Exploring Multi-Document Information Consolidation for Scientific Sentiment Summarization

Anonymous ACL submission

Abstract

Modern natural language generation systems with LLMs exhibit the capability to generate a plausible summary of multiple documents; however, it is uncertain if models truly possess the ability of information consolidation to generate summaries, especially on those source documents with opinionated information. To make scientific sentiment summarization more grounded, we hypothesize that in peer review human meta-reviewers follow a three-layer framework of sentiment consolidation to write meta-reviews and it represents the logic of summarizing scientific sentiments in meta-review generation. The framework is validated via human annotation. Based on the framework, we propose evaluation metrics to assess the quality of generated meta-reviews, and we find that the hypothesis of the sentiment consolidation framework works out empirically when we incorporate it as prompts for LLMs to generate meta-reviews in extensive experiments.[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 piece of simple instruction as a prompt 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 which 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 sentiment consolidation.

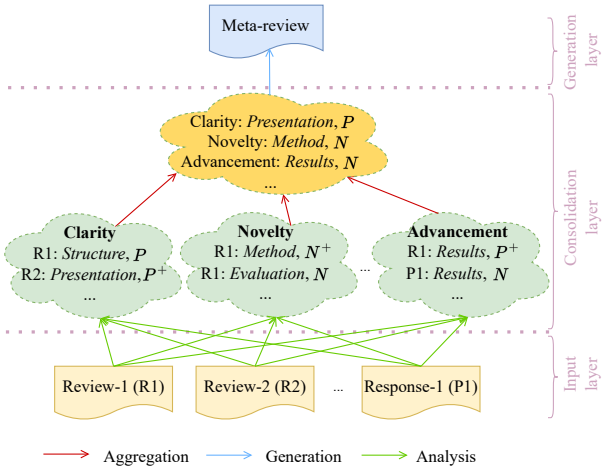

Figure 1: The three-layer framework of the underlying information consolidation logic in meta-reviewing (P: Positive, P^+: Strongly positive, N: Negative, N^+: Strongly negative).

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 not suitable for investigating information consolidation as (1) the summaries are synthetic, composed of a simple combination of extracted snippets (Amplayo et al., 2021), (2) the summary of product reviews is about extracting the majority sentiment. 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 exceptional opportunity for exploring the intricate process of multi-document information consolidation. This is because (1) meta-reviewers are supposed to understand not only all the reviews

[1] We will release the code and annotated data at ANONYMIZED upon acceptance.

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

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 argument logic has to be taken into account to arrive at sentiments in the meta-reviews where majority voting is not always working, (3) meta-reviews have to recognize and resolve conflicts and consensus among source documents (i.e., corresponding reviews and discussions).

In this paper, we hypothesize that human meta-reviewers are following a three-layer sentiment consolidation framework as shown in Figure 1 to write meta-reviews based on individual reviews and discussions in the peer review process. Human and automatic annotation is then conducted to get judgements from meta-reviews and corresponding source documents and these judgements are the foundation of our framework. With the proposed sentiment consolidation framework and annotated data, we propose two evaluation metrics which are focused more on sentiments in generated meta-reviews, and in our extensive experiments, we observe empirical validation of the framework hypothesis when integrated as prompts for LLMs to generate meta-reviews.

Contributions of the paper are summarized as follows:

- With the initialization of scientific sentiment summarization, we hypothesize that human meta-reviewers follow a three-layer framework that describes the sentiment consolidation process in the meta-reviewing process;
- We collect human annotations on meta-reviews and corresponding source documents based on the sentiment consolidation framework;
- We propose two automatic metrics (reference-free and reference-based) to evaluate the sentiment in the generated meta-reviews.
- Experiments show that the framework works out empirically 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.

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

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	360.53
#Words/Meta-review	16	648	150.87

Table 2: Statistics of the human annotated data.

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. Most work in the text summarization community is only trying to improve the generation quality of text summarization, such as relevance and faithfulness without considering the generation process. 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, and this is still under exploration. Moreover, it is also of limited exploration that the human information consolidation logic in summarization could be used to enhance the generation ability of LLMs and make generations more grounded.

