A Sentiment Consolidation Framework for Meta-Review Generation

Miao Li[1] **Jey Han Lau**[1] **Eduard Hovy**[1,2]
[1]School of Computing and Information Systems, The University of Melbourne
[2]Language Technologies Institute, Carnegie Mellon University
miao4@student.unimelb.edu.au,
{laujh, eduard.hovy}@unimelb.edu.au

Abstract

Modern natural language generation systems with Large Language Models (LLMs) exhibit the capability to generate a plausible summary of multiple documents; however, it is uncertain if they truly possess the capability of information consolidation to generate summaries, especially on documents with opinionated information. We focus on meta-review generation, a form of sentiment summarisation for the scientific domain. To make scientific sentiment summarization more grounded, we hypothesize that human meta-reviewers follow a three-layer framework of sentiment consolidation to write meta-reviews. Based on the framework, we propose novel prompting methods for LLMs to generate meta-reviews and evaluation metrics to assess the quality of generated meta-reviews. Our framework is validated empirically as we find that prompting LLMs based on the framework — compared with prompting them with simple instructions — generates better meta-reviews.[1]

1 Introduction

Notable strides have been made in abstractive text summarization (El-Kassas et al., 2021) with the advancement of Large Language Models (LLMs) (Zhao et al., 2023) over recent years. With even a simple instruction such as "*tl;dr*" or "*please write a summary*", these models can generate plausible summaries which are found more preferred over those written by humans (Pu et al., 2023). However, it is uncertain if these models truly possess the ability of information consolidation, especially when summarizing documents that are composed of opinionated information. The models may take shortcuts to generate texts instead of correctly understanding and aggregating information from the source documents (Gehrmann et al., 2023) and they may generate abstractive summaries with incorrect overall sentiment.

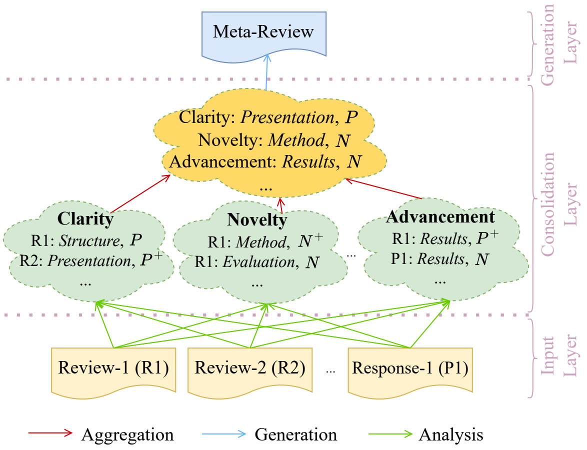


Figure 1: The three-layer framework of the underlying information consolidation logic in meta-reviewing (P: Positive, P^+: Strongly positive, N: Negative, N^+: Strongly negative).

Automated sentiment summarization holds significant importance (Kim et al., 2011) and there have been sentiment summarization datasets; however, most of them are in the product review domain. These datasets are less interesting for investigating information consolidation as (1) the summaries are extractive, composed of a simple combination of extracted snippets (Amplayo et al., 2021), and (2) the summary of product reviews is about extracting the majority sentiment (which is a simple consolidation function). To address this, in this paper, we propose the task of scientific sentiment summarization, taking the meta-reviews in scientific peer review as summaries.[2] The investigation of meta-review generation (Li et al., 2023a) presents an exciting opportunity for exploring the intricate process of multi-document information consolidation that involves complex judgement.

[1] The code and annotated data are accessible at https://github.com/oaimli/MetaReviewingLogic.

[2] The representative peer review platform which is publicly available is www.openreview.com.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10158–10177
August 11-16, 2024 ©2024 Association for Computational Linguistics

This is because (1) meta-reviewers are supposed to understand not only all the reviews from different reviewers but also the multi-turn discussions between the reviewers and the author and write their comments to support the acceptance decision of the manuscript, (2) the logic of arguments (from reviewers and authors) has to be taken into account to arrive at the final sentiment in the meta-reviews and it is not a matter of majority voting and (3) meta-reviews have to recognize and resolve conflicts and consensus among reviewers.

In this paper, we hypothesize that human meta-reviewers follow a three-layer sentiment consolidation framework as shown in Figure 1 to write meta-reviews based on reviews and multi-turn discussions in the peer review process. Human and automatic annotation is then conducted to extract sentiments and expressions on various review facets (e.g., novelty and soundness) from corresponding source documents (i.e., reviews and discussions) and these judgements play a critical role in generating the meta-reviews. We also propose two evaluation metrics which focus on assessing sentiments in generated meta-reviews, and experiments empirically validate our proposed three-layer framework when they are integrated as prompts for LLMs to generate meta-reviews.

Contributions of our paper:

- We hypothesize that human meta-reviewers follow a three-layer sentiment consolidation framework when writing meta-reviews;
- We collect human annotations on meta-reviews and corresponding source documents based on the consolidation framework;
- We propose two automatic metrics (reference-free and reference-based) to evaluate the sentiment in the generated meta-reviews.
- Experiments validate the empirical effectiveness of the framework when we incorporate it as prompts for LLMs to generate meta-reviews.

2 Related Work

In this section, we discuss large-scale information consolidation in abstractive summarization, and automated sentiment summarization.

2.1 Large-Scale Information Consolidation

Natural language generation systems are expected to not only have high-quality generations but also have the ability to comprehend the input information, especially for conditional text generation such as multi-document summarization which has to integrate and aggregate information from different source documents (Gehrmann et al., 2023). Most work in the text summarization community only attempts to improve the generation quality of text summarization, such as relevance and faithfulness, without considering the intricate generation process (Phang et al., 2022; El-Kassas et al., 2021; Xiao et al., 2022). For example, Li et al. (2023b) use heterogeneous graphs to represent source documents and borrow the idea of graph compression to train the summarization model to get improvement of the generated summaries. However, it is uncertain if these models truly possess the ability to consolidate information from different source documents.

2.2 Automated Sentiment Summarization

Sentiment summarization aims to summarise the overall sentiment given a set of documents (Hossain et al., 2023). However, most datasets for sentiment summarization are in the product review domain (Amplayo et al., 2021), and scientific sentiment summarization is under-explored. Meta-review generation, which is a typical scenario of scientific sentiment summarization, is to automatically generate meta-reviews based on reviews and the multi-turn discussions between reviewers and the author of the corresponding manuscript (Li et al., 2023a). It is mostly modelled as an end-to-end task (Bhatia et al., 2020; Wu et al., 2022; Shen et al., 2022; Chan et al., 2020). Although Li et al. (2023a) considered the conversational structure of reviews and discussions, their models do not explain how human meta-reviewers write the meta-reviews. Different from investigating checklist-guided iterative introspection for meta-review generation with prompting (Zeng et al., 2024), our work is based on a three-layer sentiment consolidation framework and focuses on various review facets, and we explicitly investigate the sentiment fusion process which is arguably an important aspect of meta-review generation.

3 Sentiment Consolidation Over Multiple Opinionated Documents

In the following section, we introduce the task of scientific sentiment summarization and our three-layer sentiment consolidation framework in meta-review generation, conduct sentiment and expres-

Component	**Definition**
Content Expression	What the sentiment is talking about
Sentiment Expression	The value of the sentiment
Review Facet	The specific review facet that the judgement belongs to
Sentiment Level	The polarity and strength of the sentiment
Convincingness Level	How well the sentiment is justified in the document

Table 1: Definitions of components in a judgement.

