A Sentiment Consolidation Framework for Meta-Review Generation

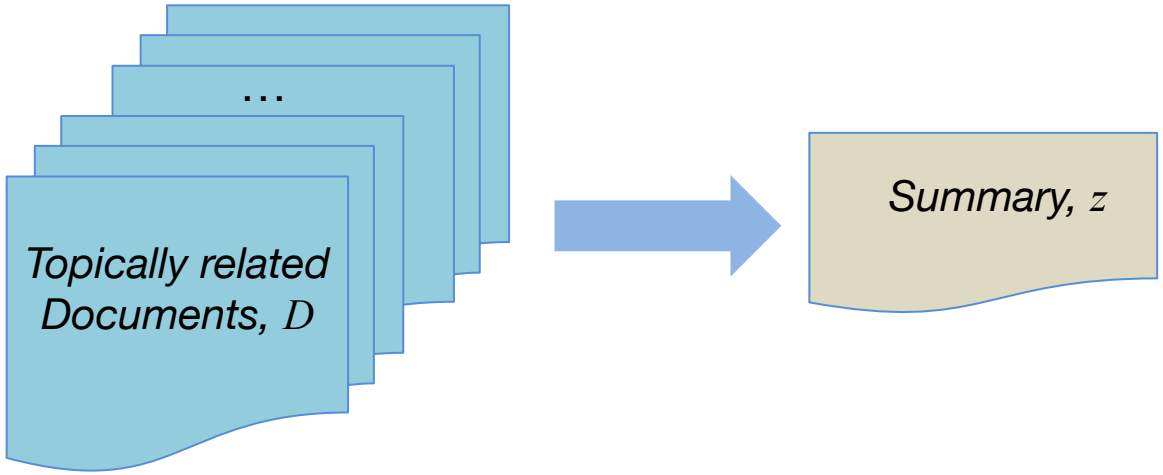
Miao Li[1], Jey Han Lau[1], Eduard Hovy[1,2]

[1]The University of Melbourne [2]Carnegie Mellon University

August 2024 Bangkok, Thailand

Multi-Document Summarization

- Lengthy input composed of multiple documents
- Complex reasoning of information consolidation
- Multi-channel nature of text summarization



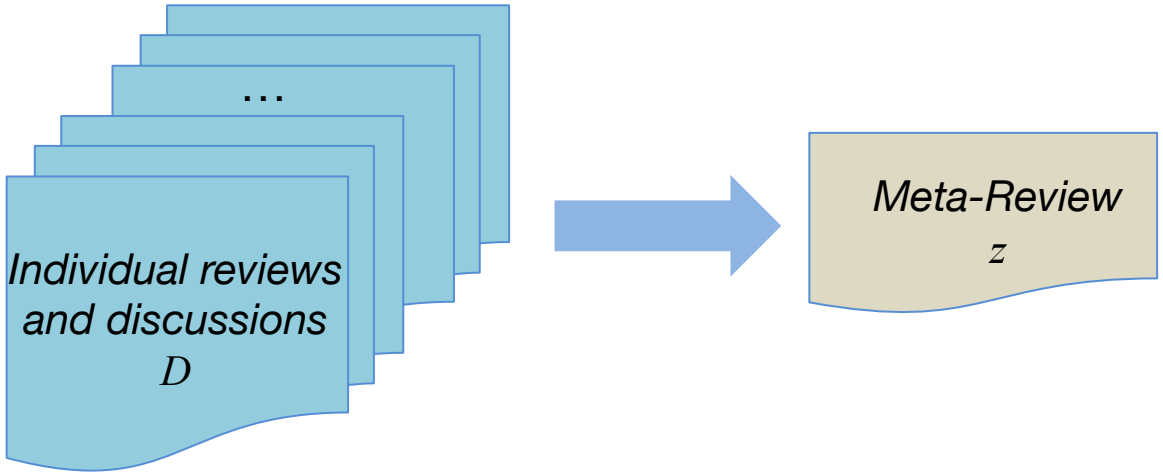
Lacking Understanding of Underlying Processes

