, a method for video stylization using a 3D CNN encoder backbone and XXX for separating motion and appearance features. The authors demonstrate the effectiveness of the proposed method on the XXX dataset.

Strengths:
- First to use 3D CNN for video stylization
- Temporally consistent results
- Quantitative evaluations of optical flow

Weaknesses:
- Motivations for using 3D CNN not well explained
- Usage of 3D CNN brings extra computational cost
- Style clip could be processed using a 2D CNN
- Intra-clip consistency measurement could be improved
- Model could be less elegant with multiple training phases
- Comparison with SOTAs not comprehensive

(Lack of Questions)

...

Paper Decision:
- Decision: Accept
- Reasons: The paper presents a novel approach to video stylization using a 3D CNN encoder and XXX for motion and appearance separation. While there are some weaknesses, such as the lack of clear motivations for using 3D CNN and the need for comprehensive comparison with SOTAs, the results demonstrate state-of-the-art performance. The authors are encouraged to address the weaknesses in future work.

GPT-3.5

Summary:
This paper introduces XXX, a method for video stylization using 3D CNNs and XXX to disentangle motion and appearance. The proposed method achieves state-of-the-art results on XXX datasets. While reviewers acknowledge the novelty and effectiveness of the approach, there are concerns about the motivation for using 3D CNN, model complexity, and the need for more comprehensive comparisons.

Strengths:
- This paper is the first to use 3D CNN for video stylization.
- The temporal and intra-loss improve the stylization stability.
- The proposed framework using 3D CNN for stylization is novel.
- Results are significantly better compared to baseline methods.
- The paper expands XXX to XXX.
- The quality of stylization results is good.

Weaknesses:
- The usage of 3D CNN is not well motivated.
- The technical contributions are weak.
- Extra computational cost compared with 2D CNN-based methods.
- The model needs to be trained with four cascaded phases.
- Lack of comprehensive comparison with state-of-the-art methods.
- Unclear necessity of four appearance subnets and Entangle Subnet.
- Lack of detailed network structure description.
- Weak paper writing and confusing logic in Section 3.
- Lack of analysis in ablation study.
- Limited comparison with related works.
- Lack of user study for subjective assessment.
- Lack of obvious improvement in quantitative comparison.
- Lack of comparison to single image stylization with optical flow motion compensation.

Questions:
- Why is 3D CNN used for stylization and what are the motivations for its usage?
- Why are four appearance subnets and the Entangle Subnet necessary?
- How does reducing the number of appearance subnets affect performance?
- How does the model training in different stages enforce network functionalities?
- How to disentangle motion and appearance using four appearance subnets?

...

Paper Decision:
- Decision: Accept
- Reasons: The paper presents a novel method for video stylization with 3D CNNs and XXX, achieving state-of-the-art results. While there are concerns about the motivation for using 3D CNN and the need for more comprehensive comparisons, the reviewers agree that the technical soundness and contributions of the paper warrant acceptance. Further improvements in addressing the reviewers' feedback could enhance the paper's impact.

SEA-S

Summary:
The paper introduces XXX, a method that utilizes 3D CNN (C3D) as the encoder backbone for video stylization, with a focus on separating motion and appearance using appearance subnets and XXX. This separation is intended to enhance the stylization stability by ensuring that motion and appearance are not entangled, which is a significant challenge in traditional 2D CNN-based methods. Results are demonstrated on the sports1M dataset, highlighting the superior performance of the proposed method compared to existing baseline methods. Despite the methodological advancements, the paper is criticized for its limited evaluation scope and unclear motivations for using 3D CNNs.

Strengths:
- The paper is the first to apply 3D CNNs to video stylization, which is a significant technical contribution.
- The use of a novel framework including an appearance subnet and Entangle network demonstrates innovative approaches in handling the stylization process.
- The proposed method has been trained in multiple phases, which ensures that the network effectively learns its functionalities.
- The results show significant improvements over baseline methods, indicating the effectiveness of the proposed approach.
- The paper introduces a large-scale dataset specifically curated for video stylization, enhancing the reproducibility and applicability of the research.

