
ReFeR: A Hierarchical Framework of Models as Evaluative and Reasoning Agents

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Assessing the quality of Natural Language Generation (NLG) outputs, such as
2 those produced by large language models (LLMs), poses significant challenges.
3 Human evaluations are not scalable, and traditional automatic metrics exhibit low
4 correlation with human judgment. In this study, we propose Review-Feedback-
5 Reason (ReFeR), a novel evaluation framework for NLG using LLM agents. The
6 proposed framework enhances the accuracy of NLG evaluation, surpassing previous
7 benchmarks by $\sim 20\%$. Moreover, feedback collected from our framework is then
8 leveraged to instruction fine-tune smaller models like Mistral-7B, yielding a better
9 correlation with human evaluations and performance nearly on par with GPT-3.5.
10 We highlight another ancillary benefit of our methodology through its application
11 on reasoning benchmarks, outperforming most of the state-of-the-art methods and
12 also beating GPT-3.5 Turbo by $\sim 11.67\%$ and GPT-4 by $\sim 1\%$ on an average.

13 1 Introduction

14 The rapid production of content by Foundation Models (FMs) [Bommasani et al., 2021], poses
15 challenges to human-centric evaluation methods and conventional linguistic metrics like BLEU,
16 ROUGE, and METEOR [Papineni et al., 2002, Lin, 2004, Banerjee and Lavie, 2005], which often
17 misalign with human judgment. Recent developments suggest using LLMs as reference-independent
18 evaluators by assessing text quality based on predicted sequence likelihoods [Chen et al., 2023] and
19 works [Liu et al., 2023b, Chiang and Lee, 2023] on improving the evaluation capability of individual
20 LLMs. Surprisingly, although an ensemble of multiple LLMs is expected to perform better, there has
21 not been much work on these lines.

22 We thus introduce the Review-Feedback-Reason (ReFeR) framework, by using LLMs as evaluators
23 and feedback providers in a system akin to academic peer review, ReFeR enables a nuanced and
24 comprehensive evaluation of NLG tasks across various domains, promoting self-improvement,
25 explainability, and robustness in complex scenarios. The paper outlines ReFeR’s methodology,
26 including its unique evaluation schema (that diverges from existing benchmarks as outlined by Liu
27 et al. [2023b] and Chiang and Lee [2023]) and the strategic use of LLM agents in roles parallel to peer
28 reviewers and area chairs, facilitating hierarchical evaluation and generating constructive feedback
29 for model refinement.

30 The primary contributions of our research are as follows: (1) Introducing ReFeR, a NLG evaluation
31 framework inspired by academic peer review system. (2) Development of a novel evaluation schema,
32 incorporating an evaluation guidelines module and a critical comments module. (3) Creation of an
33 automated instruction tuning dataset from the framework’s feedback outputs, designed to enhance
34 smaller models. (4) Empirical validation of the framework’s capability to show enhanced reasoning
35 skills.

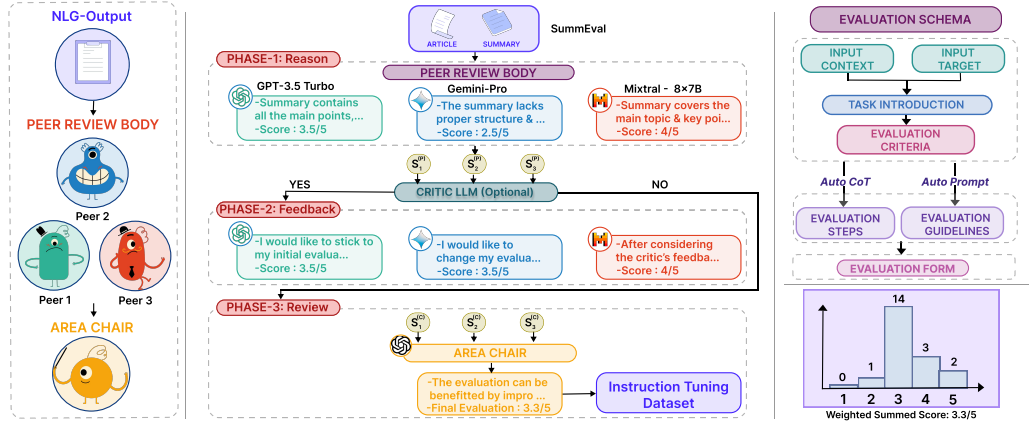


Figure 1: ReFeR Framework on SummEval dataset. A case study example is given in Appendix I

