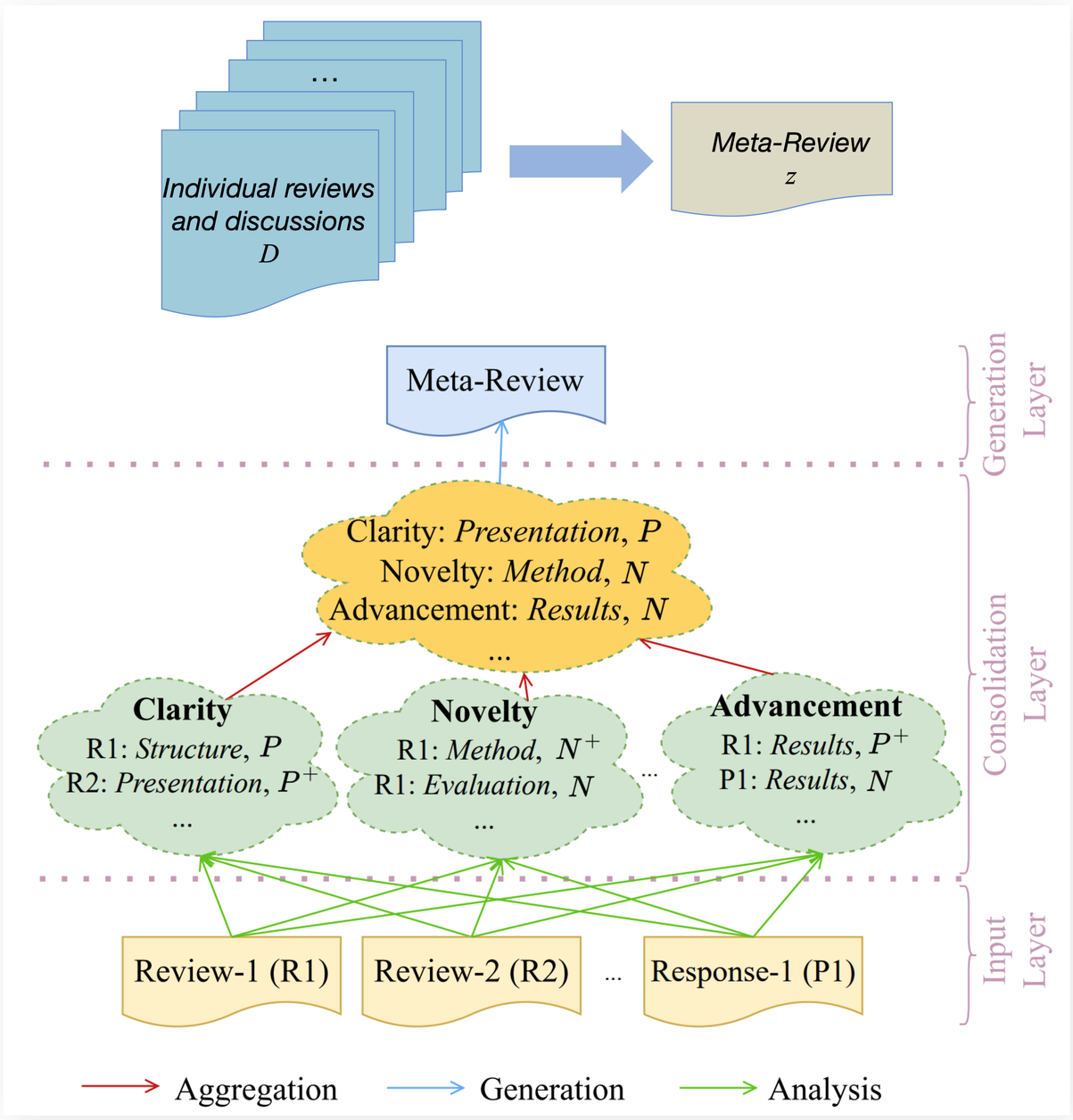
Meta-reviewers follow **a three-layer framework** to **aggregate opinions** of individual reviewers for scientific manuscripts

A Sentiment Consolidation Framework for Meta-Review Generation

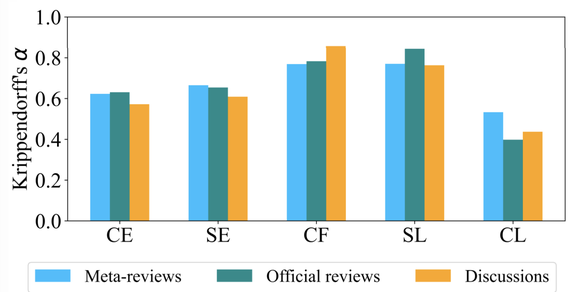
Miao Li[1], Jey Han Lau[1], Eduard Hovy[1,2]

Hierarchical Sentiment Consolidation

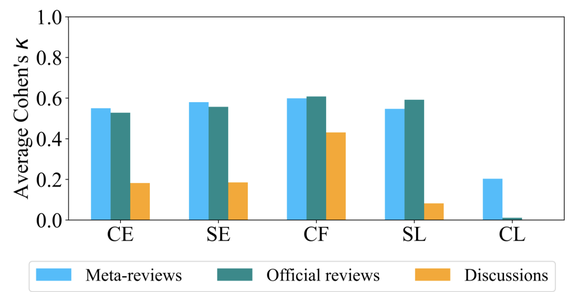


GPT-4 Sentiment Extraction and Fusion

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



Human inter-annotator agreement



GPT-4 agreement with humans

Reference-Free and Reference-Based Auto Evaluation

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

Typology of Review Facets for Scientific Manuscripts

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.

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%	

The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)
August 2024 Main conference, Long paper

THE UNIVERSITY OF MELBOURNE
Carnegie Mellon University