	Min	**Max**	**Average**
#Documents/Sample	5	30	12.4
#Words/Sample	1,541	11,901	4,260.9
#Words/Source document	10	1,562	360.5
#Words/Meta-review	16	648	150.9

Table 2: Statistics of the human annotated data.

sion extraction, and analyze the fusion process of scientific sentiments.

3.1 Hierarchical Sentiment Consolidation

The task is meta-review generation. We use the PeerSum[3] dataset where the input is reviews and discussions and the target output is the corresponding human-written meta-review. We should clarify that even though the task is to generate meta-reviews, our focus here is to get the overall sentiment in the meta-reviews to be correct. Our method and evaluation reflect this focus.

Reading the reviewer guidelines from popular academic presses such as ACM and IEEE[4], we find they are mostly about *judgements* on the quality and merit of the manuscript. These judgements are generally based on six review facets of criteria: *Novelty*, *Soundness*, *Clarity*, *Advancement*, *Compliance* and *Overall quality*. The meta-reviewers must form their final opinion based on these judgements from the reviewers and authors. Looking at the meta-reviewer guidelines for ICLR[5] and NeurIPS[6], it recommends the meta-reviewer to understand and aggregate information from the whole peer-reviewing process. That is, a human meta-reviewer should first identify judgements from reviews and discussions, and then consolidate these opinions from different review facets to write their meta-review.

To conceptualize this, we propose a three-layer framework, as shown in Figure 1. The three layers include the input layer, the consolidation layer, and the generation layer. The input layer is the input documents of different types: official reviews and multi-turn discussions. The consolidation layer represents how meta-reviewers process the documents: they first identify and extract judgements from different documents, reorganize the judgements based on review facets, and then consolidate the opinions to form the final opinions of each review facet. In the generation layer, the meta-reviewer writes the meta-review to express the final opinions that they have developed from the previous layer.

3.2 Judgement Identification and Extraction

Judgements lay the foundation of our proposed framework and the whole peer review process. A judgement here expresses sentiment on a review facet and it contains several components: Content Expression, Sentiment Expression, Review Facet, Sentiment Level, and Convincingness Level (definitions are shown in Table 1, and an example is given in Appendix Figure 5). To automate judgement identification and extraction, we first conduct human annotation, and then leverage in-context learning of LLMs to perform more (automatic) annotation.

In human annotation, there are three types of documents including meta-reviews, official reviews, and discussions (the same definition used in Li et al. (2023a)) to be annotated. We recruit two annotators[7] to do this annotation (annotation instructions and design are detailed in Appendix B). 30 samples (i.e., one sample = one meta-review and its corresponding reviews and discussions) are annotated[8],

[3] https://github.com/oaimli/PeerSum

[4] The complete table of official guidelines that we consider is in Appendix A.

[5] https://iclr.cc/Conferences/2024/SACguide

[6] https://nips.cc/Conferences/2020/PaperInformation/AC-SACGuidelines

[7] The two annotators are senior PhD students who are familiar with the peer-review process.

[8] Annotating one sample takes about one hour on average and it costs about 60 hours and 2,100 US dollars in total.

and in total, we have 1,812 and 1,744 judgements from the two annotators. The statistics of these 30 samples are presented in Table 2. We present the agreement of the two annotators in Figure 2.[9] Generally, we see a moderate to high agreement, suggesting that the annotation task is robust and reproducible.

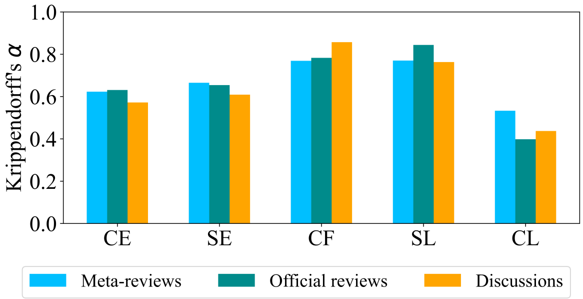


Figure 2: Inter-annotator agreement on meta-reviews, official reviews and discussions in terms of Krippendorff's α for different judgement components including Content Expression (CE), Sentiment Expression (SE), Review Facet (RF), Sentiment Level (SL), and Convincingness Level (CL).

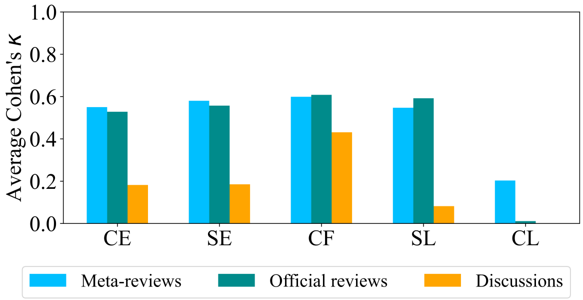


Figure 3: The averaged GPT-4's agreement with two human annotators on meta-reviews, official reviews and discussions in terms of Krippendorff's α for different judgement components including Content Expression (CE), Sentiment Expression (SE), Review Facet (RF), Sentiment Level (SL), and Convincingness Level (CL).

To get more annotated judgements for further experiments and analysis and investigate whether LLMs can be prompted to identify and extract judgements, we split the annotation task into two sub-tasks, extracting content and sentiment expressions and predicting other components of judgements, and use GPT-4 (OpenAI, 2023) with in-context learning (see full prompts in Appendix D and E respectively for the two sub-tasks).[10] We present the average agreement of GPT-4 with the two human annotators in Figure 3.[11] We can see GPT-4 has a moderate agreement with human annotators for meta-reviews and official reviews, but a low agreement for discussions. We suspect this may be because the discussions often contain rebuttals which have a different language to reviews or meta-reviews and extracting judgements from them may be more difficult. Interestingly, we also see that GPT-4 has a poor agreement in terms of convincingness (Figure 3), although the human inter-annotator agreement isn't strong in the first place (Figure 2). These observations suggest convincingness is perhaps a subjective assessment.

Facets	%Judgements	%Documents
Advancement	0.2545	0.8000
Soundness	0.2786	0.7833
Novelty	0.1817	0.6833
Overall	0.1414	0.5833
Clarity	0.1264	0.4500
Compliance	0.0174	0.0667

Table 3: Frequency of different review facets in meta-review judgements and meta-review documents.

3.3 Sentiment Fusion for Consolidation

With all the annotated judgements extracted by humans and GPT-4, we next dive more into the process of sentiment aggregation. Among all the review facets, we find that *Soundness* and *Advancement* are the two most important review facets when the meta-reviewers write their meta-reviews, while *Compliance* is rarely an issue in meta-reviews (shown in Table 3). This is consistent with our understanding of the peer-reviewing process.

More importantly, we find that human meta-reviewers do not always follow the majority review sentiment. We find that in PeerSum there are 23.7% samples where the meta-reviewer's acceptance decision is not consistent with the prediction based on majority voting by review ratings (a sample is defined as consistent when the number of reviews whose rating ≥ 5 is larger than the number of reviews whose rating < 5 and the final decision is *Accept*). We present an example in Table 4 where the meta-review does not follow the majority view on *Novelty* from the reviews.

[9] Calculation details and more results in terms of both Cohen's κ and Krippendorff's α are in Table 11, Table 12 and Table 13 in Appendix C.

[10] The version of GPT-4 we use is gpt-4-0613.