Weaknesses:
- The motivation and advantages of using 3D CNNs for video stylization are not well articulated, making the technical contributions seem weaker.
- The paper lacks a clear explanation of why four appearance subnets are necessary, and the role of the Entangle Subnet is unclear.
- The description of the network structure and its functionalities is confusing, particularly in Section 3 of the paper.
- The computational cost associated with using 3D CNNs compared to 2D CNNs is not addressed, raising concerns about the scalability and efficiency of the proposed method.
- Comparisons with state-of-the-art (SOTA) methods are not comprehensive, focusing only on optical flow metrics which might not adequately capture the quality of style transfer.
- There is a lack of user studies, which are critical for evaluating the subjective quality of the stylization results.

Questions:
- Can the authors clarify the necessity and role of the multiple appearance subnets and the Entangle Subnet in the proposed method?
- How would the performance of the model be affected if the number of appearance subnets were reduced?
- Why is a 3D CNN necessary when a 2D CNN could potentially suffice for extracting style features?
- Could the authors elaborate on how the proposed method disentangles motion and appearance, particularly in the context of the training phases?
- In the context of stylization, what are the advantages of using a 3D CNN over a 2D CNN with optical flow motion compensation?
- Could the authors provide more detailed comparisons with other SOTA methods, including qualitative and quantitative assessments across a broader range of metrics?

...

Paper Decision:
- Decision: Accept
- Reasons: Despite initial skepticism about the novelty of the approach, the paper managed to convince reviewers during the discussion phase that using 2D CNNs for the task introduces artifacts, which justifies the use of 3D CNNs. The introduction of a large-scale dataset for video stylization is seen as a significant contribution that will encourage further research in this domain. However, concerns about the clarity of the paper and the need for a more comprehensive evaluation were noted. The decision to accept is based on the overall strengths of the paper, the technical soundness, and the potential impact on the field, even though the evaluation scope and presentation clarity need improvement.

Figure 6: Examples of standardization for Mistral-7B, GPT-3.5 and SEA-S.

Mistral-7B

Summary:

The paper proposes XXX and XXX, two algorithms for risk-sensitive reinforcement learning (RL) in low-rank Markov Decision Processes (MDPs) with nonlinear function approximation. XXX is a representation learning algorithm that optimizes the CVaR metric using an MLE oracle and UCB-type bonuses, while XXX is a computationally efficient planning oracle that improves the computational complexity of XXX. Both algorithms are shown to be the first provably efficient algorithms for CVaR RL in low-rank MDPs.

Strengths:

* The paper extends the application of risk-sensitive RL to low-rank MDPs with nonlinear function approximation, which is an important step towards applying risk-sensitive RL to real-world problems where the state space is large and function approximation is necessary.
* The proposed XXX algorithm is the first provably sample-efficient algorithm for CVaR RL in low-rank MDPs, and the theoretical analysis of the algorithm demonstrates that it would provide an ₩(₩epsilon₩)-optimal CVaR with ₩(₩tilde{O}(1/₩epsilon^{2})₩) samples.
* The introduction of the XXX algorithm improves the computational complexity of XXX planning, making it only require polynomial running time with an MLE oracle.

Weaknesses:

* The paper assumes the underlying model transitions admit a low-rank decomposition, but it does not provide a method for estimating the rank of the underlying transition kernel.
* The paper assumes the availability of an MLE oracle, which may not be practical in all real-world applications.

Questions:

* How can we estimate the rank of the underlying transition kernel in low-rank MDPs without assuming it is known?
* Are there practical methods for implementing an MLE oracle in real-world applications?

Soundness:

The paper's soundness is good, as the algorithms proposed are based on solid theoretical foundations and the theoretical guarantees provided demonstrate the sample complexity and computational efficiency of the algorithms.

Presentation:

The paper's presentation is excellent, as it is well-written, clear, and easy to follow. The use of mathematical notation is clear and concise, and the results are presented in a logical and organized manner.