- Existing models are mostly black boxes
  - Uncertain if models truly posses the ability of multi-document information consolidation

- Evaluation metrics only focus on textual quality without considering the underlying reasoning process
  - Popular metrics
    - ROUGE (Lin et al. 2001), BERTScore (Zhang et al. 2021), UniEval (Zhong et al. 2022), SummaC (Laban et al. 2021)
  - Quality aspects
    - Relevance, Coherence, Fluency, Consistency

Scientific Meta-Review Generation

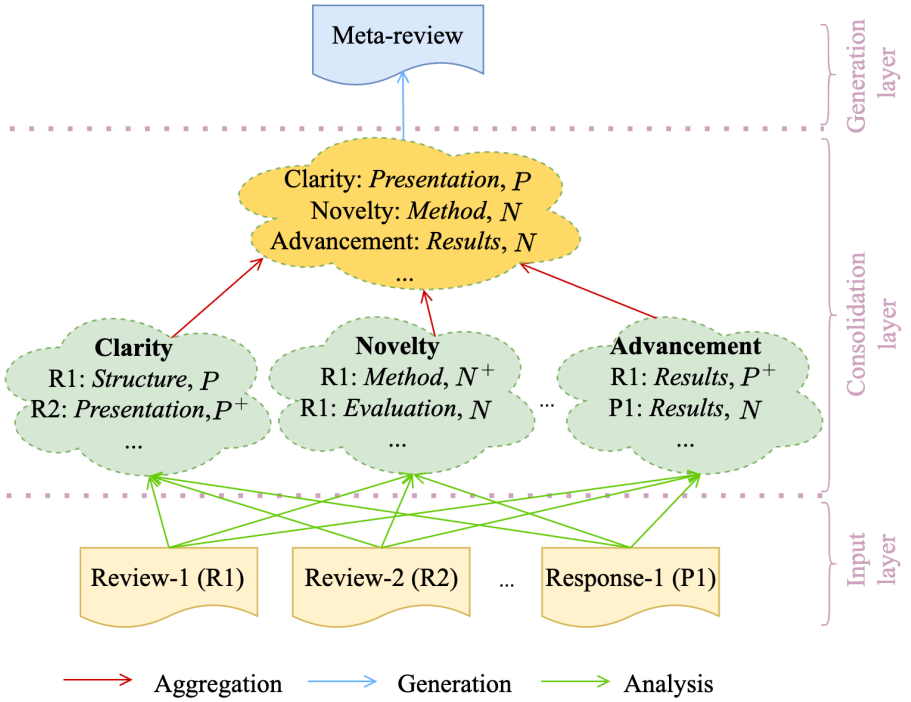
- Complex sentiment consolidation process
- Sentiment is the most important information channel