[11] More agreement results are in Table 14, Table 15 and Table 16 in Appendix C.

Human-written meta-review sentiment sentence
"Although each module in the proposed approach is **not novel**, it seems that the way they are used to address the specific problem of explainability and especially in text games is **novel** and sound."
All corresponding sentiment texts on Novelty in source reviews and discussions
"The generation of temporally extended explanations consists of a cascade of different components, **either straightfoward statistics or prior work**." "The novelty is **a bit low**." "overall novelty is **limitted**" "We contend that all steps are **individually novel as well as their combination**." "we are **the first** to use knowledge graph attention-based attribution to explain actions in such grounded environments"

Table 4: The example of a meta-review sentiment on *Novelty* which is not following majority voting of sentiments in source documents. The **green** and **red** texts indicate positive and negative sentiments, respectively.

Review Facets	**Judgements**	**Full Texts**
Advancement	0.677	**0.697**
Soundness	**0.684**	0.667
Novelty	**0.700**	0.650
Overall	**0.643**	0.631
Clarity	**0.712**	0.645
Compliance	0.555	**0.593**

Table 5: Accuracy of GPT-4 in predicting the sentiment levels in meta-reviews for each facet, using either only the annotated judgements ("Judgements") or the full text ("Full Texts") from the source documents.

To understand how well the judgements from the source documents (i.e., reviews and discussions) predict the overall sentiments in the meta-reviews for each review facet, we next formulate a text classification task where the output is the sentiment level of a content expression for a review facet in the meta-review, and the input is either: (1) the annotated judgements for the facet from the source documents; or (2) the full text of the source documents. We (zero-shot) prompt GPT-4 (full prompt detailed in Appendix F) with either input to predict 100 randomly sampled human-annotated instances and present the results in Table 5. Using judgements only as input, we see that it works better in 4 out of 6 facets — this preliminary result suggests our framework of extracting these judgements as an intermediate step may help generate better meta-reviews.

4 Sentiment-Aware Evaluation on Information Consolidation

In this section, we focus more on how to evaluate the sentiments of the generated summaries or meta-reviews in meta-review generation based on our proposed framework. We propose FacetEval and FusionEval which are reference-based and reference-free metrics, respectively.

4.1 Measuring Sentiment Similarity to Human-Written Meta-Review

To assess the quality of generated meta-reviews, we propose a reference-based evaluation metric, FacetEval, measuring the sentiment consistency c between the generated meta-review and the corresponding human-written meta-review in all review facets. Different from the generic evaluation metrics for abstractive summarization or text generation which mostly adopt surface-form matching, we focus more on review facets and their corresponding sentiment levels.

Specifically, we use the distribution of sentiments in all review facets to represent the meta-review and use the cosine similarity of the two vectors as the final score s.

$$s = \cos(\boldsymbol{m}_h, \boldsymbol{m}_g) \tag{1}$$

$$\boldsymbol{m} = \|_f [P_f^+, P_f, N_f^+, N_f, O_f] \tag{2}$$

where $\|$ denotes concatenation of representations for different facets, $\boldsymbol{m}_h$ and $\boldsymbol{m}_g$ are representations of the human-written and model-generated meta-reviews respectively. The representation $\boldsymbol{m}$ of the meta-review is the concatenation of vector representations of all review facets. Each facet of the document is represented by the frequency of different sentiment levels on the facet. The facet f is represented by a five-dimension vector $[P_f^+, P_f, N_f^+, N_f, O_f]$ where P_f^+ denotes the frequency of *Strongly positive* for the facet f, P_f the frequency of *Positive*, N_f^+ the frequency of *Strongly negative*, N_f the frequency of *Negative*, and O_f whether this facet is involved in the document. All the sentiments are obtained with GPT-4 following in-context learning in Section 3.2.

Following the similarity of meta-reviews, we could also calculate sentiment consistency among official reviews. Specifically, for every two official reviews i and j, the consistency in the facet f is the cosine similarity between two vector represen-

Review Facet	w/ Conflicts	w/o Conflicts
Advancement	0.463 (0.135)	0.551 (0.137)
Soundness	0.526 (0.158)	0.501 (0.110)
Novelty	0.300 (0.159)	0.357 (0.168)
Overall	0.433 (0.147)	0.597 (0.172)
Clarity	0.317 (0.133)	0.337 (0.145)
Compliance	0.827 (0.071)	0.771 (0.118)

Table 6: Sentiment consistency among different official reviews. (Variances are in the brackets.)

tations of documents.

$$c_{ij}^f = \cos\left(\boldsymbol{d}_i, \boldsymbol{d}_j\right) \tag{3}$$

where $\boldsymbol{d}^f = [P_f^+, P_f, N_f^+, N_f, O_f]$. Results shown in Table 6 suggest that different reviews are consistent in the sentiment to *Compliance* while there is much lower consistency in *Clarity* and *Novelty*. Moreover, we find that conflict reviews[12] would prefer showing conflicts in *Advancement*, *Novelty*, *Clarity* and *Overall*. This is also consistent with our typical understanding of peer reviews and occasional conflicts among them.

4.2 Measuring Sentiment Fusion for Individual Facets

Sentiments in the generated meta-reviews should be in line with the aggregate sentiment from the individual source documents including reviews and discussions. Seeing GPT-4 can predict the overall sentiment using judgements from source documents (Section 3.3), we introduce a reference-free evaluation metric, FusionEval, which assesses the consistency between the sentiments in the generated meta-review and that predicted by GPT-4 (with zero-shot prompting) from the source documents. Higher consistency implies the overall sentiment in generated meta-reviews are representative of the sentiments in the reviews and discussions (source documents).

Specifically, we first extract judgements from the generated meta-review following Section 3.2, and these judgements consist of *Content Expressions*, E and *Sentiment Levels*, L, and the corresponding *Review Facets*, F. Next, for each expression, $e \in E$, we predict the *Sentiment Level*, l', using GPT-4 (zero-shot) based on all judgements for the same *Review Facet* in the source documents following Section 3.3, and we get predicted *Sentiment Levels* for all judgements, L'. Lastly, FusionEval computes an accuracy score by evaluating L' against L. FusionEval only considers the precision instead of the recall for meta-review sentiments as it is reference-free and we have no information about the count of judgements that should be synthesized.

5 Enhancing LLMs with Explicit Information Consolidation

In this section, we propose two prompting methods to integrate the sentiment consolidation framework to generate meta-reviews. We compare the two methods with other prompting strategies including naive prompting and prompting with LLM-generated logic. We also run experiments on open-source models besides OpenAI ones to investigate the influence of different prompting methods on different models. The experiments are based on automatic and human annotation on 500 samples from PeerSum.[13]

5.1 Prompting LLMs with Sentiment Consolidation Logic

Following the process in Figure 1 we propose decomposing the meta-review generation process in the following steps: (1) Extracting content and sentiment expressions of judgements from source documents; (2) Predicting *Review Facets*, *Sentiment Levels*, and *Convincingness Levels*; (3) Clustering extracted judgements for different review facets; (4) generate a “mini summary” for judgements on the same review facet; and (5) Generating the final meta-review based on the mini summaries for all review facets.

We explore two methods to integrate this process for prompting an LLM. (1) Prompt-Ours: we describe the five steps in a single prompt and ask GPT-4 to generate the final meta-reviews (full prompt in Appendix G.1); (2) Pipeline-Ours: we create one prompt for each of the five steps, where the input for the intermediate step is the output from the previous step (full prompts in Appendix G.2).

