

REVIEWER2: Optimizing Review Generation Through Prompt Generation

Anonymous ACL submission

Abstract

Recent developments in LLMs offer new opportunities for assisting authors in improving their work. In this paper, we envision a use case where authors can receive LLM-generated reviews that uncover weak points in the current draft. While initial methods for automated review generation already exist, these methods tend to produce reviews that lack detail, and they do not cover the range of opinions that human reviewers produce. To address this shortcoming, we propose an efficient two-stage review generation framework called REVIEWER2. Unlike prior work, this approach explicitly models the distribution of possible aspects that the review may address. We show that this leads to more detailed reviews that better cover the range of aspects that human reviewers identify in the draft. As part of the research, we generate a large-scale review dataset of 27k papers and 99k reviews that we annotate with aspect prompts, which we make available as a resource for future research.

1 Introduction

Asking fellow group members to critique a draft is widely regarded as a valuable way of improving scientific writing, and the lack of access to such peers outside of well-resourced research groups is a key source of inequality (Merton, 1968; Nielsen and Andersen, 2021; Kozłowski et al., 2022). Furthermore, even in well-resourced groups, the frequency with which authors can receive feedback is limited. In this paper, we thus develop techniques for generating automated feedback via LLMs to aid authors in enhancing the quality of their work before it enters the formal peer review. This helps level the playing field, and it promises to reduce pressure on the peer review process (Lee et al., 2012) after experiencing exponential increases in submissions (Björk and Solomon, 2013; Bornmann and Mutz, 2014; Kelly et al., 2014).

The ability of LLMs to reason about complex tasks gives them the potential to provide automated feedback on papers (Liu and Shah, 2023; Liang et al., 2023). A key asset is that we already have substantial amounts of supervised data from peer reviews (Kang et al., 2018a; Yuan et al., 2021; Shen et al., 2022; Dycke et al., 2023), containing paper-review pairs across different years, venues, and subjects. Prior approaches to review generation (Yuan et al., 2021; Lin et al., 2023) focus on fine-tuning a pre-trained language model based on these datasets. However, unlike typical instruction following tasks (Ouyang et al., 2022; Touvron et al., 2023), we argue that open-ended review generation is under-specified in a way that makes it difficult to align language models for instruction following. In particular, asking an LLM to generate a review without specifying which aspects of the paper to focus on exposes the model to substantial uncertainty. This leads to shortcomings along the following dimensions:

Specificity. Peer reviews exhibit varying levels of specificity from general (e.g., "the paper is technically sound.") to precise (e.g., "the paper has a good theoretical basis based on the derivation in section 3."). A good review should provide detailed justifications for its assessment, especially when stating the weaknesses of the paper (Yuan et al., 2021). In addition, justifications make the review more constructive as they provide direct instructions on how to improve the paper (Xiong and Litman, 2011). However, our experiments reveal that standard fine-tuning diminishes the specificity of the generated reviews. An example is shown in Table 1 where we generate reviews based on a model that is fine-tuned over increasing numbers of training steps. The generated review is significantly more generic at step 2000 compared to the one at step 500.

Coverage and Control. Different human reviewers are likely to focus on different aspects of

Training Step 500	Training Step 1000	Training Step 2000
The paper proposes a simple and efficient differentiable data generation pipeline.	The authors have done extensive experiments to validate the effectiveness of the proposed method.	The paper is well-written and straightforward. The method is technically sound.

Table 1: Generated reviews from different steps using a generic prompt (1 epoch \approx 1000 steps).

a paper. An automated review-generation system should thus cover the range of issues that human reviewers may identify. We find that standard fine-tuning of LLMs for review generation often leads to a form of regression-to-the-mean, where the generated reviews do not cover the full range of aspects. We argue that an ideal system should actively control coverage, and give authors the ability to ask for feedback on specific aspects.

To address these issues, we propose an efficient two-stage review generation framework for papers called REVIEWER2. REVIEWER2 includes two fine-tuned language models. The first LLM analyzes the paper and produces a set of aspects that the reviews should focus on. Each of these aspects takes the form of a prompt that is the input for the second stage. The second LLM generates a review based on the paper and the aspect prompt. We implement REVIEWER2 based on LongLoRA (Chen et al., 2023) to enable 32k context length, avoiding the use of extractive summaries of the paper that was necessary in prior work due to limitations in context length (Gehrmann et al., 2018; Chen and Bansal, 2018; Dou et al., 2021; Yuan et al., 2021; Lin et al., 2023).

Unfortunately, existing peer-review datasets do not include aspect prompts, providing insufficient data for training either stage of REVIEWER2. To address this issue, we develop a Prompt Generation with Evaluation (*PGE*) pipeline to generate a variety of high-quality aspect prompts. *PGE* generates prompts given the review and uses a self-evaluation step to ensure the quality of the generated prompts. Based on *PGE*, we construct a large-scale review dataset of 27k papers and 99k reviews from six different venues with corresponding aspect prompts.

Extensive experiments on multiple venues demonstrate that our REVIEWER2 framework trained on *PGE*-generated aspect prompts substantially outperforms existing methods in terms of review quality, specificity, and coverage. The major contributions of this paper are summarized below:

- We propose REVIEWER2, a novel framework for joint aspect prompt and review generation that improves coverage and enables control.
- We implement REVIEWER2 based on Lon-

gLoRA, enabling 32k context length with low memory requirement for fine-tuning.

- We design two new metrics for evaluating specificity and coverability. We compare REVIEWER2 with various baseline methods and find that it substantially improves review generation.
- We propose *PGE*, a novel pipeline for augmenting existing review datasets with aspect prompts, and we construct the first large-scale peer review dataset that includes aspect prompts.

2 Related Work

Instruction generation and tuning. Previous works have demonstrated the efficacy of instruction fine-tuning in enhancing both task performance and adaptability to unseen tasks (Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022). However, these approaches depend heavily on human-written instruction data, which is often constrained in terms of quantity and diversity. Several works have explored using large language models (LLMs) to automatically generate instructions. Honovich et al. (2022) prompts a language model with seed examples of instructions to generate additional instructions, inputs, and outputs. Wang et al. (2023) adopts a similar approach while filtering the generated instructions to ensure diversity and quality.

Self-alignment. Self-alignment of LLMs is an emerging area of research that utilizes the model to improve itself and align with human values with minimal human supervision. This field primarily consists of two approaches: unsupervised data generation and post-hoc output refinement. In Li et al. (2023), prompts and responses are generated according to a small set of human-written principles, while Sun et al. (2023) focuses on generating synthetic prompts derived from human-written documents. On the other hand, Madaan et al. (2023) employs an iterative process to refine its output through generated feedback.