2.2 Automated Sentiment Summarization

Sentiment summarization is to get a summary of sentiments in the input texts (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 exploration. 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 dis-

cussions 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). Although Li et al. (2023a) take consideration of the in-nature conversational structure of source documents, their models have no knowledge of how human meta-reviewers write the meta-reviews and they do not consider the judgements which would be the most important content in the peer-reviewing process in their modelling or evaluation. Because this task features multiple source documents with complex relationships which should be considered in the corresponding meta-review, it is a suitable scenario to investigate the ability of models to consolidate information in the summarization process.

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 judgement identification and extraction, analyze the fusion process of scientific sentiments.

3.1 Hierarchical Sentiment Consolidation

We propose the task of scientific sentiment summarization, synthesizing reviews and discussions in the peer-reviewing process into the corresponding meta-review. While we use the PeerSum[3] dataset where the input is reviews and discussions and the target output is the corresponding human-written meta-review, scientific sentiment summarization would focus more on sentiments in the summary or meta-review generation process.

In the domain of peer review, based on reviewer guidelines from the most popular academic presses such as ACM and IEEE[4], we find that all comments in the whole process of peer-reviewing are mostly about judgements from different participants on the quality and merit of the manuscript. These opinions or judgements are based on six facets of review criteria (*Novelty*, *Soundness*, *Clarity*, *Advancement*, *Compliance* and *Overall quality*) which are the foundations of the peer-reviewing process. The meta-reviewers must form their final judgements based on those from the reviewers and authors. Similarly, based on meta-reviewer guidelines for ICLR[5] and NeurIPS[6], we find that the meta-reviewer needs to understand and aggregate information from comments in the whole peer-reviewing process. They may first identify judgements from different review and discussion documents, then aggregate opinions in different criteria facets, and lastly based on this aggregation they write the meta-review.

To conceptualize the logic of scientific sentiment summarization from multiple documents, we propose a three-layer framework, taking meta-review generation as an example shown in Figure 1. The three layers include the input layer, the consolidation layer, and the generation layer. The input layer is to understand individual judgements in input documents of different types: official reviews and multi-turn discussions. The consolidation layer is to represent what the meta-reviewers are supposed to do to aggregate sentiments from the input documents. Specifically, they first identify and extract judgements from different documents, then reorganize the judgements, and lastly do judgement fusion to form the final opinions of each criteria facet. In the generation layer, the meta-reviewer writes the meta-review to express the judgements that they develop based on consolidation of the whole peer-reviewing process.

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 criteria facet with sometimes its justification, and it contains several components: Content Expression, Sentiment Expression, Criteria Facet, Sentiment Level, and Convincingness Level (definitions are shown in Table 1, and examples are in Figure 5 in human annotation). 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 to Li et al. (2023a)) to be annotated. We recruit two human annotators[7] to do this annotation. The whole human

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

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

[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.

annotation instruction is in Appendix B. As the annotation is time-consuming, we finish 30 samples and it costs about one hour to annotate per sample and it costs about 60 hours and 2,100 US dollars in total. Two annotators get 1,812 and 1,744 judgements, respectively. The statistics of these 30 samples are present in Table 2. Agreement of the two annotators on these three types of documents is shown in Figure 2 in terms of Krippendorff's α.[8] It is clear that there is high agreement between the two human annotators. This shows that judgements on these criteria facets in these documents really exist and this makes it possible to analyze the sentiment consolidation among source documents.

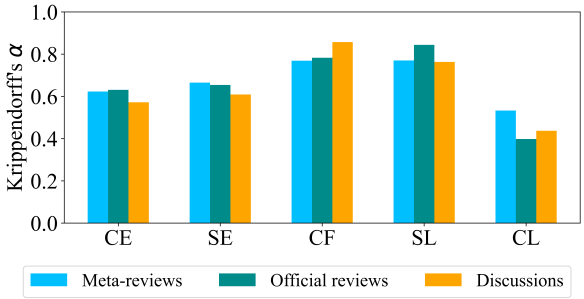

Figure 2: Inter-annotator agreement on meta-reviews, official reviews and discussions in terms of Krippendorff's α for different judgement components including Content Expression (CE), Sentiment Expression (SE), Criteria Facet (CF), Sentiment Level (SL), and Convincingness Level (CL).