2 ReFeR Methodology for NLG Evaluation

In this section, we introduce ReFeR, a versatile framework for evaluating NLG tasks by using LLM agents in roles analogous to peer reviewers and area chairs, as found in the academic peer review process.

2.1 ReFeR Framework

The challenge of evaluating work without a predefined correct answer, such as determining the quality of a research paper, is traditionally addressed in academia through the peer review system. In this process, subject-matter experts called peer reviewers, independently review submissions. Authors then have an opportunity to address any concerns raised. Finally, senior researchers serving as area chairs review the adjusted feedback and make the final decision on whether to accept or reject the submission. Our framework draws inspiration from this process, and aims to evaluate NLG outputs replicating this academic review methodology.

The framework is structured into three distinct modules, as depicted in Fig. 1. The first module, the Peer Review Body, consists of three LLM agents. Each agent independently evaluates a specific NLG output, providing a comment and a rating. The following module is the Critic Module (optional), wherein another LLM agent, emulating a critic, assesses the evaluations made by the peer reviewers. The peer reviewers can then revisit these interactions and can adjust their assessments before forwarding their final reviews to the Area Chair Module. The final module features an LLM agent acting as an Area Chair, who considers the conclusive reviews to perform the ultimate evaluation of the NLG output. We **R**ead using the LLM Agents as peers and Area Chairs, take **F**eedback of peers and pass it to the area chair and finally give a **R**eview or score to the NLG text. Hence our framework is named as **ReFeR**.

2.2 Evaluation Schema

An important aspect of assessing NLG outputs with LLM agents involves crafting prompts that elicit the highest quality evaluations. Prior work G-Eval by Liu et al. [2023b] introduced a structured evaluation schema, which organized the prompt into sections: task introduction, evaluation criteria, steps for evaluation, input presentation, and an evaluation form designed to output a numerical rating. Subsequently, Chiang and Lee [2023] demonstrated that a Chain of Thought (CoT) approach does not consistently yield the most accurate correlations with human judgment. They proposed an adjusted schema named Analyze-Rate, which prioritizes an analytical review followed by the scoring. This method showed improved performance over the G-Eval schema.

To further refine this approach, we introduce "evaluation guidelines" to enhance the peer reviewer's understanding of the scoring criteria, much like guidelines provided in traditional academic review processes. This modification posits that clear guidelines can improve evaluation accuracy by standardizing the scoring rationale. Evaluation guidelines can be automatically generated by prompting