We experiment with four open-source and close-source LLMs: GPT-4, GPT-3.5, LLaMA2-70B and LLaMA2-7B.[14]

[12] The same as in PeerSum (Li et al., 2023a), if any two reviews have ratings where the gap is larger than 4 they are conflict reviews.

[13] To avoid data contamination, we only use samples which were produced in and after 2022.

[14] Precise model names for them are: gpt-4-0613, gpt-3.5-turbo-1106, LLaMA2-70B-Chat, LLaMA2-7B-Chat. Note

LLM	Evaluation Metric	Prompt-Naive	Prompt-LLM	Prompt-Ours	Pipeline-Ours
GPT-4	FusionEval	50.14	48.90	53.62	**57.43**
	FacetEval	35.42	40.54	41.98	**42.36**
	ROUGE-1	27.16	27.49	**28.02**	24.91
	ROUGE-2	**6.63**	6.03	6.57	4.57
	ROUGE-L	24.78	24.75	**25.51**	22.70
GPT-3.5	FusionEval	48.35	49.66	51.40	**55.96**
	FacetEval	38.44	36.83	**39.88**	39.50
	ROUGE-1	28.22	25.04	**29.56**	28.92
	ROUGE-2	06.63	05.79	**6.95**	5.52
	ROUGE-L	25.36	22.77	**26.69**	16.13
LLaMA2-7B	FusionEval	46.85	46.83	50.18	**52.68**
	FacetEval	35.89	32.49	38.07	**38.35**
	ROUGE-1	25.94	23.88	**27.00**	19.39
	ROUGE-2	6.04	4.50	**6.86**	4.12
	ROUGE-L	23.57	21.59	**24.59**	17.37
LLaMA2-70B	FusionEval	47.35	48.53	50.24	**52.80**
	FacetEval	35.90	36.40	36.64	**36.82**
	ROUGE-1	26.61	16.60	**26.98**	26.41
	ROUGE-2	**6.56**	3.13	5.58	4.48
	ROUGE-L	**24.62**	14.63	24.20	23.71

Table 7: Performances of different LLMs with different prompting methods. For all metrics, a larger value denotes better performance. The bold and underlined values are the best and second in each row, respectively (×0.01)

Competition Groups	Preferred	IAA
Prompt-Naive LLaMA2-70B	46.67%	0.64
Prompt-Ours LLaMA2-70B	53.33%	
Prompt-Ours GPT-4	73.33%	0.74
Human-Written	26.67%	

Table 8: Two groups of human evaluation results based on human preferences: (1) comparing generated meta-reviews by Prompt-Naive and Prompt-Ours, and (2) comparing human-written meta-reviews and those generated by Prompt-Ours. IAA denotes inter-annotator agreement calculated with nominal Krippendorff's α.

5.2 Baselines

As baselines, we include two more methods: (1) Prompt-Naive: which prompts an LLM with a simple instruction to generate the meta-review (full prompt in Appendix G.3); and (2) Prompt-LLM: where we ask an LLM to self-generate the detailed steps for meta-review generation and we include these steps in the final prompt for meta-review generation (full prompt in Appendix G.4).

5.3 Reference-Based and Reference-Free Automatic Evaluation

For automatic evaluation, we adopt ROUGE (Lin and Hovy, 2003)[15], FacetEval (Section 4.1) and FusionEval (Section 4.2).[16]

We present the results in Table 7. Most LLMs perform better with either of our prompting methods (Prompt-Ours and Pipeline-Ours) than the baselines (Prompt-Naive and Prompt-LLM). Comparing between Prompt-Ours and Pipeline-Ours, the former seems to do best for ROGUE while the latter for FusionEval/FacetEval. As FusionEval/-FacetEval assesses the accuracy of the overall sentiment specifically, Pipeline-Ours is marginally better when it comes to getting the overall sentiment correct. Comparing different LLMs, GPT-4 seems to work best, although that is mostly true for FusionEval/FacetEval.

that for Pipeline-Ours, we always use GPT-4 for the first two steps, as we find that the other LLMs perform poorly for these tasks.

[15] We use the implementation of the algorithm in https://pypi.org/project/rouge-score/

[16] We do not consider other metrics such as BERTScore (Zhang et al., 2020), UniEval (Zhong et al., 2022), or G-Eval (Liu et al., 2023) as they have only been validated for summarization in news or a more general domain.

Human-Written Meta-Review
This paper suggests a method for defending against adversarial examples and out-of-distribution samples via projection onto the data manifold. The paper suggests a new method for detecting when hidden layers are off of the manifold, and uses auto encoders to map them back onto the manifold. The paper is well-written and the method is novel and interesting. However, most of the reviewers agree that the original robustness evaluations were not sufficient due to restricting the evaluation to using FGSM baseline and comparison with thermometer encoding (which both are known to not be fully effective baselines). After rebuttal, Reviewer 4 points out that the method offers very little robustness over adversarial training alone, even though it is combined with adversarial training, which suggests that the method itself provides very little robustness.
Meta-Review Generated by GPT-4 with Prompt-Naive
The reviewers have raised several concerns about the empirical evaluation and effectiveness of the proposed fortified network model. They note that the benefits of the fortified layers seem to be marginal and the experimental evaluation doesn't sufficiently demonstrate that the increased robustness is due to the fortified layers rather than the adversarial training. Furthermore, most of the reviewers would like to see benchmarks against state-of-the-art defenses such as PGD. Among other issues, reviewers have questioned the use of an autoencoder and the decision to use a denoising module in the hidden layers. Simply put, the reviewers would like a stronger reasoning for the methodological choices. There are also requests for clarification on the denoising terms and a comparison with a simple regularization loss. Moreover, a reviewer also raised a question on the effectiveness of the defense method against other types of attacks that might potentially compromise the proposed network. Most reviewers felt that the paper does not provide enough evidence that defends the proposed fortified network against adversarial attacks. Finally, concerns about the presentation of results and tables, and grammar errors have also been pointed out. In conclusion, while the reviewers acknowledge the relevance and importance of the research problem, the manuscript requires significant amendments to convince the experts about the effectiveness of the proposed method in tackling the problem. The authors must address these concerns and improve the manuscript to make a compelling case for their proposed approach.
Meta-Review Generated by GPT-4 with Prompt-Ours
The paper at hand proposes a fortified network model as a defense against adversarial attacks, based on the concept of denoising autoencoders. A key novelty in the presented approach is the placement of the denoising feature within the hidden layers, rather than just the input layer.The clarity and structure of the document have been appreciated by a number of reviewers, particularly the experimental data presentation and robust bibliography. There are, however, a few areas noted for improvement - predominantly around grammatical mistakes and confusing sentence structure. A significant source of debate among reviewers centers on the empirically evaluated efficacy of the fortified model. The authors propose that their model improves robustness by up to 5% over the baseline, but several reviewers argue that this improvement is marginal and does not sufficiently justify the implementation of an additional objective. This argument is most prominent when looking at the study's use of non-standard models and a PGD baseline that performs below the state-of-the-art, leading to questions about the accuracy of the baseline evaluation. Further, accuracy results against an Linf adversary of eps>0.1 on CIFAR-10 raises questions around the experiment's setup given known limitations relating to image perturbations. Suggestions were made to explore a nearest neighbor attack from the test set, multiple gradient queries per PGD step, and report standard non-robust accuracies to ensure robustness was not an artefact of model expressivity.Lastly, some apprehension was expressed about the motivation behind and efficacy of using a denoising autoencoder, with calls for comparison to a simpler baseline to justify the approach. In conclusion, while the fortified network model presents an intriguing approach to defending against adversarial attacks, the reviewers do not overwhelmingly agree that the empirical results sufficiently demonstrate advancement over existing methods. The majority find the defensive gain too marginal given the additional complexity and question the setup of the empirical evaluation. Further clarity in method and expanded empirical evidence would facilitating a stronger case for the proposed model.