Contribution:

The paper's contribution is excellent, as it proposes the first provably efficient algorithms for risk-sensitive RL in low-rank MDPs, filling a gap in the existing body of knowledge.

Rating:

The paper's rating is strong accept, as it makes a significant contribution to the field of risk-sensitive RL and provides theoretical guarantees for the sample complexity and computational efficiency of the proposed algorithms.

Paper Decision:

Accept. The paper makes a significant contribution to the field of risk-sensitive RL and provides theoretical guarantees for the sample complexity and computational efficiency of the proposed algorithms. The weaknesses and questions identified can be addressed through future work.

Mistral-7B-Random

Summary: This paper studies risk-sensitive RL in low-rank MDPs with nonlinear function approximation. A UCB-bonus driven algorithm is proposed and proved to have a sample complexity of O(H2A2d4τ2ϵ2) to achieve an ϵ-optimal CVaR. To improve the computational efficiency, the authors also propose a computationally efficient planning oracle, which has a polynomial computational cost with respect to the dimension of representations.

Strengths: 1. This paper is well-motivated. The proposed algorithm is the first provably efficient CVaR RL algorithm in low-rank MDPs.   2. The algorithm and the theoretical analysis are sound.   3. The paper is well-written and easy to follow.

Weaknesses: 1. In Theorem 4.1, the sample complexity has a dependency of H7. It would be helpful if the authors could explain the reason for this.   2. In the related work section, the authors claim that "Unlike linear MDPs, since the underlying ψ and ϕ are unknown, we need to carefully balance representation learning, exploration, and worst-case failure stakes in low-rank MDPs." It seems that the only difference is that the parameters are unknown in low-rank MDPs. Does this mean that the theoretical analysis is non-trivial?

Questions: See weaknesses.

Soundness: 3 good

Presentation: 3 good

Contribution: 3 good

Rating: 6: marginally above the acceptance threshold

Mistral-7B-3.5

Summary:
This paper introduces a novel XXX algorithm for risk-sensitive reinforcement learning in low-rank Markov Decision Processes (MDPs) with nonlinear function approximation. The algorithm aims to maximize the Conditional Value at Risk (CVaR) with a fixed risk tolerance. The paper addresses the challenge of extending CVaR RL to settings with large state spaces and proposes a computationally efficient planning oracle. Theoretical analyses provide insights into the algorithm's sample complexity and computational complexity.

Strengths:
- This paper extends the CVaR RL algorithm to the low-rank MDP setting with nonlinear function approximation, which is a significant advancement.
- The paper provides a comprehensive theoretical analysis of the algorithm, demonstrating its sample efficiency.
- The authors introduce a computationally efficient planning oracle, improving the computational complexity of the algorithm.
- The paper is well-structured and clearly written, making it easy to follow.
- The authors provide a comprehensive theoretical analysis of the algorithm, demonstrating its sample efficiency.
- The paper addresses the challenge of extending CVaR RL to settings where the state space is large, which is a significant limitation of existing methods.
- The authors propose a novel discretized Least-Squares Value Iteration (LSVI) algorithm for the CVaR objective, which is computationally efficient.
- The paper provides theoretical guarantees on the sample complexity and computational complexity of the proposed algorithm.
- The paper is well-written and easy to follow.
- The paper is the first to study CVaR RL in low-rank MDPs with nonlinear function approximation.
- The paper provides a thorough discussion of the theoretical properties of the proposed algorithm.

Weaknesses:
- The paper does not provide a comparison of its algorithm with existing methods in the field.
- The paper does not provide empirical results to demonstrate the performance of the proposed algorithm.
- The paper does not discuss the limitations of the proposed algorithm.
- The paper does not provide a comprehensive comparison of its results with existing methods in the field.
- The paper does not provide empirical evaluation of the algorithm.
- The paper does not provide a comparison of the proposed algorithm with existing methods in the field.
- The paper does not discuss the limitations of the proposed algorithm.
- The paper does not provide empirical evaluation of the algorithm.
- The paper does not provide empirical evaluation of the algorithm.
- The paper does not provide empirical evaluation of the algorithm.