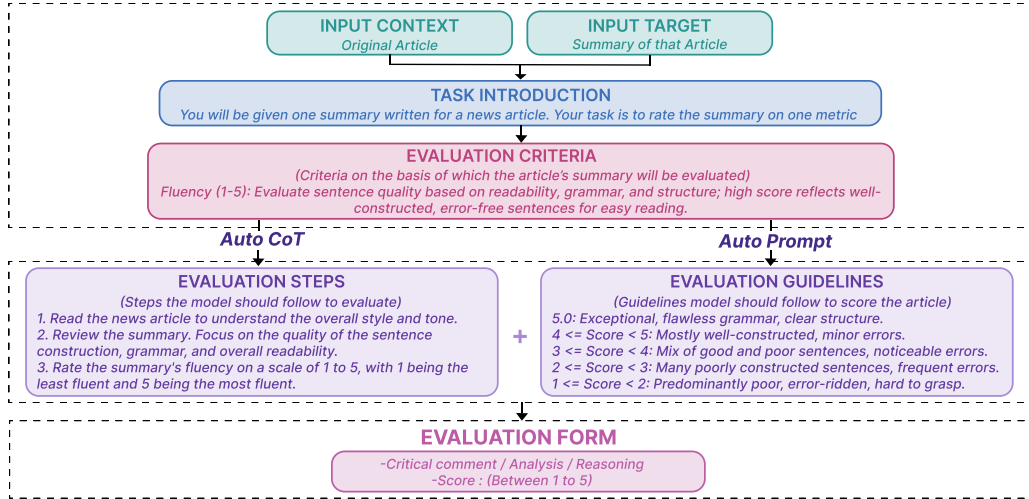


Figure 2: Evaluation Schema for ReFeR’s prompt.

71 an LLM with examples from the dataset. We call this process ‘Auto Prompt’. Another possible way
 72 to include evaluation guidelines is to use manually written human annotation guidelines of the dataset.
 73 We also changed the evaluation form to include a critical comment or reasoning for the given score.
 74 The proposed evaluation schema is shown in Fig. 2. This method has improved the performance, as
 75 was previously shown by Chiang and Lee [2023].

76 3 Experiments and Results for NLG Evaluation

77 3.1 Baselines

78 While the current landscape of models for evaluating NLG responses includes reference-free methods
 79 such as BERTScore, GPTScore and UniEval [Zhang et al., 2020, Fu et al., 2023, Zhong et al., 2022],
 80 we do not consider these models as since they were clearly surpassed by G-Eval [Liu et al., 2023b]
 81 and later works. Given our work primarily proposes a LLM based evaluation, we do a comparative
 82 analysis primarily against G-Eval [Liu et al., 2023b] and Analyze-Rate [Chiang and Lee, 2023].
 83 **G-Eval** uses a single LLM with Auto-CoT reasoning and a form-filling approach to evaluate NLG
 84 outputs. **Analyze-Rate** enhances this by adding a preliminary analysis phase before scoring to
 85 improve the evaluation process.

Table 1: Spearman Correlation (ρ) and Kendall-Tau Correlation (τ) on SummEval. The best results per column have been highlighted in bold.

	Models	Coherence		Consistency		Fluency		Relevance		Avg	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Baselines	GPT-3.5	0.354	0.288	0.311	0.283	0.289	0.249	0.283	0.232	0.309	0.263
	Mixtral	0.416	0.333	0.385	0.345	0.350	0.314	0.367	0.303	0.380	0.324
	Gemini	0.341	0.266	0.319	0.296	0.166	0.142	0.352	0.205	0.295	0.227
	Analyze-Rate (GPT-3.5, n=20)	0.558	0.413	0.404	0.327	0.394	0.312	0.442	0.328	0.449	0.345
	G-Eval (GPT-3.5, n=20)	0.420	0.311	0.287	0.234	0.310	0.228	0.421	0.315	0.359	0.272
	ReFeR(Ours)	0.562	0.413	0.406	0.327	0.411	0.328	0.509	0.379	0.472	0.362

86 3.2 Experimental Setup

87 Due to G-Eval not releasing TopicalChat prompts, Chiang and Lee [2023] created new ones based on
 88 the original G-Eval design, which we used in our experiments to ensure consistency and address the
 89 impact of prompt changes on results.

90 Our experimental framework employs GPT-3.5 Turbo (2023-06-13) [OpenAI, 2023], Gemini-Pro
 91 [Team et al., 2023], and Mixtral 8x7B [Jiang et al., 2024] as peer evaluators, and GPT-3.5 Turbo (with

Table 2: Spearman Correlation (ρ) and Kendall-Tau Correlation (τ) on TopicalChat. Best and second-best per column have been highlighted with bold and underline respectively. ReFeR results are without Critic LLM. For ReFeR with different Critic results see Appendix H.1.