Automation in peer review. Automated systems have played a significant role in various aspects of the review process. Numerous algorithms (Stelmakh et al., 2019; Kobren et al., 2019; Cohan et al., 2020) have been developed to evaluate the expertise of potential reviewers, optimizing reviewer-

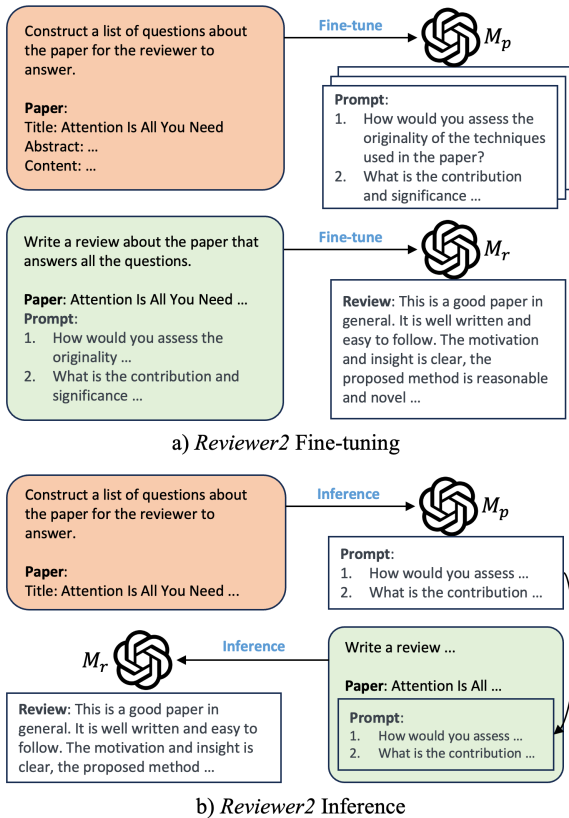


Figure 1: Illustrations of REVIEWER2. **a)** REVIEWER2 fine-tunes two models: M_p generates aspect prompts based on paper, and M_r generates reviews based on the paper and a prompt. **b)** REVIEWER2 utilizes a two-stage inference to generate an aspect prompt and generate the review based on the generated prompt.

paper assignments. In addition, several algorithms have been proposed to ensure the submissions adhere to appropriate guidelines, such as plagiarism detection (Foltýnek et al., 2019) and desk rejection prediction (Ghosal et al., 2019). Recently, efforts have been directed towards the development of algorithms for review generation (Yuan et al., 2021; Lin et al., 2023), leveraging papers as input and fine-tuning on LLMs for review generation.

3 REVIEWER2 for Review Generation

In this section, we introduce our REVIEWER2 pipeline for generating reviews. The key idea is to insert explicit control into the pipeline to ensure that the generated reviews cover the full range of aspects that human reviewers may comment on. We demonstrate that this improves both coverage and specificity of the generated reviews.

Figure 1(a) illustrates how we train the two stages of REVIEWER2. For the first stage, we fine-tune an LLM

$$M_p : p \rightarrow \{x^1, \dots, x^k\}$$

to produce a set of aspect prompts x^1, \dots, x^k for paper p that cover the aspects that a reviewer may comment on for this paper. For the second stage of REVIEWER2, we fine-tune another LLM

$$M_r : (p, x) \rightarrow y$$

to produce a review y for paper p that addresses aspect x . When generating a review for a new paper p' , we first query M_p for an aspect prompt x . We then query M_r to produce a review y for the generated aspect prompt. This inference process is depicted in Figure 1(b). We will provide evidence that this two-stage pipeline not only provides explicit control of aspect coverage, it also avoids a type of regression-to-the-mean (Barnett et al., 2004) that makes single-stage pipelines produce generic reviews with little specificity.

An illustrative example is shown in Figure 2 which contains three reviews, $\{y_i^1, y_i^2, y_i^3\}$, for paper p_i . All three reviews comment on either or both theoretical and empirical justifications, representing the general aspects. However, the reviews provide different suggestions for improvement, which are considered as specific parts. We find that a single-stage pipeline that is trained without aspect prompts tends to only generate the general components of the review, as illustrated in Figure 2(b), since such "mean reviews" align closely with all three reviews. On the other hand, by adding aspect prompts $\{x_i^1, x_i^2, x_i^3\}$ derived from the paper, the augmentation diversifies the aspects that are addressed, aligning it more effectively with the variability seen in the human reviews. Note that the prompt space now better captures the variability between reviewers, which reduces the noise when mapping to generated reviews. This reduction in noise enables the generation of more specific reviews, \hat{y}_i , during inference as shown in Figure 2(c). The geometric intuition behind our illustration is detailed in Appendix A.

To enable efficient long context fine-tuning and inference, we adapt LoRA⁺ and S²-Attn from Chen et al. (2023). LoRA⁺ extends on top of LoRA (Hu et al., 2021) by making the embedding and normalization layers trainable, and S²-Attn groups input tokens to address the quadratic complexity of self-attention.

4 Review Dataset with Aspect Prompts

Training REVIEWER2 requires a dataset of papers and reviews that is augmented with aspect prompts.

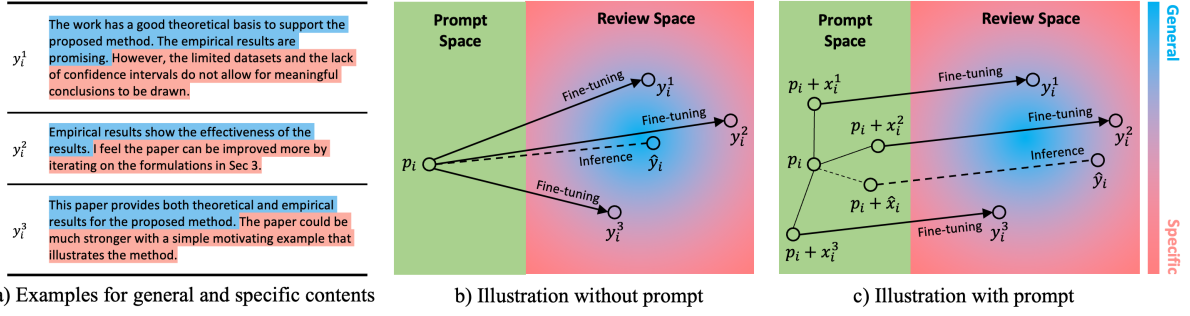


Figure 2: Illustrations of the effect of aspect prompts. **a)** General content is highlighted in blue, while specific content is highlighted in red. **b)** Fine-tuning without aspect prompts causes the generated contents to be general during inference. **c)** Fine-tuning with aspect prompts allows specific content generation during inference.

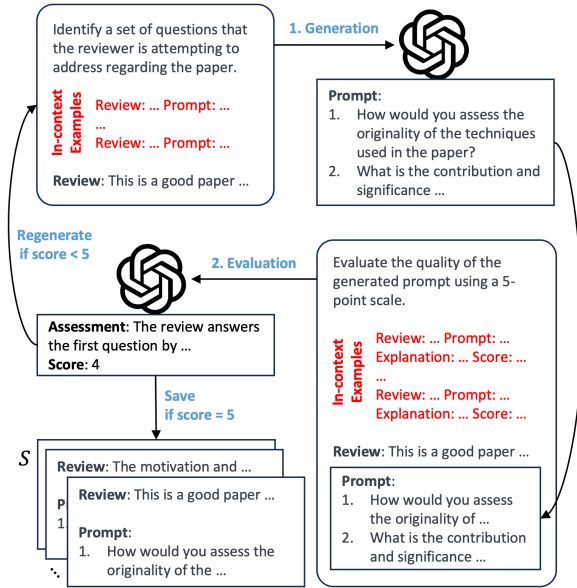


Figure 3: *PGE* includes two steps: generation and evaluation. The prompt is regenerated if the score is below 5 on a 5-point scale, otherwise, it is saved to S .