Table 9: Human-written meta-review and the corresponding meta-reviews generated by GPT-4 with the naive prompt and the prompt based on the sentiment consolidation. (The green spans are positive sentiment values, red spans are negative sentiment values, while blue spans are the content expressions.)

5.4 Reference-Free Human Evaluation

To further validate the effectiveness of our prompting methods, we conduct human evaluations to assess the quality of meta-reviews generated by different prompting methods or written by human meta-reviewers. We recruited three volunteer annotators who are senior PhD students familiar with artificial intelligence research and the peer review process. They are asked to select their preferred meta-reviews based on their own understanding of high-quality meta-reviews without knowing the source.[17]

Prompt-Naive vs Prompt-Ours We randomly select 30 samples and the annotators are asked to compare the generated meta-reviews by Prompt-Naive and Prompt-Ours (using LLaMA2-70B) and select which one is better. Table 8 shows that the meta-reviews generated by Prompt-Ours are selected more by the annotators.

Prompt-Ours vs Human-Written We repeat the same experiments, but this time comparing meta-reviews generated by Prompt-Ours (GPT-4) vs. written by humans. Looking at Table 8, interestingly Prompt-Ours are much more preferred by the annotators. We suspect this may be because the generated meta-reviews tend to be more consistent in terms of the amount of detail it writes for each review facet, where else there is more variance for the human-written meta-reviews.

[17] We use majority voting to get the final human preference.

5.5 Case Study on Generated Meta-Reviews

To dive deeper into what difference the integration of sentiment consolidation framework makes, we also conduct a case study on generated meta-reviews with different prompting methods. We find that generated meta-reviews all seem plausible and machine-generated meta-reviews are much longer than human-written ones. In machine-generated meta-reviews, there are more details which are sometimes unnecessary or redundant. As shown in the example in Table 9, details such as "PGD" or "CIFA-10" are not essential to form the meta-review.

Our proposed Prompt-Ours tend to have a more balanced judgements. For example, in Table 9, Prompt-Naive does not talk about the positive aspects for *Clarity* and only highlights some issues, but Prompt-Ours comments on both the strengths and weaknesses for *Clarity*. This is consistent with the finding in Table 7 that Prompt-Naive gets worse sentiments than Prompt-Ours.

6 Conclusions and Future Work

In this paper, we explore sentiment-focused multi-document information consolidation within the task of scientific sentiment summarization. We introduce a three-layer framework of sentiment consolidation to focus on generating meta-reviews and it considers the sentiments for each review facet in the reviews and discussions. We also propose automatic evaluation metrics that assess the overall sentiments in the generated meta-reviews. Experiments on meta-review generation show that prompting LLMs by following the processes in the three-layer framework results in better meta-reviews, providing an empirical validation of our framework for describing the meta-review writing process. As the sentiment consolidation also exist in other domains where human reviews or comments exist such as politics and advertisement, we will explore adapting our proposed sentiment consolidation framework into other domains in the future.

Limitations

Although integration of the sentiment consolidation framework could improve the generation results, there are still some limitations of this work.

- As in other areas peer review data is not publicly available, we use the data only from some artificial intelligence conferences, and this may make the models biased. We hope that more data from diverse areas could be included.
- Experiments are only in English texts rather than other languages.
- We only inject the information consolidation logic into prompting based models instead of fine-tuning based models. We will investigate leveraging the information consolidation framework to improve fine-tuned models in the future.
- Although GPT-4 can predict meta-review sentiments based on source judgements to some extent, we have to understand more about how these models achieve this and what makes them fail in error cases.
- Meta-review generation is not only about sentiment prediction, future work has to consider more information such as argumentation in source reviews and justification in meta-reviews.

Ethics Statement

While our experiments demonstrate that the models exhibit potential in generating satisfactory meta-reviews to a certain degree, we strongly advise against solely relying on the generated results without manual verification and review, as instances of hallucinations exist in the generations. It is important to emphasize that we do not advocate for replacing human meta-reviewers with LLMs. However, it is noteworthy that these models have the capacity to enhance the meta-reviewing process, rendering it more efficient and effective.

Acknowledgements

We would thank all the anonymous reviewers for their great comments to improve the work. It is impossible to have this work without the volunteer contributions of the annotators on human evaluations.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *AAAI*, pages 12489–12497.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *SIGIR*, pages 1653–1656.

Hou Pong Chan, Wang Chen, and Irwin King. 2020. A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In *SIGIR 2020*, pages 1191–1200.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *JAIR*, 77:103–166.

Md. Murad Hossain, Luca Anselma, and Alessandro Mazzei. 2023. Exploring sentiments in summarization: Sentitextrank, an emotional variant of textrank. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, volume 3596.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023a. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of EMNLP*.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023b. Compressed heterogeneous graph for abstractive multi-document summarization. In *AAAI*.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*, pages 71–78.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *CoRR*, abs/2208.04347.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

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 *Findings of ACL*, pages 2521–2535.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *CIKM*, pages 2189–2198.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Meta-review generation with checklist-guided iterative introspection. *CoRR*, abs/2305.14647.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *EMNLP*, pages 2023–2038.

A Review Criteria in Different Reviewer Guidelines

Academic Press	Review guidelines
ACM	https://dl.acm.org/journal/dgov/reviewer-guidelines
ACL Rolling Review	https://aclrollingreview.org/reviewertutorial
IEEE	https://conferences.ieeeauthorcenter.ieee.org/understand-peer-review/
Springer	https://www.springer.com/gp/authors-editors/authorandreviewertutorials/howtopeerreview/evaluating-manuscripts/10286398
NeurIPS	https://neurips.cc/Conferences/2021/Reviewer-Guidelines
ICLR	https://iclr.cc/Conferences/2023/ReviewerGuide#Reviewinginstructions
ACL	https://2023.aclweb.org/blog/review-acl23/
Cambridge University Press	https://www.cambridge.org/core/services/aop-file-manager/file/5a1eb62e67f405260662a0df/Refreshed-Guide-Peer-Review-Journal.pdf

Table 10: Review guidelines from different academic presses.

B Annotation Instructions for Human Annotation

The screenshots of the two-page annotation instruction for human annotation are shown in Figure 4 and Figure 5 in the last two pages of the Appendix.

C Inter-Annotator Agreement Among Human Annotators and GPT-4

We describe how we calculate inter-annotator agreement among human annotators and GPT-4 here. For Content Expression and Sentiment Expression, as they are highlighted text spans we calculate the character-level agreement with Krippendorf's α and Cohen's κ. Specifically, for each document, two annotators may highlight different text spans for Content Expression and Sentiment Expression. We construct two vectors of the same length as the characters to represent the highlighting behaviours of any two annotators. This agreement shows whether annotators identify sentiments from similar text spans.