Questions:
- Could you provide a comparison of your algorithm with existing methods in the field?
- Could you provide empirical results to demonstrate the performance of your proposed algorithm?
- Could you discuss the limitations of your proposed algorithm?
- Could you provide a comprehensive comparison of your results with existing methods in the field?
- Could you provide empirical evaluation of your algorithm?
- Could you provide empirical evaluation of your algorithm?
- Could you provide empirical evaluation of your algorithm?
- Could you provide empirical evaluation of your algorithm?

Soundness:
3 good

Presentation:
3 good

Contribution:
3 good

Rating:
7 accept, but needs minor improvements

Paper Decision:
- Decision: Accept
- Reasons: The paper makes a significant contribution by extending CVaR RL to low-rank MDPs with nonlinear function approximation. The proposed algorithm is theoretically sound and computationally efficient. While the paper lacks empirical evaluations and comparisons with existing methods, the thorough theoretical analysis and well-structured presentation support its acceptance.

Figure 7: Examples of generated reviews for Mistral-7B, Mistral-7B-Random and Mistral-7B-GPT-3.5.

SEA-E

Summary:
The paper investigates the application of Conditional Value at Risk (CVaR) in low-rank Markov Decision Processes (MDPs), specifically focusing on the integration of exploration bonuses and risk-sensitive policies. The authors propose an algorithm called ELA (ELLA in the appendix), which leverages an MLE oracle to optimize CVaR and achieve near-optimal policies with polynomial sample complexity. The algorithm is further refined by introducing a computationally efficient planning oracle, ELLA, which enhances computational efficiency. The theoretical framework provided includes proofs of sample complexity and regret bounds, although these are critiqued for their clarity and the assumptions made.

Strengths:
- The paper introduces a novel algorithm for CVaR RL in low-rank MDPs, which is the first of its kind to address this specific setting.
- The algorithm is designed to be computationally efficient, with theoretical guarantees for both regret and sample complexity.
- The paper is well-written, with clear presentation of the algorithm and its theoretical underpinnings, making it accessible to readers.
- The authors provide a comprehensive discussion on the computational complexity of their algorithms, which is a significant contribution to the field.
- The paper addresses a relevant problem in risk-sensitive RL, which is crucial for practical applications in finance and risk management.

Weaknesses:
- The paper's theoretical analysis, particularly the proofs of regret and sample complexity, is not well-explained, leading to potential misunderstandings about the assumptions and the implications of the results.
- There is a lack of clarity in the definitions and theorems, which could hinder understanding and reproducibility of the results.
- The paper does not sufficiently discuss the limitations of its approach, particularly in terms of computational efficiency and the practical applicability of the assumptions made.
- The paper does not provide empirical evidence or simulations to support the claims made about the algorithm's performance, which could undermine the credibility of the results.
- The paper's reliance on a low-rank assumption and the need for a strong realizability assumption could limit the generalizability of the results.

Questions:
- Can the authors clarify the assumptions and implications of the regret and sample complexity results, particularly in terms of the assumptions made and the conditions under which the results hold?
- How does the algorithm handle the computational efficiency issues, especially in large-scale MDPs?
- Is there a possibility to extend the algorithm to handle non-discrete reward distributions, and if so, what are the challenges involved?
- How does the algorithm compare to existing algorithms in terms of computational efficiency and sample complexity?
- Could the authors provide more details on the computational efficiency of the algorithm, particularly in scenarios where the state space is large?
- What are the practical implications of the assumptions made in the paper, and how do these assumptions affect the generalizability of the results?