While there is ample data on papers and their associated reviews, these datasets contain generic review prompts that do not capture which aspects the human reviewer chose to focus on. We therefore developed the following methodology for augmenting existing review datasets with aspect prompts.

The result is the first review dataset that is annotated with aspect prompts, and we make this dataset available as a new resource. It consists of up-to-date crawls of publicly available reviews from NeurIPS and ICLR, and we also augment the datasets from PeerRead (Kang et al., 2018b) and NLPeer (Dyck et al., 2023).

4.1 PGE: Prompt Generation with Evaluation

In order to generate the corresponding prompt for each review, we propose Prompt Generation with Evaluation (*PGE*) pipeline consisting of a generation step and an evaluation step, as shown in

Figure 3. Specifically, given a set of m papers $P = \{p_1, p_2, \dots, p_m\}$ and corresponding reference reviews $Y = \{y_i^n | 1 \leq i \leq m, 1 \leq n \leq n_i\}$ where n_i is the number of reviews for paper i , the goal of the pipeline is to generate a set of prompts $X = \{x_i^n | 1 \leq i \leq m, 1 \leq n \leq n_i\}$ that one prompt corresponds to one review.

For a review y_i^n , the generation step generates a prompt, x_i^n , and the evaluation step evaluates the generated prompt based on a 5-point scale. If x_i^n achieves a score of 5, the pair (x_i^n, y_i^n) is stored in the set S , $S = S \cup \{(x_i^n, y_i^n)\}$, otherwise the prompt is regenerated. This two-step iterative approach resolves the problem of the absence of ground-truth prompts for reviews and ensures the quality of prompt generation without human supervision. The prompts we used for generation and evaluation are shown in Appendix B.

Prompt Generation. We initialize S with human-annotated examples that will be used as initial in-context examples during generation. To construct these examples, we use Llama-2-70B-Chat (Touvron et al., 2023) to generate prompts for a randomly selected subset of 100 reviews in a zero-shot fashion. Then, we manually refine the prompts by removing irrelevant questions, adding missing questions that are covered in the review, and refining to align with the open-ended format of review questions. An example of a review-prompt pair is shown in Appendix C.

To enhance the performance of prompt generation, we apply in-context learning (ICL) (Dong et al., 2023) in the process. The in-context examples are randomly sampled from S . As more prompts are generated and saved to S , the pool of available examples also expands, ensuring the diversity of the prompts. We always sample the maximum possible number of in-context examples while satisfying the context length constraint.

Table 2: Dataset Statistics

	CONLL-16	ACL-17	COLING-20	ARR-22	ICLR-17-23	NeurIPS-16-22	total
# papers	22	137	89	476	16,327	10,754	27,805
# words per paper	4,325	4,679	4,230	4,850	6,959	5,236	6,229
# reviews	39	275	112	684	58,933	39,684	99,727
# words per review	418	440	414	397	512	482	487
# prompts	37	270	108	676	58,107	38,762	97,960
# words per prompt	56	60	45	46	52	51	53
% accepted	50%	67%	93%	100%	32%	98%	55%
domain	NLP/CL	NLP/CL	NLP/CL	NLP/CL	ML	ML	multi

Table 3: Dataset Comparison

	# papers	# reviews	prompts
PeerRead (Kang et al., 2018b)	3,006*	10,770	✗
ASAP-Review (Yuan et al., 2021)	8,877	28,119	✗
MReD (Shen et al., 2022)	7,894	30,764	✗
NLPeer (Dycke et al., 2023)	5,672	11,515	✗
Ours	27,805	99,727	✓

*Number of papers that have reviews.

Prompt Evaluation. Similar to generation, we also apply ICL during the evaluation step. We use Llama-2-70B-Chat to evaluate the review-prompt pair based on a 5-point scale with five in-context examples for each score from 1 to 5. The in-context examples (shown in Appendix D) are manually constructed and remain consistent across all evaluations. Inspired by chain-of-thought prompting (Wei et al., 2023), we prompt the LLM to generate an explanation for the score before producing the final score to encourage more accurate assessments.

Regeneration. To ensure the quality of the generated prompt, the pipeline regenerates the prompt if the score is not 5. Since the in-context examples for generation are randomly sampled rather than a fixed set, the regeneration step is guaranteed to generate a different prompt compared to the previous generations, minimizing redundancy. We use a limit of 5 generations per review, and the review is excluded from further generation if it exceeds the limit. 93.60% of the reviews take less than or equal to 3 generations to reach a score of 5.

4.2 Dataset Details

We incorporate parts of the PeerRead and NLPeer datasets. **CONLL-16** and **ACL-17** from PeerRead contain papers and reviews from the NLP domain. The reviewing process is double-blind and the formats of the review are unstructured. NLPeer’s **COLING-20** and **ARR-22** are collected

via a donation-based workflow in NLP domain with formats in free-form reports and standardized structured review forms.

In addition to the prior datasets, we crawl ICLR papers from 2017 to 2023 through OpenReview¹ and NeurIPS papers from 2016 to 2020 through NeurIPS Proceedings² and from 2021 to 2022 through OpenReview. The resulting datasets are **ICLR-17-23** and **NeurIPS-16-22**. For each paper’s review, we follow the format of the previous datasets to keep as much metadata information as possible including reference and meta reviews from official reviewers, and final decisions.

Unification. The diverse sources of datasets are converted into a unified format to enhance accessibility and consistency. For each paper, we include the full text of the paper, metadata, and corresponding reviews and prompts. For the contents of the paper, we use Science Parse³ from AllenAI to parse the PDFs of the papers into construct structured JSON files. Each paper is accompanied by detailed metadata, providing essential information about the paper. The detailed sections of paper and metadata are shown in Appendix E. The reviews contain both textual components and scores that are divided into different sections based on the venue-specific formats. In addition, we employ our *PGE* pipeline to construct a prompt for each review. For simplicity, we only use the text part of the review for prompt generation and review generation.

Analysis. The statistics of our dataset are shown in Table 2. Our dataset consists of more than 27k papers and 99k reviews in various domains. The average paper length spans from 4k to 7k, demonstrating substantial variability. The review length and prompt length exhibit smaller variances, averaging from 400 to 500 and 45 to 60 respectively. Compared to other review datasets (Table 3), our dataset has the largest number of papers and re-

¹<https://openreview.net/>

²<http://papers.neurips.cc/>

³<https://github.com/allenai/science-parse>

Table 4: Results of the model variations using three metrics across six venues (SS-E0: SINGLES-E0, SS-E: SINGLES-E, SS: SINGLES, R2-E: REVIEWER2-E, R2: REVIEWER2). The best-performing model for each venue and metric is highlighted in bold.