	Models	Coherence		Engagingness		Groundedness		Naturalness		Avg	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Baselines	(Peer) GPT-3.5	0.417	0.350	0.519	0.439	0.527	0.493	0.416	0.348	0.470	0.407
	(Peer) Mixtral	0.424	0.358	0.532	0.456	0.443	0.410	0.451	0.376	0.463	0.400
	(Peer) Gemini	0.363	0.303	0.477	0.398	0.539	0.504	0.398	0.333	0.444	0.385
	Analyze-Rate (GPT-3.5, n=20)	0.506	0.384	0.637	0.480	0.646	0.546	0.522	0.391	0.578	0.450
	G-Eval (GPT-3.5, n=20)	0.472	0.356	0.618	0.474	0.456	0.377	0.501	0.373	0.512	0.395
	ReFeR (Ours)	0.514	0.390	0.651	0.502	0.678	0.590	0.544	0.414	0.597	0.474

92 number of responses generated per prompt $n = 20$) acting solely as the Area Chair and Critic LLM.
 93 Appendix E provides details of LLM hyper-parameters. Following [Fu et al., 2023, Liu et al., 2023b],
 94 we primarily report Spearman correlations (ρ) between the scores generated by our framework and
 95 those annotated by humans and use this as the primary differentiator to find the best model.

96 **3.3 Main Results for NLG Evaluation**

97 We assess ReFeR’s performance through a series of experiments, employing a diverse array of LLM
 98 agents as peers and an Area Chair.

99 Tables 1 and 2 show ρ and τ for SummEval and TopicalChat, respectively. Results are shown for
 100 individual performance by each of the 3 peers, and G-Eval and Analyze-Rate (both with GPT-3.5,
 101 n=20) as baseline methods. Our framework’s results here does not use the Critic LLM Phase but
 102 we show results using 4 variations of the ReFeR framework with critic phase in Appendix H.1. By
 103 juxtaposing the ReFeR framework’s outcomes against those derived from G-Eval, Analyze-Rate
 104 and contrasting these findings with individual peers’ scores, we get insights into the substantial
 105 enhancements by our framework. Specifically, ReFeR surpasses the average Spearman correlation by
 106 $\sim 20\%$ on the SummEval dataset and by $\sim 3\%$ on the TopicalChat dataset when compared to best
 107 baseline performances.

108 We chose to report main results without the (optional) critic module because as highlighted by Laban
 109 et al. [2024], existing LLMs often exhibit fluctuating stances under scrutiny regarding their response
 110 confidence, suggesting a propensity for opinion revision. This observation implies that, for optimal
 111 correlation scores, it may be advantageous to bypass the critic module until such foundational issues
 112 within LLMs are addressed, at which point its incorporation could yield further benefits. However,
 113 for completeness, we report results using multiple critic LLMs in Appendix H.1.

114 Further, prompt sensitivity is a fundamental constraint of LLMs [Sclar et al., 2024, Loya et al.,
 115 2023]. A poorly constructed prompt can skew results, leading to outcomes that deviate from expected
 116 benchmarks. This observation is further validated by the findings of Chiang and Lee [2023] in the
 117 automated evaluation domain. When the prompt is not optimized the results can be very misleading
 118 and can confuse researchers with the thinking that a certain method is not accurate. So to get any
 119 conclusive results, we always need a very well-crafted, manually engineered prompt.

120 We also investigated the best prompt that should be used for each LLM Agent. We add all prompt
 121 and performance ablations in the Appendix G due to constraints of space in the main paper.

122 **4 Instruction-Tuning of Small LLMs using Area Chair Outputs**

123 We enhance smaller LLMs through instruction-tuning using feedback from larger LLMs (“Area
 124 Chairs”) within the ReFeR framework. This fine-tuning improves the performance of smaller models
 125 like Mistral-7B, making them competitive with larger models. For training, we used the same 200
 126 test samples from SummEval and 45 from TopicalChat, with the remaining data used for training and
 127 development. Mistral-7B was chosen due to its lower operational cost compared to GPT-3.5 Turbo.
 128 A case study example is given in Appendix J.

129 Notably, the fine-tuned Mistral-7B model clearly surpasses the baseline established by its non-fine-
 130 tuned counterpart. These results also illustrate the competitive edge that fine-tuned, smaller models
 131 gain against the considerably larger and more resource-intensive GPT 3.5 Turbo model (25-fold

Table 3: Performance Comparison of Finetuned vs Non-Finetuned Models (Spearman Correlation (ρ) metric) on SummEval (left) and TopicalChat (right) datasets.