For *Review Facet*, *Sentiment Level*, and *Convincingness Level*, we calculate Krippendorf's α and Cohen's κ in a common way. We first identify whether two annotators recognize sentiment from the same text span with a ROUGE threshold (the summation of ROUGE-1, ROUGE-2 and ROUGE-L between highlighted text spans for sentiment is larger than 2.0), and calculate agreement on the predicted values.

Inter-annotator agreement between two human annotators for human annotation in Section 3.2 are present in Table 11, Table 12, and Table 13. Averaged agreement of GPT-4 with the two human annotators are present in Table 14, Table 15, and Table 16.

D Prompt to Get Content and Sentiment Expressions with GPT-4

Annotation	Cohen's κ	Krippendorf's α
Content Expression	0.623	0.623
Sentiment Expression	0.666	0.665
Review Facet	0.769	0.769
Sentiment Level	0.770	0.770
Convincingness Level	0.534	0.533

Table 11: Human annotator agreement on annotating meta-reviews.

Annotation	**Cohen's κ**	**Krippendorff's α**
Content Expression	0.631	0.631
Sentiment Expression	0.654	0.654
Review Facet	0.783	0.783
Sentiment Level	0.844	0.844
Convincingness Level	0.405	0.398

Table 12: Human annotator agreement on annotating official reviews.

Annotation	**Cohen's κ**	**Krippendorff's α**
Content Expression	0.572	0.572
Sentiment Expression	0.609	0.609
Review Facets	0.857	0.857
Sentiment Levels	0.764	0.763
Convincingness Levels	0.455	0.437

Table 13: Human annotator agreement on annotating discussions.

Annotation	*A*	*B*	**Avg**
Content Expression	0.558	0.542	0.550
Sentiment Expression	0.565	0.594	0.580
Review Facets	0.588	0.610	0.599
Sentiment Levels	0.552	0.541	0.547
Convincingness Levels	0.213	0.192	0.203

Table 14: GPT-4 agreement in terms of Cohen's κ with human annotators *A* and *B* on annotating meta-reviews.

Annotation	*A*	*B*	**Avg**
Content Expression	0.522	0.534	0.528
Sentiment Expression	0.544	0.569	0.557
Review Facets	0.579	0.637	0.608
Sentiment Levels	0.594	0.589	0.592
Convincingness Levels	0.008	0.013	0.011

Table 15: GPT-4 agreement in terms of Cohen's κ with human annotators *A* and *B* on annotating official reviews.

Annotation	*A*	*B*	**Avg**
Content Expression	0.176	0.187	0.182
Sentiment Expression	0.182	0.188	0.185
Review Facets	0.480	0.381	0.431
Sentiment Levels	0.123	0.046	0.082
Convincingness Levels	0.0	0.0	0.0

Table 16: GPT-4 agreement in terms of Cohen's κ with human annotators *A* and *B* on annotating discussions.

```
Please read the document:

{{source_document}}

This task requires you to analyze the above document which is used to express opinions on the quality of a scientific manuscript. You are good at understanding the sentiment information with judgements in the document.
Please first identify the sentence with judgements only on the quality of scientific manuscripts based on the review facets for scientific peer-review: novelty, soundness, clarity, advancement, compliance and overall quality within the given document.
Once you have found a sentence that provides judgement in one or more of these areas, you then need to extract the specific expression of sentiment and the content it refers to.

The process can be broken into two steps:
1) Identify a judgement sentence that focuses on the quality of the manuscript based on the given criteria.

2) From the identified judgement sentence, extract two pieces of information: the sentiment expression and the content expression. The sentiment expression is the specific term or phrase that conveys the sentiment or opinion. The content expression pertains to the content that this sentiment is referring to.

Please provide the data in the following format:
{"judgement_sentence": "sentence", "content_expression": "content", "sentiment_expression": "sentiment"}

Here are a few examples for your reference:
{"judgement_sentence": "The writing of the paper is not well-written.", "content_expression": "The writing of the paper", "sentiment_expression": "not well-written"}
{"judgement_sentence": "Experimental results are not sufficiently substantiated.", "content_expression": "Experimental results", "sentiment_expression": "not sufficiently substantiated"}
{"judgement_sentence": "This paper presents two novel approaches to provide explanations for the similarity between two samples based on 1) the importance measure of individual features and 2) some of the other pairs of examples used as analogies.", "content_expression": "approaches", "sentiment_expression": "novel"}

The predicted judgments (following the same jsonline format of the above example):
```

E Prompt to Get Judgement Component Predictions with GPT-4

```
Please first read the document below:
```

```

{{source_document}}


Please predict the facet that the given judgements are talking about. You can refer to the context in the above source document.

Possible facets:

Novelty: How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).

Soundness: (1) Empirical: how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted. (2) Theoretical: whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness and the methodology (e.g., mathematical approach) and the analysis is correct.

Clarity: The readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.

Advancement: Importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.

Compliance: Whether the manuscript fits the venue, and all ethical and publication requirements are met.

Overall: Overall quality of the manuscript, not for specific facets.


You are also good at understanding sentiment information in the judgements.

Please predict the original expresser of the sentiment in the judgement sentence. You can refer to the context in the source document.

Possible sentiment expressers:

- Self: the sentiment is from the speaker
- Others: the sentiment is quoted from others


Please predict how well the sentiment in the judgement sentence is justified in the document in your understanding. You can refer to
```

```
the context in the source document.

Possible sentiment convincingness:

- Not applicable: the sentiment is explicitly excerpted from others.
- Not at all: not convincing at all or when there is no justification. How well the sentiment is justified in the document in your understanding
- Slightly Convincing: there is some evidence or logical reasoning, but it might not be comprehensive.
- Highly Convincing: leaving little room for doubt.


Please predict the polarity and strength of the sentiment in the judgement sentence. You can refer to the context in the source document.

Possible sentiments polarities:

- Strong negative: very negative
- Negative: minor negative
- Positive: minor positive
- Strong positive: very positive


Judgements:
{{judgement_expressions}}

Your predictions for the above judgements (following the same jsonlines format, return the same number of lines, and keep the same content and sentiment expressions):
```

F Prompts to Predict Meta-Review Sentiment Levels

F.1 Prediction with Judgements of Source documents

The judgements are extracted from source documents, and they are in the same review facet to the target meta-review judgement.

```
You will be given source judgements from reviewers for a scientific manuscript. Your task is to implicitly write a meta-review for these judgements and predict the sentiment level based on these judgements.

Source Judgements:

{{source_judgements}}

Candidate Sentiment Levels:

- Strong negative
- Negative
- Positive
- Strong positive

Content Expression:

{{content_expression}}

Predict the sentiment level of the given content expression based on the above judgements. You must follow the following format.
{"Content Expression": the above content expression, "Sentiment Level": your predicted sentiment level}
```

F.2 Prediction with Full Texts of Source documents

The source texts are the concatenation of the source documents.

```
You will be given multiple review documents for a scientific manuscript. Your task is to implicitly write a meta-review and predict the sentiment level based on these documents.

Source Documents:

{{source_texts}}

Candidate Sentiment Levels:

- Strong negative
- Negative
- Positive
- Strong positive

Content Expression:

{{content_expression}}

Predict the sentiment level of the given content expression based on related information in the above documents. You must follow the following format.
{"Content Expression": the above content expression, "Sentiment Level": your predicted sentiment level}
```

G Prompts for Meta-Review Generation with Integration of Information Consolidation Logic

G.1 Prompt with Descriptive Consolidation Logic

```
    Your task is to write a meta-review based on the following reviews and discussions for a scientific manuscript.

{{input_documents}}

Following the underlying steps below will get you better generated meta-reviews.
```

```

1. Extracting content and sentiment expressions of judgements in all above review and discussion documents;

2. Predicting Review Facets, Sentiment Levels, and Convincingness Levels;
Candidate review facets: Novelty, Soundness, Clarity, Advancement, Compliance, and Overall quality
Candidate sentiment levels: Strong negative, Negative, Positive and Strong positive
Candidate convincingness levels:  Not at all, Slightly Convincing, Highly Convincing

3. Reorganize extracted judgements in different clusters for different review facets;

4. Generate a small summary for judgements on the same review facet with comparison and aggregation;

5. Aggregate judgements in different review facets and write a meta-review based on the aggregation.


You may follow these steps implicitly and only need to output the final meta-review. The final meta-review:
```

G.2 Prompts Used in the Pipeline Generation

Prompts for the first two steps, getting content and sentiment expressions and predicting other judgement components, are the same as prompts in Appendix D and Appendix E, respectively.