		Method	BLEU (max)	ROUGE (max)			BertScore (max)
				R-1	R-2	R-L	
In-domain	ICLR	SS-E0	8.15	29.93	7.14	13.76	68.45
		SS-E	12.53	39.63	10.19	19.76	79.40
		R2-E	13.32	40.06	10.59	20.34	80.11
		SS	15.08	40.77	11.78	21.09	81.18
		R2	16.94	44.58	13.56	22.62	83.61
	NeurIPS	SS-E0	8.29	28.96	6.98	13.63	67.82
		SS-E	11.72	39.54	9.75	19.67	79.17
		R2-E	12.91	39.87	10.02	19.81	80.17
		SS	14.44	40.62	11.22	20.8	81.83
		R2	16.24	42.15	13.11	22.52	83.23
Cross-domain	ACL	SS-E0	5.02	30.77	6.28	12.69	68.90
		SS-E	4.67	35.23	7.07	16.53	78.15
		R2-E	4.82	36.44	7.98	16.73	80.03
		SS	5.40	35.73	7.94	16.94	80.25
		R2	6.49	36.88	8.04	17.77	83.65
	ARR	SS-E0	6.01	32.48	7.89	13.91	69.34
		SS-E	6.89	38.30	9.67	18.67	79.09
		R2-E	6.96	39.17	10.94	19.53	80.69
		SS	6.73	38.93	11.22	19.61	81.03
		R2	7.46	40.18	12.04	20.76	82.29
COLING	SS-E0	3.66	30.51	6.49	12.83	69.19	
	SS-E	2.65	35.31	6.92	16.5	77.92	
	R2-E	3.01	35.09	7.34	17.74	78.15	
	SS	3.34	34.57	8.11	17.14	80.21	
	R2	4.37	37.13	9.18	18.91	83.35	
CONLL	SS-E0	5.18	32.01	6.32	12.75	69.45	
	SS-E	3.41	35.16	6.89	16.18	78.39	
	R2-E	3.59	34.28	6.74	16.82	80.15	
	SS	5.09	33.85	6.88	16.52	79.83	
	R2	6.07	35.38	7.40	18.22	83.13	

views and is the only dataset that includes aspect prompts.

Licensing and Personal Data All datasets are distributed under an open Creative Commons license and compiled with explicit consent or sourced from materials with an open license. We attribute authors of the papers in our dataset while excluding personal and reviewer metadata.

5 Experiments

In the following section, we evaluate review quality, review specificity, and aspect coverage as key properties of the generated reviews. We provide extensive ablation experiments that identify how much each novel contribution of our approach contributes to improved performance. In particular, we compare REVIEWER2 against the following baselines:

- REVIEWER2-E: Following (Yuan et al., 2021), we apply a cross-entropy (CE) extraction method to extract a diverse set of sentences from the pa-

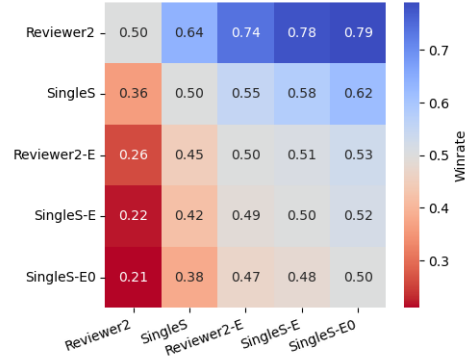


Figure 4: Pairwise winrates on faithfulness, coverage, coherence, and specificity among REVIEWER2 and baselines using GPT4 as a judge.

per to represent the content of the paper. The framework is the same as REVIEWER2 while we only use the extracted part instead of the full paper: $M_p^E : e \rightarrow \{x^1, \dots, x^k\}$, $M_r^E : (e, x) \rightarrow y$ where e is the extracted content from paper p . This ablation is used to evaluate the difference between using the full paper compared to an extractive summary.

- SINGLES: We fine-tune a single-stage model to directly generate reviews from the full context of the paper without an aspect prompt, $M_r^S : p \rightarrow y$. Prompts are neither used in fine-tuning nor inference. This ablation is designed to evaluate the effect of aspect prompts.
- SINGLES-E: This variant involves fine-tuning a single model to generate reviews only from extractive summaries of papers, $M_r^{SE} : e \rightarrow y$. This method aligns with commonly employed pipelines in previous papers and serves as a baseline representing the state-of-the-art.
- SINGLES-E0: This zero-shot approach prompt an LLM to generate a review from the extracted context directly without aspect prompts. This baseline evaluates the effect of fine-tuning.

We use SINGLES-E and SINGLES as proxies for the baseline methods proposed by (Yuan et al., 2021; Lin et al., 2023). Llama-2-70B-Chat (Touvron et al., 2023) is used as the instruction-following model for PGE and Llama-2-7B-Chat is used for REVIEWER2 and the single stage baselines. More experimental details are shown in Appendix F. We randomly select 80% of ICLR and NeurIPS papers for training, 10% for validation, and 10% for testing while using all the papers in other venues for testing. Since the other venues have review formats different from ICLR and NeurIPS, this allows us to test adaptability to different review formats.

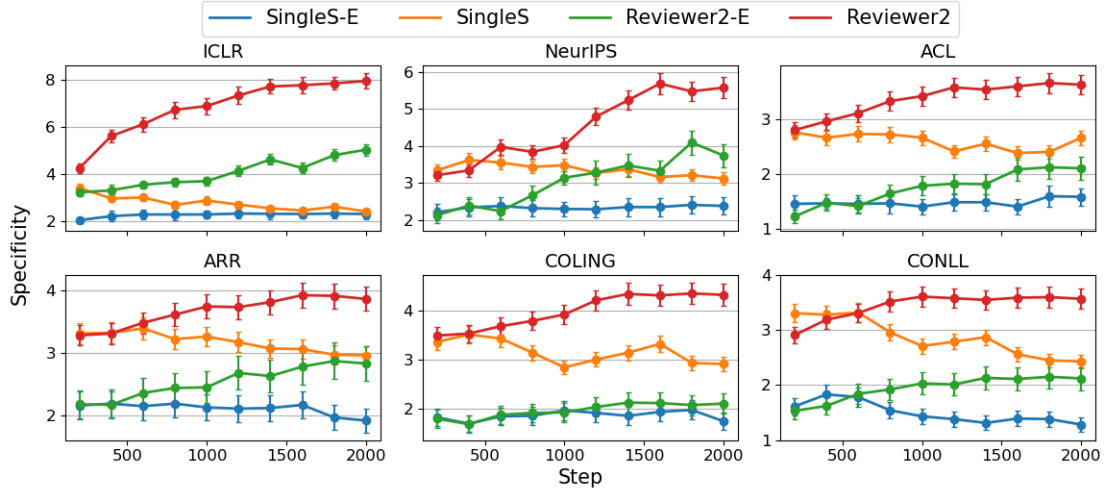


Figure 5: Specificity plots of four methods for 2000 steps across six venues.