Models	Coh	Con	Flu	Rel	Avg	Models	Coh	Eng	Gro	Nat	Avg
Mistral-7B-non-finetuned	0.284	0.210	0.158	0.240	0.223	Mistral-7B-non-finetuned	0.136	0.205	0.086	0.087	0.128
GPT-3.5 (n=1)	0.357	0.363	0.237	0.279	0.309	GPT-3.5 (n=1)	<u>0.437</u>	0.531	0.497	0.544	0.502
Mistral-7B-finetuned (ReFeR)	0.372	0.255	0.289	0.258	0.293	Mistral-7B-finetuned (ReFeR)	0.544	<u>0.389</u>	0.287	0.385	0.401

132 larger size) by employing feedback-driven fine-tuning in lieu of relying on larger models like GPT-4
 133 for equivalent levels of evaluative accuracy.

134 5 Collective Reasoning through ReFeR

135 The ancillary benefit of our framework is its ability to enhance collective reasoning. We tested
 136 ReFeR’s reasoning performance aligned with the methodologies outlined by Chen et al. [2024]
 137 (scores reported on 100 test samples per dataset) on GSM8K [Cobbe et al., 2021] testing math
 138 reasoning, StrategyQA [Geva et al., 2021] testing deductive reasoning, and CSQA [Talmor et al.,
 139 2019] testing commonsense reasoning. We crafted prompts (Appendix M) mirroring our evaluation
 140 schema, with peer agents generating answers and corresponding reasoning that were subsequently
 141 consolidated by an area chair into a final response.

142 The results of our experiments are shown in Table 4, with ReConCile’s results sourced from [Chen
 143 et al., 2024], using GPT-3.5, Claude-2, and Bard. While a direct comparison isn’t feasible due to
 144 different peer groups, we assess how ReFeR’s collective intelligence surpasses individual agents
 145 reasoning limitations. Some examples of how ReFeR improves the collective reasoning of each peer
 146 is given in Appendix K.

Table 4: Accuracy (%) for the reasoning tasks across the GSM8k, StrategyQA and CSQA benchmarks.

Models	GSM8k	StrategyQA	CSQA
GPT-3.5	71	70	72
Mixtral	74	77	71
Gemini-Pro	77	71	73
Reconcile	85	79	75
GPT-4	95	72	78
ReFeR	87	81	80

147 We evaluated ReFeR’s collective reasoning against individual models and the ReConCile framework
 148 [Chen et al., 2024]. ReFeR outperforms ReConCile on StrategyQA (81.0 vs. 75.6) and CSQA (80
 149 vs. 74.7), while coming close to GPT-4 on GSM8K (87.0 vs. 95.0) and surpassing on GPT-4 on
 150 StrategyQA and CSQA. This demonstrates ReFeR’s strong reasoning ability and cost-efficiency
 151 compared to GPT-4 OpenAI [2024].

152 6 Conclusion

153 We introduce ReFeR, an NLG evaluation framework inspired by the academic peer review process,
 154 which enhances both evaluation and collective reasoning capabilities. ReFeR’s three-stage evaluation
 155 system—peer review, optional critic phase, and final evaluation by area chairs—demonstrates a 20%
 156 improvement on the SummEval dataset and 3% on TopicalChat, compared to previous benchmarks.
 157 The instruction-tuning datasets created using ReFeR further improve smaller models like Mistral-7B,
 158 achieving a 31% and 213% increase in performance on SummEval and TopicalChat, respectively.
 159 ReFeR also outperforms state-of-the-art reasoning frameworks on GSM8K, CSQA, and StrategyQA
 160 benchmarks and the ReFeR’s collective reasoning also outperformed models like GPT-3.5 Turbo by
 161 $\sim 11.67\%$ and GPT-4 by $\sim 1\%$ on an average.