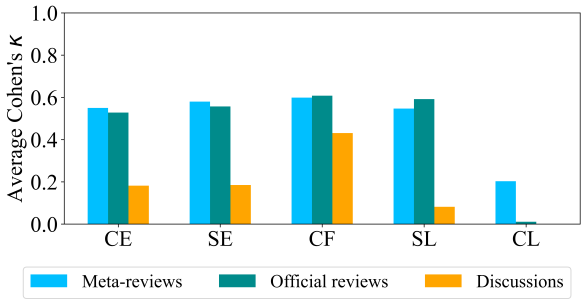

Figure 3: The averaged GPT-4's agreement with two human annotators on meta-reviews, official reviews and discussions in terms of Krippendorff's α for different judgement components including Content Expression (CE), Sentiment Expression (SE), Criteria Facet (CF), Sentiment Level (SL), and Convincingness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

To get more annotated judgements, we split the annotation task into two phases: extracting content and sentiment expressions, and predicting other components of judgements, and leverage GPT-4 to

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 criteria facets in meta-review judgements and meta-review documents.

accomplish this with in-context learning as there is not enough data to train specific models. Specifically, we design prompts for GPT-4[9], which is currently one of the most powerful LLMs, with few-shot prompting to get the content and sentiment expressions which express the judgements in annotation documents (the prompt template is in Appendix D), and get the predictions of other judgement components (the prompt template is in Appendix E). We present average agreement of GPT-4 with the two human annotators in Figure 3[10]. We can see that annotation results by GPT-4 have a high agreement with human annotation results, which means that GPT-4 can be qualified to do the annotation on meta-reviews and official reviews. It is also clear that GPT-4 cannot have high agreement with human annotators on annotating discussions. This may be because that judgements in discussion documents are usually to rebuttal and not easy to detect, which is consistent in human annotation. Moreover, it is clear that GPT-4 has lowest agreement with human annotators on convincingness levels in Figure 3 while human annotators have the lowest agreement on convincingness levels in Figure 2. This may be because that convincingness is subjective.

3.3 Sentiment Fusion for Consolidation

With all the annotated judgements, we next dive more into the process of sentiment aggregation. Among all the criteria facets, we find that *Soundness* and *Advancement* are the two most important criteria 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-written meta-reviews are not always following simple majority voting to develop sentiments. Specifically, in

[8]Calculation details and more results in terms of both Cohen's κ and Krippendorff's α are in Table 10, Table 11 and Table 12 in Appendix C.

[9]The version of GPT-4 in the paper is gpt-4-0613 in default.

[10]More results are in Table 13, Table 14 and Table 15 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.

the document level, we find that in PeerSum there are 23.7% samples where human meta-reviewer's acceptance decision is not consistent with the prediction based on majority voting by review ratings (if the number of ratings which are higher than 5 for a paper is larger than the number of ratings which are not higher than 5, the majority voting will get *Accept* while the human meta-reviewer's decision may be *Reject* for the paper.); in the sentiment level, there is an example in Table 4 where the meta-reviewer is not following majority voting to form the sentiment on *Novelty*.

To learn the aggregation function of sentiments from source documents to the meta-review, we try in-context learning of LLMs as there are not enough training data to train other models and LLMs like GPT-4 are powerful in zero-shot learning. We formulate this as a text classification task where the output is the sentiment level of the piece of *Content Expression* in a meta-review judgement and the input is judgements in all source documents which are belonging to the same criteria facet as the meta-review judgement or the full text of all source documents. Specifically, we randomly sample 100 human-annotated meta-review judgements for each criteria facet in each annotator's annotation results, and the prompt for GPT-4 is in Appendix F. Average results between two human annotators in Table 5 suggest that using judgements is better than full texts to predict the meta-review sentiments and this zero-shot classification function has the potential to predict the sentiments in the meta-review. This suggests that it may be useful to incorporate some intermediate steps to extract text segments related to a facet before generating the final sentiment (i.e. these results argue that we should not treat it as a simple text-to-text problem where the input is reviews and discussions and the output is the meta-review).