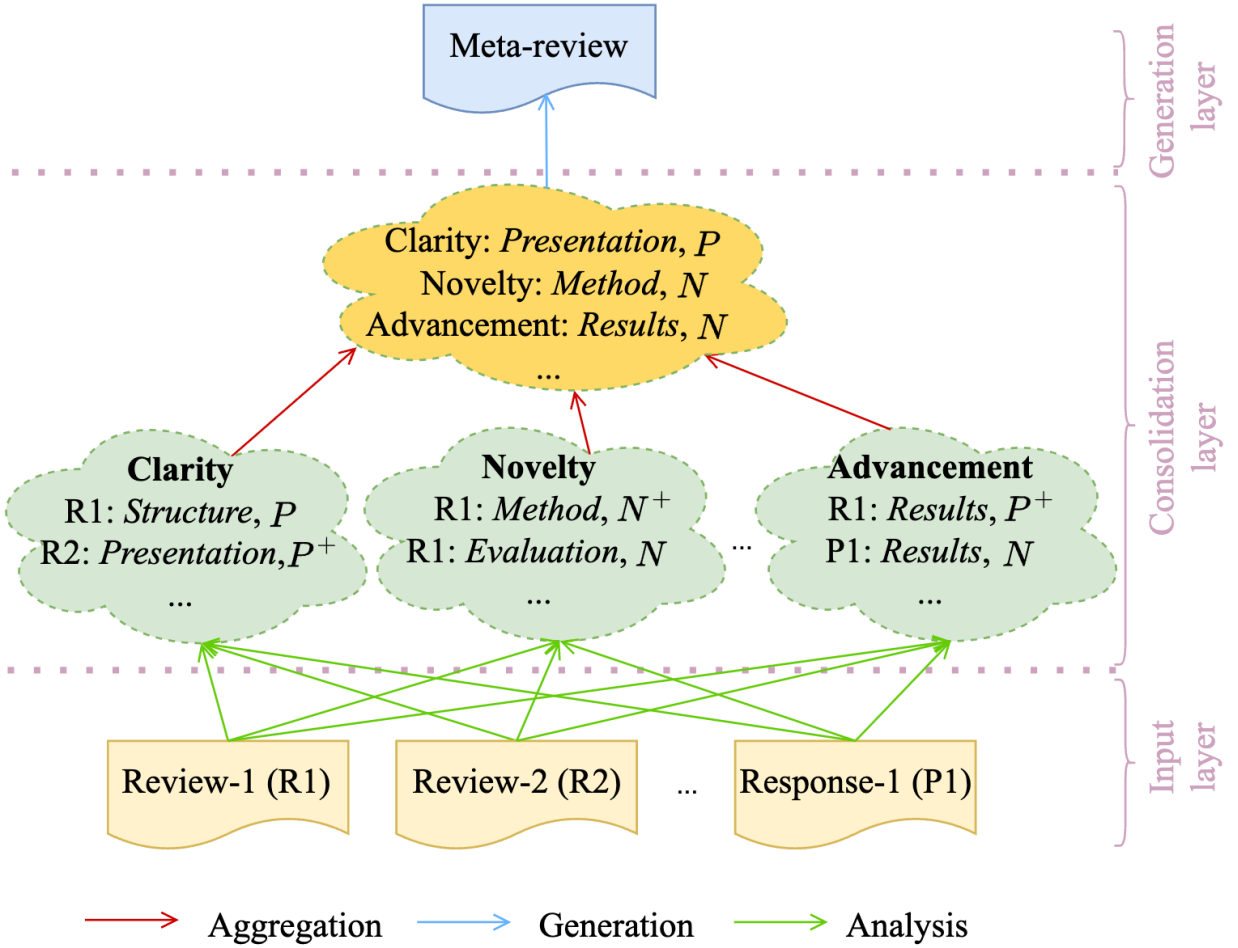
Human Consolidation of Scientific Sentiments

- Meta-reviewers should follow a sentiment aggregation logic
  - Six review facets that meta-reviewers and reviewers focus on
  - A three-layer sentiment consolidation framework