For the step of generating sub-summaries for individual facets, the prompt is as follows.

```
{{input_judgements}}

Write a summary of the above judgements on {{criteria_facet}} of a manuscript.
```

For the step of generating final meta-reviews based on sub-summaries of individual facets, the prompt is as follows.

```
{{input_sub_summaries}}

Write a meta-review to summarize the above sub-summaries of reviews and discussions in different review facets for a manuscript.
```

G.3 Prompts from Prompt-Naive

For Prompt-Naive in our experiments, the prompt we use is as follows.

```
{{input_documents}}

Write a meta-review based on the above reviews and discussions for a manuscript.
```

G.4 Prompts from Prompt-LLM

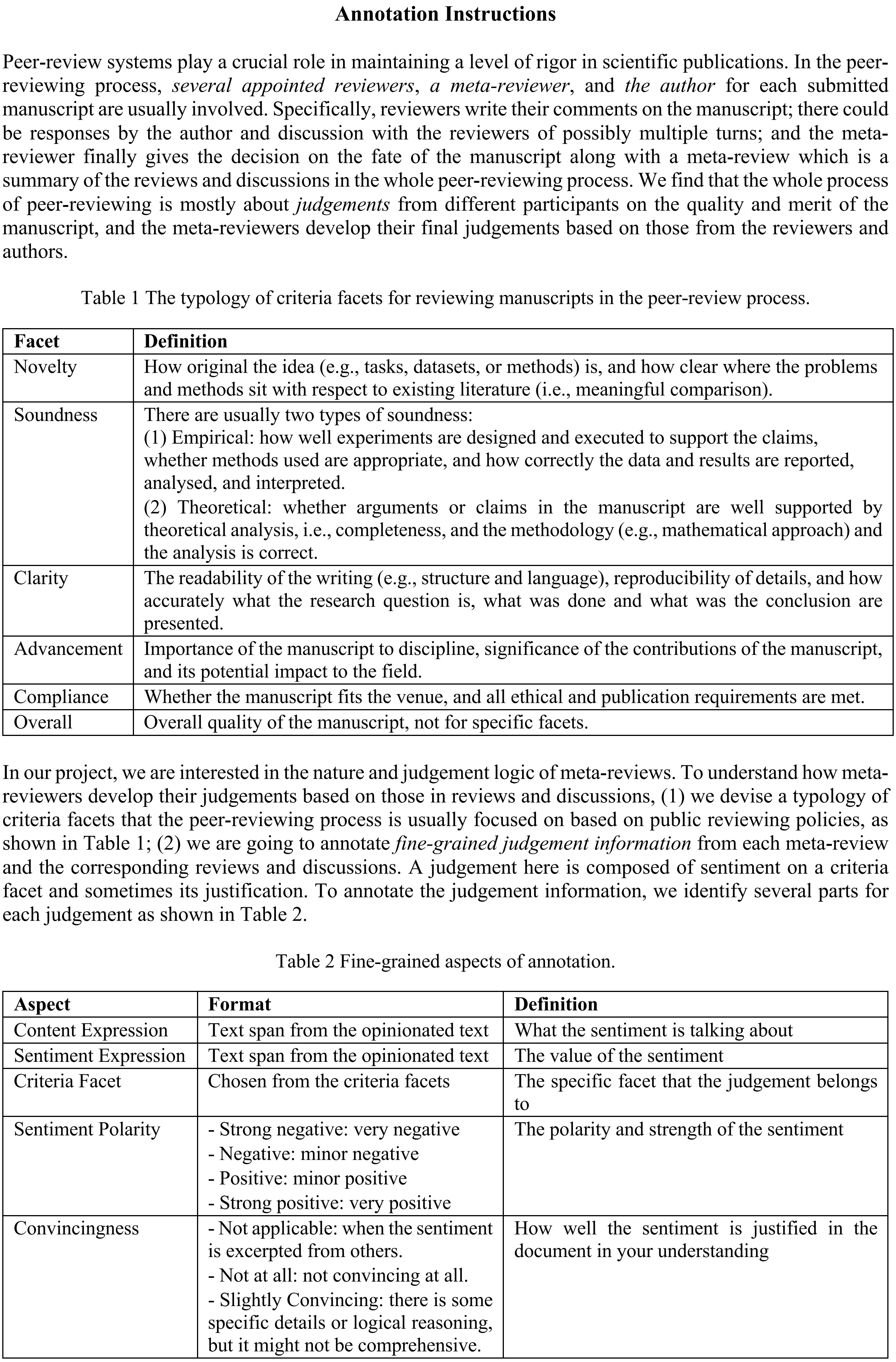
For Prompt-LLM, we have to generate first the steps with GPT-4 and then the meta-review based on the generated steps.

The prompt to generate the steps:

```
{{input_documents}}

What are the steps to write a meta-review specifically for the above reviews and discussions of a manuscript.
```

The prompt to generate the meta-review:

```
{{input_documents}}

Follow the following steps and write a meta-review based on the above reviews and discussions for a manuscript.

{{generated_steps}}
```



Annotation Instructions

Peer-review systems play a crucial role in maintaining a level of rigor in scientific publications. In the peer-reviewing process, *several appointed reviewers*, *a meta-reviewer*, and *the author* for each submitted manuscript are usually involved. Specifically, reviewers write their comments on the manuscript; there could be responses by the author and discussion with the reviewers of possibly multiple turns; and the meta-reviewer finally gives the decision on the fate of the manuscript along with a meta-review which is a summary of the reviews and discussions in the whole peer-reviewing process. We find that the whole process of peer-reviewing is mostly about *judgements* from different participants on the quality and merit of the manuscript, and the meta-reviewers develop their final judgements based on those from the reviewers and authors.

Table 1 The typology of criteria facets for reviewing manuscripts in the peer-review process.