5.1 Quality Analysis

To compare the generated reviews with the reference reviews, we employ three metrics: BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2003), and BertScore (Zhang et al., 2020). BLEU and ROUGE measure the n-gram similarity while BertScore measures the semantic similarity in the embedding space. Notably, there are several reference reviews for each paper. When computing BLEU, ROUGE, and BertScore, following (Yuan et al., 2021), we use the maximum value instead of an average since the generated reviews do not need to be closely aligned with all references, given that the reference reviews may focus on different aspects. To compare across the generated reviews, we prompt GPT4 to select the better review based on faithfulness, coverage, coherence, and specificity on 100 randomly sampled reviews for each method across six venues. More details of the evaluation is shown in Appendix G.

Result. Table 4 compares the performance of REVIEWER2 against several ablations and baselines. Overall, REVIEWER2 outperforms all methods across all metrics and datasets, demonstrating the effectiveness of leveraging both the full context of the paper and the aspect prompt. The comparisons between REVIEWER2 and SINGLES as well as REVIEWER2-E and SINGLES-E reveal consistent performance improvement through the two-stage approach. Furthermore, the comparison between REVIEWER2 and REVIEWER2-E shows that avoiding extractive summaries provides an additive benefit on top of using aspect prompts. On the cross-domain datasets (ACL, ARR, COLING, CONLL) we can observe a com-

parable BertScore with ICLR and NeurIPS using REVIEWER2, demonstrating the semantic adaptability of the method to domains that the methods was not trained on.

Figure 4 compares REVIEWER2 with the baselines on faithfulness, coverage, coherence, and specificity using GPT4 as a judge. REVIEWER2 consistently achieves higher winrates compared to the baselines, demonstrating the effectiveness of our method in producing high-quality reviews.

To further illustrate REVIEWER2, we included aspect prompts produced by M_p and a review produced by M_r in Appendix H.

5.2 Specificity Analysis

A highly specific review identifies specific issues of the given paper, and it does not look like a generic review that could apply to other papers. To formalize this into a concise metric, we measure the specificity of the review by calculating the drop in BertScore when pairing the review with the reference reviews of a different paper. A generated review with high specificity will lead to a large average drop, while a generic review will lead to a smaller drop. Formally, given papers P , reviews Y , and generated reviews $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m\}$, we define specificity (SPE \uparrow) as:

$$\text{SPE} = \frac{1}{m} \sum_{i=1}^m \max\{\text{sim}(\hat{y}_i, y_i^n) | 1 \leq n \leq n_i\} - \frac{1}{m-1} \sum_{j \neq i} \max\{\text{sim}(\hat{y}_i, y_j^n) | 1 \leq n \leq n_i\}$$

where $\text{sim}(a, b)$ denotes the BertScore between a and b and \hat{y}_j . We approximate the inner sum by Monte Carlo sampling $j \sim [1, m] \setminus i$.

Table 5: Effect of prompts for SINGLES (SS) and REVIEWER2 (R2) across six venues.

avg	SS	$\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{n=1}^{n_i} \text{sim}(M_r^S(p_i), y_i^n)$					
	R2	$\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i^2} \sum_{n=1}^{n_i} \sum_{k=1}^{n_i} \text{sim}(M_r(p_i, x_i^n), y_i^k)$					
max	SS	$\frac{1}{m} \sum_{i=1}^m \max_{1 \leq n \leq n_i} \{\text{sim}(M_r^S(p_i), y_i^n)\}$					
	R2	$\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{n=1}^{n_i} \max_{1 \leq k \leq n_i} \{\text{sim}(M_r(p_i, x_i^n), y_i^k)\}$					
	Method	ICLR	NeurIPS	ACL	ARR	COLING	CONLL
avg	SS	80.19	80.23	79.85	80.23	79.42	78.41
	R2	80.13	80.36	79.14	79.96	79.53	78.28
max	SS	81.18	81.83	80.25	81.03	80.21	79.83
	R2	83.63	83.41	83.54	82.51	83.19	83.32

Result. To obtain a reliable measure, we conducted ten random shuffles and calculated the average. The result is shown in Figure 5 along with the variance. For methods that do not make use of aspect prompts, SINGLES and SINGLES-E, the specificity drops with more training steps. This indicates that increased training without prompts leads to more generic reviews. For the methods that use prompts, REVIEWER2-E and REVIEWER2, the specificity consistently increases with a higher number of steps. Notably, the difference between REVIEWER2 and SINGLES is higher than the difference between REVIEWER2-E and SINGLES-E, suggesting that adding prompts on top of the full context leads to higher improvement comparing to adding to the extracted context.

5.3 Control Analysis

To assess how responsive REVIEWER2 is to the aspect prompts, we conduct experiments that compare REVIEWER2 and SINGLES. The M_r model in REVIEWER2 is given the prompts generated by *PGE*. We compute the average similarity of the generated review to the reference reviews for both methods as well as the maximum similarity. The detailed equations for the computations are shown in Table 5. BertScore is used for computing sim.

Result. REVIEWER2 and SINGLES have similar average similarity while REVIEWER2 has a higher maximum similarity across all six venues. This means that SINGLES generates reviews that are close to all the reference reviews, but that are not particularly close to any one of them. In contrast, REVIEWER2 is consistently able to generate reviews that closely match one of the references.

Table 6: Coverability (COV ↓) for REVIEWER2-E (R2-E) and REVIEWER2 (R2) across six venues.

Method	ICLR	NeurIPS	ACL	ARR	COLING	CONLL
R2-E	13.55	12.66	16.62	15.29	14.84	15.46
R2	4.22	3.99	3.23	2.91	5.09	4.25

5.4 Coverage Analysis

Finally, we evaluate whether authors can achieve good coverage through the choice of aspect prompts and the effect of different aspect prompts on generation. Since M_r and M_r^E are the only models that permit aspect prompts, we evaluate the effect of aspect prompts on coverage for these two models. Given papers P , reviews Y , prompts X , we define coverability (COV ↓) for M_r as:

$$\text{COV} = \frac{1}{m} \sum_{i=1}^m g_i - h_i$$

$$h_i = \frac{1}{n_i(n_i - 1)} \sum_{n=1}^{n_i} \sum_{\substack{k=1 \\ k \neq n}}^{n_i} \text{sim}(y_i^n, y_i^k)$$

$$g_i = \frac{1}{n_i(n_i - 1)} \sum_{n=1}^{n_i} \sum_{\substack{k=1 \\ k \neq n}}^{n_i} \text{sim}(M_r(p_i, x_i^n), M_r(p_i, x_i^k))$$

Here, h_i represents the pairwise similarity among the reference reviews for paper p_i while g_i is the pairwise similarity among generated reviews based on the *PGE* prompts in the dataset. The coverability for M_r^E is defined similarly but with e_i as input instead of p_i . We use BertScore to calculate the similarities. A high g_i indicates that the generated reviews are similar despite being generated from different prompts.

Result. The results are shown in Table 6. While perfectly reproducing the coverage of the human reviews would imply a value of 0, M_r exhibits significantly better coverage than M_r^E , demonstrating its effectiveness in generating tailored responses across diverse prompts for a given paper and the importance of using full context.