Criteria 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 to predict sentiment levels of human-annotated meta-review judgements based on annotated source judgements or original full texts.

4 Sentiment-Aware Evaluation on Information Consolidation

In this section, we focus more on how to evaluate 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 criteria facets. Different from the generic evaluation metrics for abstractive summarization or text generation which mostly adopt surface-form matching, we focus more on criteria facets and their corresponding sentiment levels in human-written and model-generated meta-reviews.

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

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

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

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 criteria 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

Criteria 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.)

$[P_f^+, P_f, N_f^+, N_f, O_f]$ where P_f^+ denotes the frequency of *Strongly positive* for the facet f, P_f denotes the frequency of *Positive*, N_f^+ denotes the frequency of *Strongly negative*, N_f denotes the frequency of *Negative*, and O_f denotes whether this facet is involved in the document.

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 representations of documents.

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

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[11] would prefer showing conflicts in *Advancement*, *Novelty*, *Clarity* and *Overall*. This is also consistent with our typical understanding in 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. As GPT-4 has the potential to predict sentiments in the human-written meta-reviews from judgements in source reviews and discussions in Section 3.3, we present a reference-free evaluation metric, FusionEval, which assesses the consistency between the sentiments in the generated meta-review and that predicted by the zero-shot sentiment level prediction with GPT-4 from the source judgements. The higher consistency indicates that the generated meta-reviews follow a more similar sentiment aggregation scheme to get the meta-review sentiments.

[11]The same as in PeerSum, if any two reviews have ratings where the gap is larger than 4 they are conflict reviews.

Specifically, we first extract judgements from the generated meta-review with a set of *Content Expression*, E and *Sentiment Level*, L; next, for all expressions in E, we get the predicted set of *Sentiment Level*, L' with the zero-shot sentiment prediction function based on judgements in the source documents; last, the accuracy between L and L' is the consistency and the final score of FusionEval. 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, to enhance LLMs in meta-review generation we propose two prompting methods to integrate the information consolidation logic. 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.[12]

5.1 Prompting LLMs with Sentiment Consolidation Logic

As human meta-reviewers follow the sentiment consolidation logic introduced in Figure 1, LLMs may get improved if we integrate the logic into the generation process. As we know, there are mainly two paradigms in text generation: prompting and fine-tuning. Because fine-tuned models are black-box and it is not easy to integrate the information consolidation logic. Prompting with detailed logic would make it possible to investigate intermediate results in the generation process and this could promote trust and transparency, therefore we focus on enhancing LLMs in prompting with the integration of information logic.

Based on the framework in Figure 1 and comprehensive analysis in Section 3.3, there should be several steps in the meta-review generation process. (1) Step 1: Extracting content and sentiment expressions of judgements in all source documents;(2) Step 2: Predicting *Criteria Facets*, *Sentiment Levels*, and *Convincingness Levels*; (3) Step 3: Reor-

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

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. ($\times 0.01$)

ganize extracted judgements in different clusters for different criteria facets; (4) Step 4: Generate a small summary for judgements on the same criteria facet with sentiment comparison and aggregation; (5) Step 5: Generate the final meta-review based on summaries for different criteria facets.

To integrate the information consolidation logic into prompting of LLMs, we present two methods. (1) Prompt-Ours: we inject descriptive sentiment consolidation logic into the prompt and make the model follow the steps. The complete prompt is in Appendix G.1. (2) Pipeline-Ours: we formulate the task of generating meta-reviews into a pipeline and conduct each part separately instead of a single prompt. The first two steps are finished by GPT-4[13] and all used prompts are in Appendix G.2.

5.2 Reference-Based and Reference-Free Automatic Evaluation

In automatic evaluation, we adopt ROUGE (Lin and Hovy, 2003)[14] and our proposed FacetEval and FusionEval. Besides ROUGE which is the most popular metric based on surface matching, there are some other semantic representation based metrics like BERTScore (Zhang et al., 2020), UniEval (Zhong et al., 2022), and G-Eval (Liu et al., 2023), but they are only optimized for news summarization or dialogue generation tasks instead of sentiment summarization.