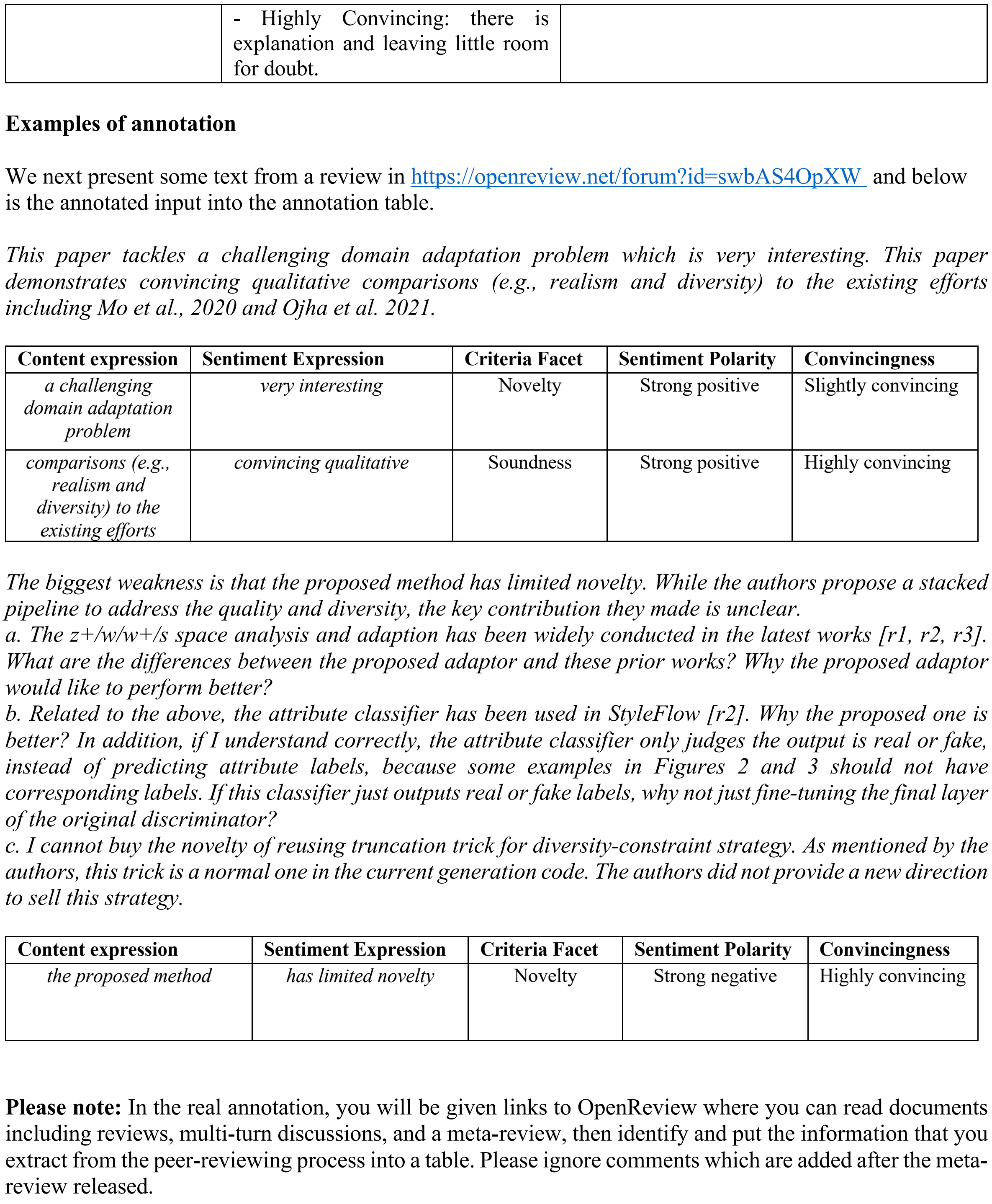
Facet	**Definition**
Novelty	How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).
Soundness	There are usually two types of soundness: (1) Empirical: how well experiments are designed and executed to support the claims, whether methods used are appropriate, and how correctly the data and results are reported, analysed, and interpreted. (2) Theoretical: whether arguments or claims in the manuscript are well supported by theoretical analysis, i.e., completeness, and the methodology (e.g., mathematical approach) and the analysis is correct.
Clarity	The readability of the writing (e.g., structure and language), reproducibility of details, and how accurately what the research question is, what was done and what was the conclusion are presented.
Advancement	Importance of the manuscript to discipline, significance of the contributions of the manuscript, and its potential impact to the field.
Compliance	Whether the manuscript fits the venue, and all ethical and publication requirements are met.
Overall	Overall quality of the manuscript, not for specific facets.

In our project, we are interested in the nature and judgement logic of meta-reviews. To understand how meta-reviewers develop their judgements based on those in reviews and discussions, (1) we devise a typology of criteria facets that the peer-reviewing process is usually focused on based on public reviewing policies, as shown in Table 1; (2) we are going to annotate *fine-grained judgement information* from each meta-review and the corresponding reviews and discussions. A judgement here is composed of sentiment on a criteria facet and sometimes its justification. To annotate the judgement information, we identify several parts for each judgement as shown in Table 2.

Table 2 Fine-grained aspects of annotation.

Aspect	**Format**	**Definition**
Content Expression	Text span from the opinionated text	What the sentiment is talking about
Sentiment Expression	Text span from the opinionated text	The value of the sentiment
Criteria Facet	Chosen from the criteria facets	The specific facet that the judgement belongs to
Sentiment Polarity	- Strong negative: very negative - Negative: minor negative - Positive: minor positive - Strong positive: very positive	The polarity and strength of the sentiment
Convincingness	- Not applicable: when the sentiment is excerpted from others. - Not at all: not convincing at all. - Slightly Convincing: there is some specific details or logical reasoning, but it might not be comprehensive.	How well the sentiment is justified in the document in your understanding

Figure 4: The first page of the annotation instruction for human judgement annotation.



	- Highly Convincing: there is explanation and leaving little room for doubt.	

Examples of annotation

We next present some text from a review in https://openreview.net/forum?id=swbAS4OpXW and below is the annotated input into the annotation table.

This paper tackles a challenging domain adaptation problem which is very interesting. This paper demonstrates convincing qualitative comparisons (e.g., realism and diversity) to the existing efforts including Mo et al., 2020 and Ojha et al. 2021.

Content expression	Sentiment Expression	Criteria Facet	Sentiment Polarity	Convincingness
a challenging domain adaptation problem	*very interesting*	Novelty	Strong positive	Slightly convincing
comparisons (e.g., realism and diversity) to the existing efforts	*convincing qualitative*	Soundness	Strong positive	Highly convincing

The biggest weakness is that the proposed method has limited novelty. While the authors propose a stacked pipeline to address the quality and diversity, the key contribution they made is unclear.
a. The z+/w/w+/s space analysis and adaption has been widely conducted in the latest works [r1, r2, r3]. What are the differences between the proposed adaptor and these prior works? Why the proposed adaptor would like to perform better?
b. Related to the above, the attribute classifier has been used in StyleFlow [r2]. Why the proposed one is better? In addition, if I understand correctly, the attribute classifier only judges the output is real or fake, instead of predicting attribute labels, because some examples in Figures 2 and 3 should not have corresponding labels. If this classifier just outputs real or fake labels, why not just fine-tuning the final layer of the original discriminator?
c. I cannot buy the novelty of reusing truncation trick for diversity-constraint strategy. As mentioned by the authors, this trick is a normal one in the current generation code. The authors did not provide a new direction to sell this strategy.

Content expression	Sentiment Expression	Criteria Facet	Sentiment Polarity	Convincingness
the proposed method	*has limited novelty*	Novelty	Strong negative	Highly convincing

Please note: In the real annotation, you will be given links to OpenReview where you can read documents including reviews, multi-turn discussions, and a meta-review, then identify and put the information that you extract from the peer-reviewing process into a table. Please ignore comments which are added after the meta-review released.

Figure 5: The first page of the annotation instruction for human judgement annotation.