162 We limited our experiments to a small number of LLMs and datasets, focusing on English-only
 163 evaluations. Future work can explore the impact of more peers, multi-lingual datasets, and additional
 164 LLMs. Prompt sensitivity remains a challenge, as suboptimal prompts significantly affect perform-
 165 ance. Additionally, improving communication strategies between LLM agents offers a promising
 166 direction for further research. We also introspected and added a Social Impact Statement in Appendix
 167 A.

168 **References**

- 169 Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved
170 correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss,
171 editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for*
172 *Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005.
173 Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909>.
- 174 Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx,
175 Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson,
176 S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel,
177 Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon,
178 John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie,
179 Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter
180 Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard,
181 Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte
182 Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya
183 Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li,
184 Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell,
185 Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Benjamin Newman, Allen
186 Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel
187 Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan,
188 Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher
189 R’e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan,
190 Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang,
191 Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A.
192 Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou,
193 and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL
194 <https://crfm.stanford.edu/assets/report.pdf>.
- 195 Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and
196 Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023.
- 197 Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference
198 improves reasoning via consensus among diverse llms, 2024.
- 199 Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. Exploring the use of large
200 language models for reference-free text quality evaluation: An empirical study. In Jong C. Park,
201 Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors,
202 *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages
203 361–374, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. URL
204 <https://aclanthology.org/2023.findings-ijcnlp.32>.
- 205 Cheng-Han Chiang and Hung-yi Lee. A closer look into automatic evaluation using large language
206 models, 2023.
- 207 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
208 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
209 Schulman. Training verifiers to solve math word problems, 2021.
- 210 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving
211 factuality and reasoning in language models through multiagent debate. *ArXiv*, abs/2305.14325,
212 2023. URL <https://api.semanticscholar.org/CorpusID:258841118>.
- 213 Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and
214 Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021.
- 215 Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023.
- 216 Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle
217 use a laptop? a question answering benchmark with implicit reasoning strategies, 2021.

- 218 Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu
219 Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded
220 Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895, 2019. doi: 10.21437/
221 Interspeech.2019-3079. URL <http://dx.doi.org/10.21437/Interspeech.2019-3079>.
- 222 Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit
223 Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the
224 solution to scaling up multilingual evaluation?, 2023.
- 225 Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa
226 Suleyman, and Phil Blunsom. Teaching machines to read and comprehend, 2015.
- 227 Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris
228 Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand,
229 Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-
230 Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le
231 Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed.
232 Mixtral of experts, 2024.
- 233 Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhui Chen. Tigerscore:
234 Towards building explainable metric for all text generation tasks, 2023.
- 235 Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of
236 translation quality, 2023.
- 237 Philippe Laban, Lidiya Murakhovs’ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure?
238 challenging llms leads to performance drops in the flipflop experiment, 2024.
- 239 Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao.
240 Multimodal foundation models: From specialists to general-purpose assistants, 2023.
- 241 Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summariza-
242 tion Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational
243 Linguistics. URL <https://aclanthology.org/W04-1013>.
- 244 Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and
245 Lifu Huang. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning
246 with auxiliary evaluation aspects, 2023a.
- 247 Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg
248 evaluation using gpt-4 with better human alignment, 2023b.
- 249 Manikanta Loya, Divya Anand Sinha, and Richard Futrell. Exploring the sensitivity of LLMs’
250 decision-making capabilities: Insights from prompt variations and hyperparameters. In *The
251 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=noPuQXVx8Y>.
- 253 OpenAI. GPT-3.5 Turbo: Language models by openai. [https://platform.openai.com/docs/
254 models/gpt-3-5-turbo](https://platform.openai.com/docs/models/gpt-3-5-turbo), 2023.
- 255 OpenAI. Gpt-4 technical report, 2024.
- 256 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic
257 evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for
258 Computational Linguistics, ACL ’02*, page 311–318, USA, 2002. Association for Computational
259 Linguistics. doi: 10.3115/1073083.1073135. URL [https://doi.org/10.3115/1073083.
260 1073135](https://doi.org/10.3115/1073083.1073135).
- 261 Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model
262 connected with massive apis, 2023.
- 263 Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer,
264 Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through
265 embeddings, 2023.