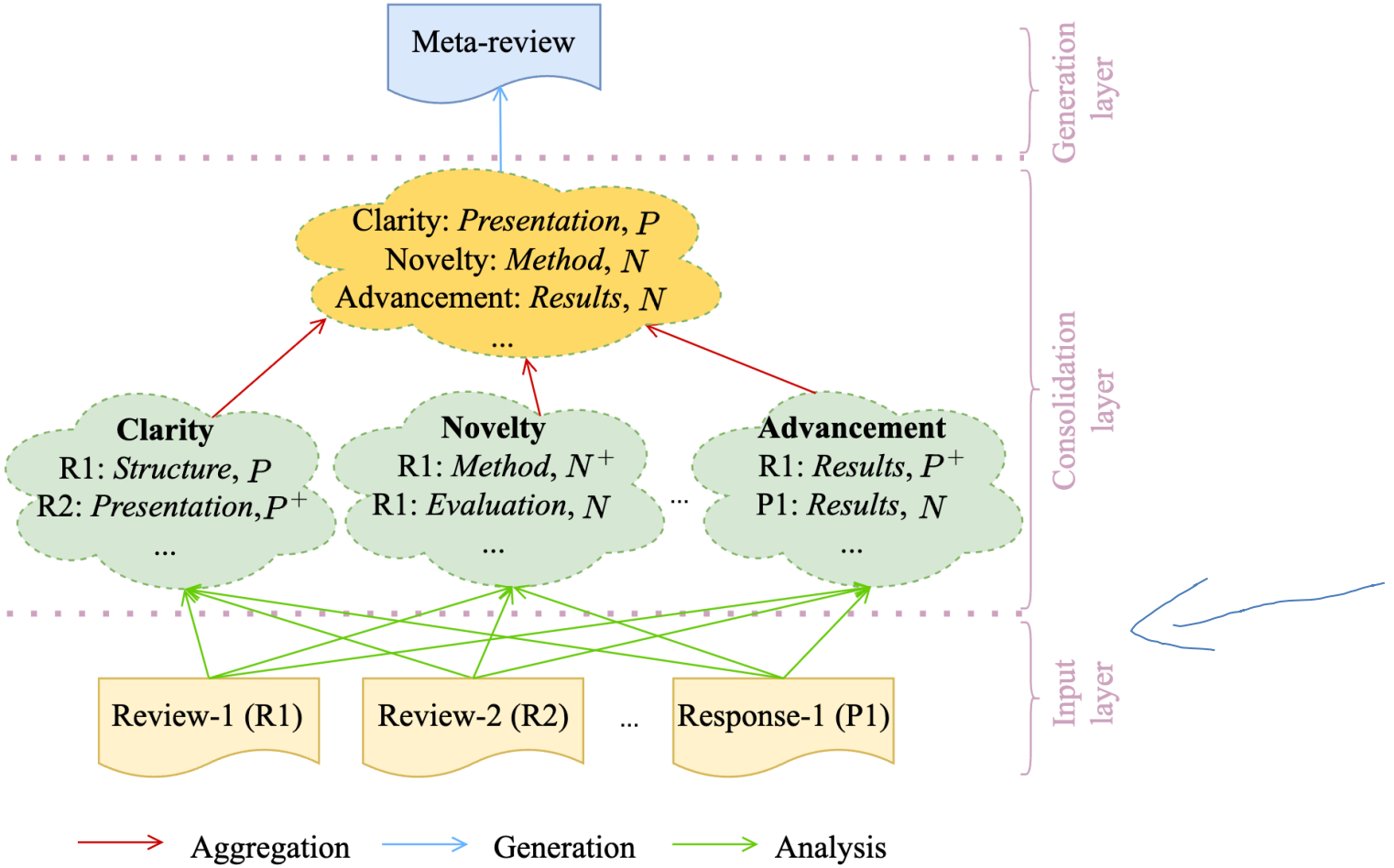
Facet	Definition
Novelty	How original the idea (e.g., tasks, datasets, or methods) is, and how clear where the problems and methods sit with respect to existing literature (i.e., meaningful comparison).
Soundness	There are usually two types of soundness: (1) Empirical: how well experiments are designed 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.



Sentiment Identification and Extraction



Sentiment Identification and Extraction



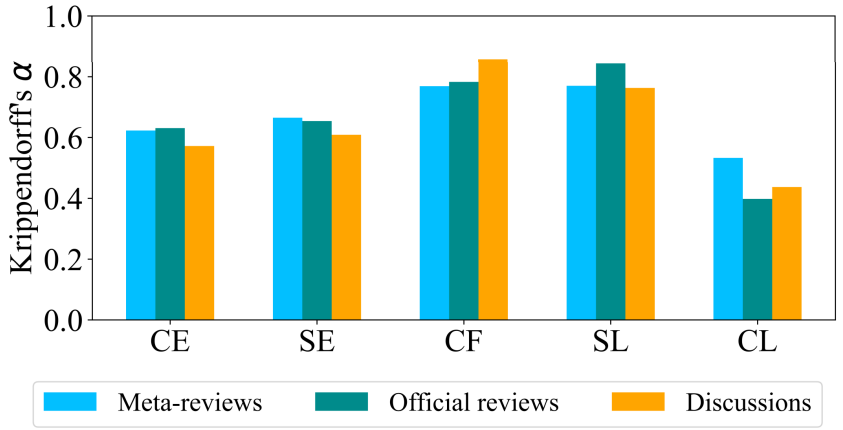
GPT-4 Judgement Identification and Extraction