6 Conclusion

We propose a two-stage review generation framework that incorporates aspect prompts. Analyses of quality, specificity, and controllability indicate that our method can generate high-quality and specific reviews while being controllable based on the aspect prompt. Furthermore, we develop a new pipeline for annotating review datasets with aspect prompts, and we make this new dataset available.

7 Limitations

In this section, we discuss some of the limitations for *PGE* and REVIEWER2.

7.1 Disjoint Processes for Generation

Our current configuration first uses *PGE* to generate prompts and subsequently fine-tunes REVIEWER2 with the generated prompts. However, this approach leads to a disjointed process, where prompt generation operates independently of review generation, reducing the effectiveness of the generated prompts. Ideally, the generated prompts should assist alignment during fine-tuning. A possible extension is to integrate the two processes together and refine the generated prompts based on the review generation pipeline.

7.2 Input Inconsistency

The input to *PGE* consists of human-written reviews, while REVIEWER2 also incorporates papers. This distinction arises from the limitation of Llama-2-70B-Chat, which only has a context length of 4,096. Although GPT-4 (OpenAI, 2023) supports up to 32,000 context length, the associated cost is high since the average context length of the papers is 6,229. The potential improvement in performance may not be worth the increased cost.

7.3 Limited Domain Knowledge

Currently, REVIEWER2 relies on its pre-trained corpus, assuming that the language model used has adequate domain knowledge. This approach might produce inaccurate reviews for papers that demand substantial in-domain expertise. A potential future work could investigate the effectiveness of second-stage pre-training or domain adaptation using the paper corpus.

8 Ethics

Automatic review generation is a complex task and bears a wide range of risks. It is crucial to emphasize that the ongoing efforts in this field are not designed to replace human reviewers; instead, they function as a valuable tool for authors and a guiding resource for human reviewers. This research is an exploratory work within this domain, and it is important to stress that the outcomes produced by the models should not be misconstrued as definitive and authentic reviews of the respective papers. In utilizing datasets, we adhere to the intended purposes outlined in previous works. The

datasets we released offer many possibilities for advancing research in NLP, including but not limited to review generation, instruction following, and self-alignment.

601
602
603
604

References

- 605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
- Adrian G Barnett, Jolieke C van der Pols, and Annette J Dobson. 2004. [Regression to the mean: what it is and how to deal with it](#). *International Journal of Epidemiology*, 34(1):215–220.
- Bo-Christer Björk and David Solomon. 2013. [The publishing delay in scholarly peer-reviewed journals](#). *Journal of Informetrics*, 7(4):914–923.
- Lutz Bornmann and Ruediger Mutz. 2014. [Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references](#). *Preprint*, arXiv:1402.4578.
- Yen-Chun Chen and Mohit Bansal. 2018. [Fast abstractive summarization with reinforce-selected sentence rewriting](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.
- Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. [Longlora: Efficient fine-tuning of long-context large language models](#). *Preprint*, arXiv:2309.12307.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. [Specter: Document-level representation learning using citation-informed transformers](#). *Preprint*, arXiv:2004.07180.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. [A survey on in-context learning](#). *Preprint*, arXiv:2301.00234.
- Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. [GSum: A general framework for guided neural abstractive summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.
- Nils Dycke, Iliia Kuznetsov, and Iryna Gurevych. 2023. [NLPeer: A unified resource for the computational study of peer review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.
- Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. [Academic plagiarism detection: A systematic literature review](#). *ACM Comput. Surv.*, 52(6).
- Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. [Bottom-up abstractive summarization](#). *Preprint*, arXiv:1808.10792.
- Tirthankar Ghosal, Ravi Sonam, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2019. [Is the paper within scope? are you fishing in the right pond? In 2019 ACM/IEEE Joint Conference on Digital Libraries \(JCDL\)](#), pages 237–240.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. [Unnatural instructions: Tuning language models with \(almost\) no human labor](#). *Preprint*, arXiv:2212.09689.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018a. [A dataset of peer reviews \(PeerRead\): Collection, insights and NLP applications](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018b. [A dataset of peer reviews \(peerread\): Collection, insights and nlp applications](#). *Preprint*, arXiv:1804.09635.
- Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. [Peer review in scientific publications: Benefits, critiques, & a survival guide](#). *EJIFCC*, page 227–243.
- Ari Kobren, Barna Saha, and Andrew McCallum. 2019. [Paper matching with local fairness constraints](#). *Preprint*, arXiv:1905.11924.
- Diego Kozlowski, Vincent Larivière, Cassidy R. Sugimoto, and Thema Monroe-White. 2022. [Intersectional inequalities in science](#). *Proceedings of the National Academy of Sciences*, 119(2):e2113067119.
- Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2012. [Bias in peer review](#). *ASIS&T*.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. [Self-alignment with instruction back-translation](#). *Preprint*, arXiv:2308.06259.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. [Can large language models provide useful feedback on research papers? a large-scale empirical analysis](#). *Preprint*, arXiv:2310.01783.
- Chin-Yew Lin and Eduard Hovy. 2003. [Automatic evaluation of summaries using n-gram co-occurrence statistics](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

715	Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Moprdr: A multidisciplinary open peer review dataset . <i>Neural Computing and Applications</i> , 35(34):24191–24206.	772
716		773
717		774
718		
719	Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing . <i>Preprint</i> , arXiv:2306.00622.	775
720		776
721		777
722	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback . <i>Preprint</i> , arXiv:2303.17651.	778
723		779
724		780
725		
726		
727		
728		
729		
730	Robert K. Merton. 1968. The matthew effect in science . <i>Science</i> , 159(3810):56–63.	
731		
732	Mathias Wullum Nielsen and Jens Peter Andersen. 2021. Global citation inequality is on the rise . <i>Proceedings of the National Academy of Sciences</i> , 118(7):e2012208118.	
733		
734		
735		
736	OpenAI. 2023. Gpt-4 technical report . <i>Preprint</i> , arXiv:2303.08774.	
737		
738	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback . <i>Preprint</i> , arXiv:2203.02155.	
739		
740		
741		
742		
743		
744		
745		
746	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation . In <i>Proceedings of the 40th Annual Meeting on Association for Computational Linguistics</i> , ACL '02, page 311–318, USA. Association for Computational Linguistics.	
747		
748		
749		
750		
751		
752	Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters . In <i>Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining</i> , KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.	
753		
754		
755		
756		
757		
758		
759		
760	Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao,	
761		
762		
763		
764		
765		
766		
767		
768		
769		
770		
771		
	Thomas Wolf, and Alexander M. Rush. 2022. Multi-task prompted training enables zero-shot task generalization . <i>Preprint</i> , arXiv:2110.08207.	772
		773
		774
	Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. MReD: A meta-review dataset for structure-controllable text generation . In <i>Findings of the Association for Computational Linguistics: ACL 2022</i> , pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.	775
		776
		777
		778
		779
		780
	Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. Peerreview4all: Fair and accurate reviewer assignment in peer review . <i>Preprint</i> , arXiv:1806.06237.	781
		782
		783
	Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision . <i>Preprint</i> , arXiv:2305.03047.	784
		785
		786
		787
		788
		789
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models . <i>Preprint</i> , arXiv:2307.09288.	790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800
		801
		802
		803
		804
		805
		806
		807
		808
		809
		810
		811
		812
	Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne . <i>Journal of Machine Learning Research</i> , 9(86):2579–2605.	813
		814
		815
	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khoshabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions . <i>Preprint</i> , arXiv:2212.10560.	816
		817
		818
		819
		820
	Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners . <i>Preprint</i> , arXiv:2109.01652.	821
		822
		823
		824
		825
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and	826
		827