To evaluate the benefit from the integration of the sentiment consolidation logic, we compare the two proposed prompting methods with the naive prompting (Prompt-Naive) which only simply make the model generate meta-reviews without detailed instructions and LLM-generated instructions (Prompt-LLM) which are based on the understanding of the model on how to generate meta-reviews. Moreover, we run experiments on closed-source (GPT-4 and GPT-3.5) and open-source models (LLaMA2-70B and LLaMA2-7B.)[15]

Results are present in Table 7. It is clear that all the experimental models perform much better with our proposed two prompting methods, Prompt-Ours and Pipeline-Ours, than other two prompting methods in terms of FusionEval, FacetEval, ROUGE. This show that prompting based on our proposed framework makes the models get similar meta-reviews to the human-written ones and achieve better sentiments in the generated meta-reviews. We find that GPT-4 is still the best model in the scientific sentiment summarization, and the steps generated by the model itself cannot always get better results than the naive prompt. Although Pipeline-Ours can get better results in terms of FusionEval and FacetEval, it performs much worse in terms of ROUGE. This is because it only considers the sentiments rather than other factual summary

[13]We only use GPT-4 to automate the extraction of judgements, as we find that other models are not well following the instructions.

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

[15]gpt-4-0613, gpt-3.5-turbo-1106, LLaMA2-70B-Chat, LLaMA2-7B-Chat

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 the naive prompting
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 the proposed prompting based on our sentiment consolidation framework
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 8: 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.)

of the manuscript, which leads to low ROUGE.

5.3 Case Study on Generated Meta-Reviews

To dive deeper into what difference the integration of sentiment consolidation logic makes, we also conduct a case study on generated results 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 and in machine-generated meta-reviews there are many more details which are unnecessary to form an effective meta-review. As shown in the example in Table 8, details about such as "PGD" or "CIFA-10" are not essential to form the meta-review.

We also find that prompts based on our sentiment consolidation framework prefer to cover more balanced judgements based on annotation like in Table 8 and compare them, while the naive prompting only mentions some sentiments. For example, in Table 8, the generation does not include positive sentiments for *Clarity* and only mentions some issues, but the generation by the promting based on our sentiment consolidation framework acknowledges the good clarity while also mentioning some clarity problems. This is consistent with that prompting with naive prompts get worse sentiments than prompting with the prompt based on the sentiment consolidation logic in Table 7.

6 Conclusions and Future Work

We initialize the task of scientific sentiment summarization. Based on the nature of the meta-reviewing process, we propose the three-layer framework of sentiment consolidation. With the framework, we propose automatic evaluation metrics to evaluate sentiments in generated summaries. Extensive experiments show that generation results of LLMs could be improved by integrating the sentiment consolidation framework into the textual prompting or the generation pipeline. 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 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 logic 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.

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.

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.

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.

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

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.

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 9: Review guidelines from different academic presses.

B Annotation Instructions for Human Annotation

The screen shots of the two-page annotation instruction for human annotation are shown in Figure 4 and Figure 5.

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 highlighting behaviours of any two annotators. This agreement shows whether annotators identify sentiments from similar text spans.

For Criteria 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 a 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 10, Table 11, and Table 12. Averaged agreement of GPT-4 with the two human annotators are present in Table 13, Table 14, and Table 15.

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

Table 10: 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
Criteria Facet	0.783	0.783
Sentiment Level	0.844	0.844
Convincingness Level	0.405	0.398

Table 11: 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
Criteria Facets	0.857	0.857
Sentiment Levels	0.764	0.763
Convincingness Levels	0.455	0.437

Table 12: Human annotator agreement on annotating discussions.

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

```
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 criteria 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"}

```

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

Table 13: 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
Criteria Facets	0.579	0.637	0.608
Sentiment Levels	0.594	0.589	0.592
Convincingness Levels	0.008	0.013	0.011

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

```
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).
```

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

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

```

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 criteria 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 Criteria Facets, Sentiment Levels, and Convincingness Levels;
Candidate criteria 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 criteria facets;

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

5. Aggregate judgements in different criteria 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 criteria facets for a manuscript.
```

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.