- Human annotation
  - Two PhD students on 30 samples
  - Each sample cost one hour for each annotator

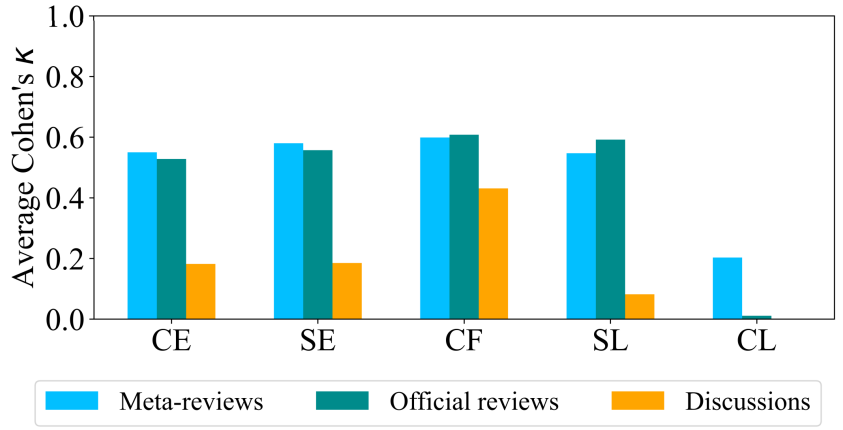
Component	Definition
Content Expression	What the sentiment is talking about
Sentiment Expression	The value of the sentiment
Criteria Facet	The specific criteria facet that the judgement belongs to
Sentiment Level	The polarity and strength of the sentiment
Convincingness Level	How well the sentiment is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

Table 1: Definitions of components in a judgement.