Soundness:
3 good

Presentation:
3 good

Contribution:
3 good

Rating:
5 marginally below the acceptance threshold

Paper Decision:
- Decision: Reject
- Reasons: The paper, while introducing a novel algorithm for CVaR RL in low-rank MDPs, suffers from several critical issues that led to its rejection. The primary concerns include the lack of clarity in the theoretical analysis, particularly the proofs of regret and sample complexity, which are not well-explained and rely on strong assumptions that may not hold in practical scenarios. Additionally, the paper does not provide empirical evidence to support the claims made about the algorithm's performance, which is essential for validating the theoretical results. Furthermore, the paper's assumptions, such as the realizability assumption and the low-rank assumption, are not adequately discussed, which raises concerns about the generalizability of the results. The decision to reject is supported by the metareview, which highlights these issues and suggests that the paper needs significant revisions to address these concerns before it can be considered for acceptance.

SEA-EA

Summary:
The paper investigates risk-sensitive reinforcement learning (RL) in the context of low-rank Markov Decision Processes (MDPs), focusing on the Conditional Value at Risk (CVaR) objective. The authors introduce an algorithm called ELA (ELLA in the appendix), which employs an Upper Confidence Bound (UCB) bonus-driven approach to manage exploration, exploitation, and representation learning. The algorithm leverages a discretized Least-Squares Value Iteration (LSVI) for computational efficiency. The paper provides theoretical analysis on the sample complexity and computational efficiency of the algorithm, showing that it achieves a sample complexity of O~(H^TAD^4/τ^2ε^2) and a computational complexity that depends on the dimension of representations rather than the state space size.

Strengths:
- The paper is the first to study risk-sensitive RL with CVaR in low-rank MDPs, providing a novel theoretical analysis that is relevant and interesting.
- The algorithm is well-explained and easy to follow, with clear mathematical notation and a comprehensive analysis of the algorithm's theoretical properties.
- The paper introduces a novel discretized Least-Squares Value Iteration (LSVI) algorithm for the CVaR objective, which is computationally efficient and achieves a polynomial running time with a Maximum Likelihood Estimation (MLE) oracle.
- The authors provide a comprehensive theoretical analysis of the algorithm, demonstrating that it would provide an ε-optimal CVaR with O(1/ε^2) samples, and show that the ELA algorithm is the first provably sample-efficient algorithm for CVaR RL in low-rank MDPs.
- The paper is well-written, with a clear structure and logical flow, and the mathematical notation is well-defined and easy to follow.

Weaknesses:
- The paper's assumption of access to exact Value Iteration (VI) is strong and not justified, especially considering the computational challenges in real-world applications.
- The paper lacks empirical validation of the algorithm, which could help in understanding its practical performance and limitations.
- The computational complexity of the algorithm, particularly in terms of the dependence on the number of episodes H, is not thoroughly discussed or justified.
- The paper does not adequately discuss the scalability of the algorithm with increasing state space size, which is a critical factor in practical applications.
- The paper does not compare its results with other state-of-the-art algorithms, which could help in understanding the relative performance and effectiveness of the proposed method.

Questions:
- Can the authors justify the assumption of access to exact Value Iteration (VI) and discuss the implications of this assumption on the practical applicability of the algorithm?
- How does the computational complexity of the algorithm scale with the number of episodes H, and what are the practical implications of this dependence?
- Could the authors provide empirical validation of the algorithm to demonstrate its practical performance and limitations?
- How does the proposed algorithm compare with other state-of-the-art algorithms in terms of computational efficiency and sample complexity?
- Could the authors discuss the potential applications of the algorithm in practical scenarios, such as finance or healthcare, and how it might be adapted to different MDP settings?

Soundness:
3 good

Presentation:
3 good

Contribution:
3 good

Rating:
7 accept, but needs minor improvements

Paper Decision:
- Decision: Accept
- Reasons: The paper presents a novel algorithm for risk-sensitive RL in low-rank MDPs, which is both theoretically sound and computationally efficient. The algorithm introduces a discretized Least-Squares Value Iteration (LSVI) for computational efficiency and achieves a sample complexity of O~(H^TAD^4/τ^2ε^2). The paper is well-written, with clear mathematical notation and a comprehensive analysis of the algorithm's theoretical properties. However, the paper lacks empirical validation of the algorithm, which could help in understanding its practical performance and limitations. The decision to accept is based on the paper's theoretical contributions and the computational efficiency of the proposed algorithm, but it is recommended that the authors address the missing empirical validation in future work.