- 828 Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*,
829 arXiv:2201.11903.
830
- 831 Wenting Xiong and Diane Litman. 2011. [Automatically predicting peer-review helpfulness](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, Portland, Oregon, USA. Association for Computational Linguistics.
832
833
834
835
836
- 837 Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. [Can we automate scientific reviewing?](#) *Preprint*,
838 arXiv:2102.00176.
839
- 840 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*,
841 arXiv:1904.09675.
842
843

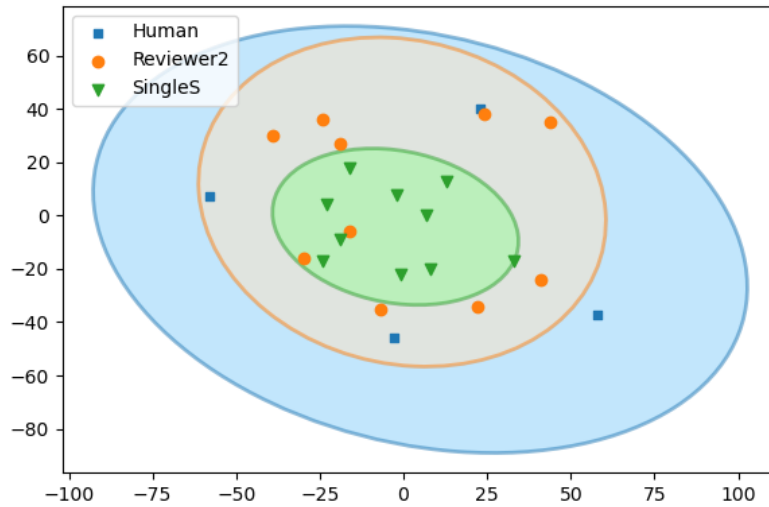


Figure 6: Embedding visualization of the reviews from human, REVIEWER2, and SINGLES. The eclipse represents the confidence interval with 2 standard deviations.

To justify the intuition of our illustration in Figure 2, we visualize the text embeddings of actual generations from REVIEWER2 and SINGLES. For a given paper, we use REVIEWER2 and SINGLES to generate 10 reviews respectively. Then, we use SFR-Embedding-Mistral⁴ as the embedding model to embed the text and visualize the embeddings using t-SNE (van der Maaten and Hinton, 2008). The SINGLES generations have a much smaller coverage and reside around the middle of the space spanned by the human-written reviews, while generations from REVIEWER2 are more dispersed and closer to the human references. This plot aligns closely with the intuition we convey in Figure 2 (b) and (c).

846

847

848

849

850

851

852

⁴Huggingface Model Card: Salesforce/SFR-Embedding-Mistral

B Prompts for *PGE*

Prompt for Generation

[INST] «SYS» You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information. «/SYS»

Analyzing the provided review, identify a set of questions that the reviewer is attempting to address regarding the paper without being too specific.

Here are some examples:

Review:

[SAMPLED REVIEW FROM *S*]

Questions to address:

[SAMPLED PROMPT FROM *S*]

Review:

[SAMPLED REVIEW FROM *S*]

Questions to address:

[SAMPLED PROMPT FROM *S*]

Review:

[SAMPLED REVIEW FROM *S*]

Questions to address:

[SAMPLED PROMPT FROM *S*]

Review:

[REVIEW FOR GENERATION]

Questions to address:[/INST]

Prompt for Evaluation

[INST] «SYS» You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information. «/SYS»

Below is a set of questions and a candidate answer. Evaluate the quality of the questions. Are the questions a good match to the candidate answer? Please assign a score using the following 5-point scale:

- 1: This score indicates that the response deviates significantly from the instruction, providing information or addressing aspects that were not required or specified.
- 2: This score suggests that the response is limited in scope, focusing on a small subset of the questions posed in the instruction. It does not comprehensively cover the entire set of questions.
- 3: This score indicates that the response covers a substantial portion of the questions outlined in the instruction but falls short of addressing all of them. It suggests a moderate level of completeness.
- 4: This score indicates that the response covers most of the questions. However, there is some irrelevant information in the answer that is not asked by any of the questions.
- 5: This score indicates that the response is comprehensive, addressing all questions in the instruction without any irrelevant information.

Here are some examples:

Questions:

[EXAMPLE PROMPT]

Answer:

[EXAMPLE REVIEW]

Assessment:

[EXAMPLE ASSESSMENT]

Score: [EXAMPLE SCORE]

Questions:

[EXAMPLE PROMPT]

Answer:

[EXAMPLE REVIEW]

Assessment:

[EXAMPLE ASSESSMENT]

Score: [EXAMPLE SCORE]

Questions:

[EXAMPLE PROMPT]

Answer:

[EXAMPLE REVIEW]

Assessment:

[EXAMPLE ASSESSMENT]

Score: [EXAMPLE SCORE]

Questions:

[PROMPT FOR EVALUATION]

Answer:

[REVIEW FOR EVALUATION]

Assessment:[/INST]

C Example Review-Prompt Pair

Review	<p>Summary Of The Paper</p> <p>This paper introduces neural matching fields into semantic correspondence. To the best my knowledge, this approach should be the first method to do the task using implicit neural representation. There are two problems: the computation for 4D matching field and the inference efficiency. Authors provide effect method to address the two problems.</p> <p>Strengths And Weaknesses</p> <p>This paper employs implicit neural representation to do semantic correspondence. This should be the major contribution. According to the statement of authors, I can follow the idea easily and this idea should work. The disadvantage of this work is the experiments. There are too many quantitative comparisons. According to the data, the performance of this method seems OK. However, authors should provide more visual experiments to convince readers.</p> <p>Questions</p> <p>I only have one concern. Traditional Implicit Neural Representation method such as LIIF and NeRF records images into the weights of neural network. One neural network represents one image or one scene. Does NeMF take a neural network to represent a semantic correspondence or a matching cost. If so, how much time will your method cost to train a network? If not so, what is the difference between your method and other semantic correspondence methods.</p> <p>Limitations</p> <p>According to my understand, NeMF takes a network to represent a matching cost. In practice, people need a method to compute different matching cost for different image pairs. How does NeMF to deal with this situation.</p>
--------	---

Questions to address	<ol style="list-style-type: none"> 1. What is the focus and contribution of the paper on semantic correspondence? 2. What are the strengths of the proposed approach in terms of neural representation? 3. What are the weaknesses for the experiment section? 4. Do you have any concerns on the semantic correspondence representation? 5. What are the limitations regarding the NeMF approach on matching cost representation?
----------------------	---