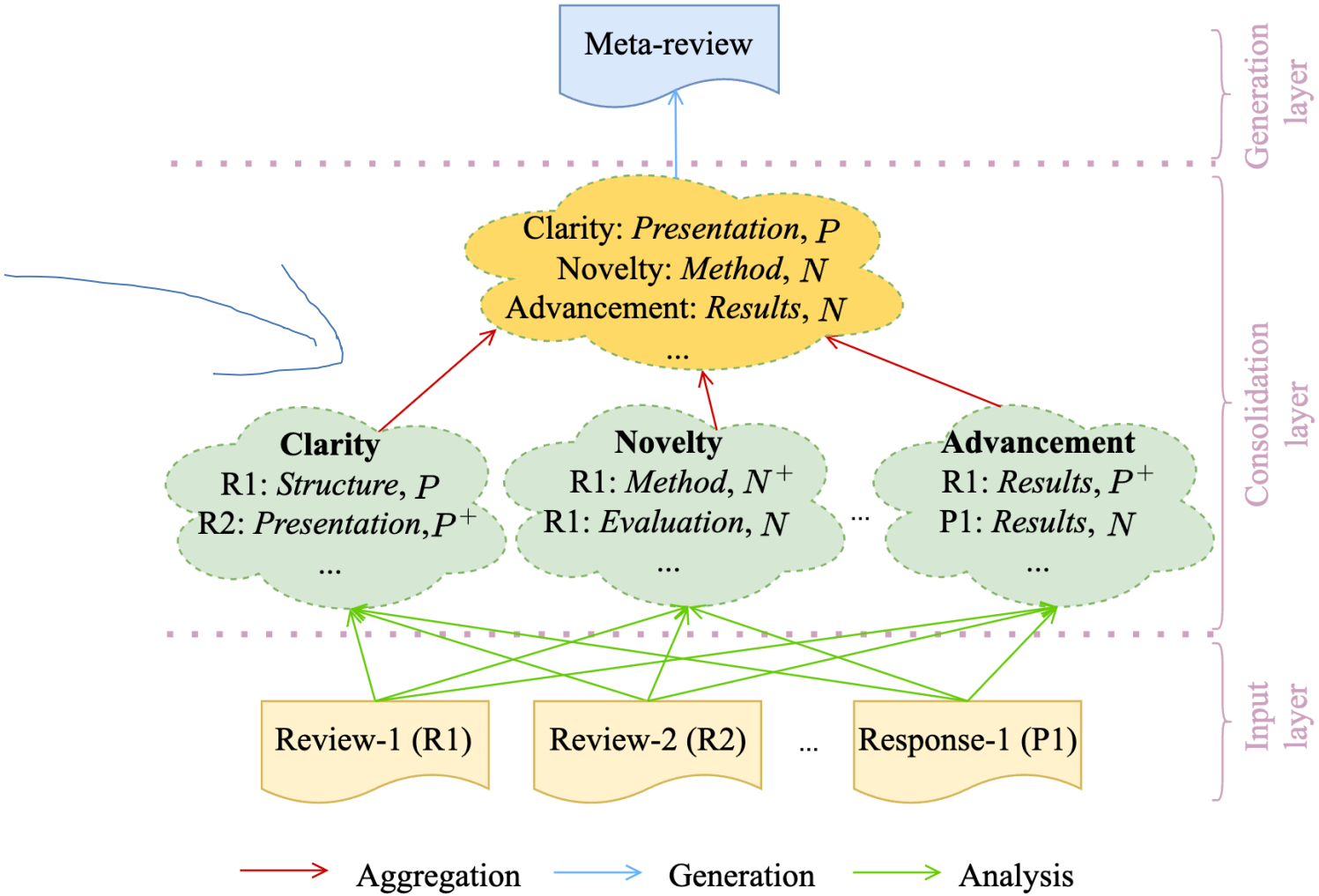
- Human inter-annotator agreement



- GPT-4 agreement with humans



Sentiment Aggregation



GPT-4 for Sentiment Aggregation

- Human meta-reviewers are not always following majority voting
  - Meta-review level: 23.7% not majority voting
  - Sentiment level, an example

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"

- Predict meta-review sentiments with GPT-4

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

Predicting with human-annotated judgements vs full texts

Sentiment-Aware Evaluation

- FacetEval
  - Similarity between the sentiment distributions in the human-written and model-generated meta-reviews
  - Reference-based

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

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

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.

- FusionEval:
  - Sentiment fusion correctness for individual facets
  - Reference-free

Evaluating steps

- *Extracting judgements from a generated meta-review*
  - *For each judgement, there are Content Expression, E and Sentiment Level, L*
- *Zero-shot Predicting sentiment level L' based on source judgements*
- *Accuracy between L and L' is the score*

Enhancing LLMs with Explicit Consolidation

- Enhancing LLMs with integrating the sentiment aggregation logic
- Based on our framework and experiments, Prompt-Ours can follow

- Step 1: *Extracting content and sentiment expressions of judgements in all source documents;*
- Step 2: *Predicting Criteria Facets, Sentiment Levels, and Convincingness Levels;*
- Step 3: *Reorganize extracted judgements in different clusters for different criteria facets;*
- Step 4: *Generate a small summary for judgements on the same criteria facet with sentiment comparison and aggregation;*
- Step 5: *Generate the final meta-review based on summaries for different criteria facets.*

- Compared with other prompting methods
  - Prompt-LLM: The prompt is generated by a powerful LLM

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. ($\times 0.01$)

Reference-Free Human Evaluation

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.

Case Study on Generated Meta-Reviews

- Model-generated meta-reviews are longer than human-written ones
  - There are unnecessary details, e.g., “CIFAR-10”
- Generations with our prompts prefer to cover more balanced sentiments and compare them

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.)

Thanks!
Questions & Answers

https://oaimli.github.io
miao4@student.unimelb.edu.au