Figure 8: Examples of generated reviews for SEA-E and SEA-EA.

SEA

Summary:

The paper introduces the SEA framework for automated peer reviewing in academic conferences, designed to streamline the review process by integrating multiple reviews into a standardized format. The framework consists of three modules: Standardization, Evaluation, and Analysis. The Standardization module distills data standardization capabilities of GPT-4 to integrate reviews into a unified format, the Evaluation module fine-tunes Mistral-7B to generate constructive reviews, and the Analysis module introduces a new evaluation metric called the mismatch score to assess consistency between paper contents and reviews. Extensive experiments are conducted to demonstrate the effectiveness of the SEA framework, showing improvements in consistency and review quality compared to existing methods.

Strengths:

- The paper is well-written and easy to follow, making complex concepts accessible to a broad audience.
- The introduction of the SEA framework represents a significant advancement in the automation of peer reviewing, offering a more efficient and standardized approach to the review process.
- The SEA framework is innovative in its use of large language models (LLMs) for automated reviewing, showcasing the potential of these models to improve the quality and consistency of reviews.
- The proposed evaluation metric, the mismatch score, is a novel and effective method for assessing the consistency between paper contents and reviews, providing a valuable tool for ensuring the accuracy and reliability of generated reviews.
- The framework demonstrates a clear improvement in the consistency and quality of reviews generated by LLMs, as evidenced by experimental results and comparisons with baseline models.

Weaknesses:

- The novelty of the framework is somewhat limited as it primarily represents an application of existing LLM capabilities to a specific task, without introducing fundamentally new methodologies.
- The paper does not adequately address the potential biases and limitations of the LLMs used, which could impact the fairness and accuracy of the generated reviews.
- The methodology of the framework, particularly the use of self-correction strategy and the evaluation of consistency, is not sufficiently detailed or explained, which could hinder understanding and replication of the results.
- The paper does not provide sufficient qualitative analysis or examples of generated reviews, which could help in understanding the specific improvements and limitations of the SEA framework.

Questions:

- Could the authors provide more details on the specific LLMs used in the SEA framework, including their capabilities and limitations in the context of academic reviewing?
- How does the SEA framework handle the potential biases and limitations of the LLMs used, and what measures are in place to ensure fairness and accuracy in the generated reviews?
- In the self-correction strategy, how is the threshold for review regeneration determined, and what are the implications of this decision on the quality and consistency of the generated reviews?
- How does the SEA framework ensure that the generated reviews are not simply rephrased versions of existing reviews, which could undermine the novelty and usefulness of the reviews?
- Could the authors provide more detailed examples or case studies of the generated reviews, along with human evaluations or comparisons with existing reviews to better understand the impact and effectiveness of the SEA framework?

Soundness:
3 good

Presentation:
3 good

Contribution:
3 good

Rating:
7 accept, but needs minor improvements

Paper Decision:

- Decision: Accept
- Reasons: The paper presents a well-structured framework for automated peer reviewing, demonstrating its effectiveness through extensive experiments. The introduction of the mismatch score as a new evaluation metric adds a valuable contribution to the field. Despite some concerns about the novelty and potential biases of the LLMs used, the paper is well-written and the experimental results are robust, showing improvements over existing methods. The decision to accept is supported by the review content, which highlights the paper's contribution and the overall positive reception. The decision also aligns with the suggestions for minor improvements to enhance clarity and address some of the identified limitations.

Figure 9: The review generated by applying the SEA framework to this paper.

Table 10: Instruction for generating standardized review based on multiple reviews for each paper.