D In-Context Example for Evaluation

Questions	<ol style="list-style-type: none"> 1. What is the main contribution of the paper on dictionary learning? 2. What are the strengths of the paper in the theoretical analysis? 3. Do you have any questions regarding the assumptions, theorems, and algorithm of the paper? 4. Could you access the reproducibility of the paper?
Answer	<p>The paper proposes an alternating minimization algorithm for dictionary learning, and theoretical guarantees are also given. In each step the algorithm first uses an l_1, l_2 and l_∞ algorithm with thresholding to get an estimate of the coefficients, and then use another gradient step to update the dictionary.</p> <p>To me two shining points of the paper:</p> <ol style="list-style-type: none"> 1. Guarantee holds for the overcomplete dictionary. 2. Improved the sparsity level requirement by a factor of $\log d$. <p>Obviously the NIPS format is too short for the arguments the authors are making, and a lot of details are moved to the appendix. Due to time limit I cannot read all the details of the proof. Below are some questions:</p> <ol style="list-style-type: none"> 1. In A1 you have a μ-incoherence assumption, but μ is not shown in your theorem 3. Is it hidden somewhere? 2. In assumption B1 you mentioned, and I agree that there is a fast random initialization so that the condition holds. Can you give some details about your initialization procedure and guarantees? 3. How do you handle the permutation invariance of A? 4. In your algorithm 1, line 3, the MUS algorithm has a return, but in your definition (equation 2), the return is not specified. Actually the returned should be θ instead of (θ, t, u). 5. “w_k^t is the k^{th} covariate at step t”? Why w_k^t is called the k^{th} covariate? 6. Any simulation result verifying your convergence rate?
Assessment	<p>The answer addresses the first question by summarizing the main contribution of the paper. For the second question, the answer gives two strong points of the paper in its theoretical justifications. The answer address the third question by providing six different questions converging the assumptions, theorems, and the algorithm of the paper. However, the answer fails to address the fourth question.</p>
Score	<p>Since the answer fails to address all of the questions, it receives a score of 3.</p>

856	E Dataset Details	
857	E.1 Paper Contents	
858	• title: title of the paper	
859	• authors: list of author names	
860	• emails: list of author emails	
861	• sections: list of sections of the paper	
862	– heading: heading of the section	
863	– text: text of the section	
864	• references: list of references of the paper	
865	– title: title of the reference	
866	– author: list of author names of the reference	
867	– venue: venue of the reference	
868	– citeRegex: citation expression	
869	– shortCiteRegex: short citation expression	
870	– year: publication year of the reference	
871	• referenceMentions: the location of the reference	
872	in the paper	
873	– referenceID: numerical reference id	
874	– context: context of the reference in the pa-	
875	– per	
876	– startOffset: start index of the context	
877	– endOffset: end index of the context	
878	• year: year of publication	
879	• abstractText: abstract of the paper	
880	E.2 Metadata Contents	
881	• id: unique id of the paper	
882	• conference: venue for the paper	
883	• decision: final decision for the paper (ac-	
884	cept/reject)	
885	• url: link to the PDF of the paper	
886	• review_url: link to the review of the paper	
887	• title: title of the paper	
888	• authors: list of the authors of the paper	

F Experimental Details	889
REVIEWER2 and SINGLES have a context length	890
of 32,768 while other models have a 4,096 context	891
length. All of the models excluding SINGLES-E0	892
are fine-tuned with 8 A6000 GPUs using Deep-	893
Speed (Rasley et al., 2020) stage 2, batch size	894
64, gradient accumulation 8, and warm-up steps	895
100 for 2 epochs. We use the AdamW opti-	896
mizer with a learning rate $1e - 5$ searched from	897
$[5e - 6, 1e - 5, 2e - 5, 5e - 5, 1e - 4, 2e - 4]$.	898
We perform supervised fine-tuning (SFT) for RE-	899
VIEWER2 and each baseline methods that require	900
training. The setup is detailed below:	901
• REVIEWER2	902
– Prompt generation model (M_p)	903
Model: Llama-2-7B-Chat	904
Input: Full paper	905
Output: Aspect prompts	906
– Review generation model (M_r)	907
Model: Llama-2-7B-Chat	908
Input: Full paper and aspect prompt	909
Output: Review	910
• REVIEWER2-E	911
– Prompt generation model (M_p^E)	912
Model: Llama-2-7B-Chat	913
Input: Extracted sentences from the pa-	914
– per	915
Output: Aspect prompts	916
– Review generation model (M_r^E)	917
Model: Llama-2-7B-Chat	918
Input: Extracted sentences from the pa-	919
– per and aspect prompt	920
Output: Review	921
• SINGLES	922
– Review generation model (M_p^S)	923
Model: Llama-2-7B-Chat	924
Input: Full paper	925
Output: Review	926
• SINGLES-E	927
– Review generation model (M_p^{SE})	928
Model: Llama-2-7B-Chat	929
Input: Extracted sentences from the pa-	930
– per	931
Output: Review	932

G Winrate Details

933

We are using gpt-4-0613 checkpoint for winrate computations.

934

Prompt for Winrate

Which of the following reviews provide a more accurate review of the given paper? Judge based on faithfulness, coverage, coherence, and specificity.

Paper:

{{Paper}}

Review A:

{{reviewa}}

Review B:

{{reviewb}}

Instructions:

FIRST provide a comparison of the two reviews, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice. Your response should use the format:

Comparison: <comparison and explanation >

Preferred: <"A" or "B">

H Generated Aspect Prompts and Review for this Paper

Aspect	What is the focus and contribution of the paper on peer review generation?
Prompts	<p>What are the strengths of the proposed approach, particularly in terms of its two-stage architecture?</p> <p>What are the weaknesses of the paper regarding its experimental setup and proposed metrics?</p> <p>Do you have any concerns or suggestions for improving the paper’s clarity and organization?</p>

Review	<p>Summary Of the Paper. This paper proposes a two-stage review generation framework that incorporates aspect prompts. They design a prompt generation with evaluation (PGE) pipeline to generate a variety of high-quality aspect prompts, and they make this new dataset available. They evaluate review quality, specificity, and aspect coverage as key properties of the generated reviews. They provide extensive ablation experiments that identify how far each novel contributions in their approach contributes to improved performance. Finally, they conclude by highlighting the advantages of leveraging both the full context of the paper and the aspect prompt.</p> <p>Strengths. The paper clearly states the problem of existing methods that only generate generic reviews and lacks specificity and control. The authors propose a two-stage architecture that explicitly models the distribution of possible aspects that the review may address. They show that this leads to more detailed reviews that better cover the range of aspects that human reviewers identify in the draft.</p> <p>Weaknesses. The authors evaluate the effectiveness of the proposed method with several baselines. However, the baselines are not very convincing. The paper lacks clear motivation for why they use a two-stage architecture. The paper needs to provide more details on how they evaluate the generated prompts.</p>
--------	---