INSTRUCTION:

As an experienced academic paper reviewer, you are presented with different review contents for the same paper. Please analyze these contents carefully and consolidate them into a single review. The review should be organized into nine sections: Summary, Strengths, Weaknesses, Questions, Soundness, Presentation, Contribution, Rating and Paper Decision. Below is a description of each section:

1. Summary: Combine the 'Summary' sections from all reviews into a cohesive summary, aiming for a length of about 100-150 words.
2. Strengths/Weaknesses/Questions: Combine the Strengths/Weaknesses/Questions sections from all reviews into a unified, cohesive bullet-point list that avoids redundancy while preserving the specific details and depth of each point.
3. Soundness/Presentation/Contribution: Aggregate the Contribution/Soundness/Presentation score from each review to determine a suitable overall score (the score must be an **integer**), then, match this integer score to the corresponding criterion from the list below and provide the result. For example, if the score is 3, the result should be '3 good'. The possible scores and their criteria are:

1 poor \n 2 fair \n 3 good \n 4 excellent

4. Rating: Aggregate the 'Rating' from each review to determine a suitable overall Rating (the Rating must be an **integer**), then, match this integer Rating to the corresponding criterion from the list below and provide the result. For example, if the Rating is 1, the result should be '1 strong reject'. The possible Ratings and their criteria are:

1 strong reject
2 reject, significant issues present
3 reject, not good enough
4 possibly reject, but has redeeming facets
5 marginally below the acceptance threshold
6 marginally above the acceptance threshold
7 accept, but needs minor improvements
8 accept, good paper
9 strong accept, excellent work
10 strong accept, should be highlighted at the conference

5. Paper Decision: It must include the Decision itself (Accept or Reject) and the reasons for this decision which is based on Meta-review, the criteria of originality, methodological soundness, significance of results, and clarity and logic of presentation, etc. Please ensure your Decision (Accept/Reject) matches the value of the 'Decision' key in the JSON, if present.

Here is the template for a review format, you must follow this format to output your review result:
Summary: \n <Summary content> \n

Strengths: \n <Strengths result> \n
Weaknesses: \n <Weaknesses result> \n
Questions: \n <Questions result> \n

Soundness: \n <Soundness result> \n
Presentation: \n <Presentation result> \n
Contribution: \n <Contribution result> \n
Rating: \n <Rating result> \n

Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content

Table 11: Instructions for generating review comments based on the content of the paper.

INSTRUCTION:

You are a highly experienced, conscientious, and fair academic reviewer. Please help me review this paper. The review should be organized into nine sections:
1. Summary: A summary of the paper in 100-150 words.
2. Strengths/Weaknesses/Questions: The Strengths/Weaknesses/Questions of paper, which should be listed in bullet points, with each point supported by specific examples from the article where possible.
3. Soundness/Contribution/Presentation: Rate the paper's Soundness/Contribution/Presentation, and match this score to the corresponding criterion from the list below and provide the result. The possible scores and their criteria are:
   1 poor
   2 fair
   3 good
   4 excellent
4. Rating: Give this paper an appropriate rating, match this rating to the corresponding criterion from the list below and provide the result. The possible Ratings and their criteria are:
   1 strong reject
   2 reject, significant issues present
   3 reject, not good enough
   4 possibly reject, but has redeeming facets
   5 marginally below the acceptance threshold
   6 marginally above the acceptance threshold
   7 accept, but needs minor improvements
   8 accept, good paper
   9 strong accept, excellent work
   10 strong accept, should be highlighted at the conference
5. Paper Decision: It must include the Decision itself (Accept or Reject) and the reasons for this decision which is based on the criteria of originality, methodological soundness, significance of results, and clarity and logic of presentation.

Here is the template for a review format, you must follow this format to output your review result:
Summary: \n <Summary content> \n

Strengths: \n <Strengths result> \n
Weaknesses: \n <Weaknesses result> \n
Questions: \n <Questions result> \n

Soundness: \n <Soundness result> \n
Presentation: \n <Presentation result> \n
Contribution: \n <Contribution result> \n
Rating: \n <Rating result> \n

Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content

Please ensure your feedback is objective and constructive. The paper is as follows: <paper content>