- 266 Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke
267 Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach
268 themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*,
269 2023. URL <https://openreview.net/forum?id=Yacmpz84TH>.
- 270 Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity
271 to spurious features in prompt design or: How i learned to start worrying about prompt formatting.
272 In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RIu51yNXjT>.
- 274 Lei Shu, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, and Lei Meng.
275 Fusion-eval: Integrating evaluators with llms, 2024.
- 276 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question
277 answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and
278 Tamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter
279 of the Association for Computational Linguistics: Human Language Technologies, Volume 1
280 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for
281 Computational Linguistics. doi: 10.18653/v1/N19-1421. URL [https://aclanthology.org/
282 N19-1421](https://aclanthology.org/N19-1421).
- 283 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu,
284 Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav
285 Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen,
286 Emily Pitler, Timothy Lillicrap, and many others. Gemini: A family of highly capable multimodal
287 models, 2023.
- 288 Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu,
289 Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study, 2023.
- 290 Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the
291 emergent cognitive synergy in large language models: A task-solving agent through multi-persona
292 self-collaboration, 2024.
- 293 Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. To-
294 wards reasoning in large language models via multi-agent peer review collaboration. *ArXiv*,
295 abs/2311.08152, 2023. URL <https://api.semanticscholar.org/CorpusID:265157805>.
- 296 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating
297 text generation with bert, 2020.
- 298 Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and
299 Jiawei Han. Towards a unified multi-dimensional evaluator for text generation, 2022.

300 A Social Impacts Statement

301 The ReFeR framework presents significant potential for the future of Natural Language Generation
302 (NLG) evaluation and reasoning tasks. By emulating an academic peer review process, this system
303 leverages large language models (LLMs) to enhance both the accuracy and interpretability of auto-
304 mated evaluations, surpassing traditional benchmarks. The broader social impact of this work lies
305 in its ability to democratize access to high-quality model evaluation, enabling smaller models to
306 perform evaluations on par with larger, resource-intensive models. This reduces computational costs
307 and environmental footprints, promoting sustainable AI practices.

308 Moreover, ReFeR’s capacity for constructive feedback generation has implications for improving the
309 explainability and transparency of AI systems, which is critical for building trust in applications de-
310 ployed in high-stakes domains like healthcare, education, and legal decision-making. The framework
311 also fosters collective reasoning, providing a more holistic evaluation of AI outputs, which could
312 prevent biased or inaccurate evaluations that single-model approaches might miss.

313 However, we acknowledge the potential risks associated with misuse, such as manipulating the system
314 to generate biased or unjust evaluations. Ongoing work is essential to ensure that ethical guidelines

315 are followed in deploying ReFeR, especially in contexts where the results of AI evaluations directly
316 impact human lives or societal outcomes.

317 **B Datasets**

318 Following previous works [Zhong et al., 2022, Liu et al., 2023b, Chiang and Lee, 2023], our meta-
319 evaluations predominantly utilize two datasets (SummEval, TopicalChat), each designed to test
320 distinct evaluation capabilities of our framework for Summarization and Dialog Generation tasks.

321 **SummEval** [Fabbri et al., 2021] provides human assessments on four critical dimensions of summa-
322 rization quality: fluency, coherence, consistency, and relevance, utilizing the CNN/DailyMail dataset
323 [Hermann et al., 2015] as its foundation. Considering computational budget and time constraints, we
324 selected only 200 samples for our experiments.

325 **TopicalChat** [Gopalakrishnan et al., 2019] establishes a framework for the meta-evaluation of
326 evaluators in dialogue response generation systems, specifically those that incorporate knowledge
327 elements. Our approach adheres to the methodology outlined in Zhong et al. [2022], employing
328 human ratings to assess dialogues on four attributes: coherence, engagingness, groundedness, and
329 naturalness.

330 **C Related Work**

331 **Evaluation using LLMs.** Fu et al. [2023] proposes GPTScore, a framework that evaluates texts
332 with generative pre-training models like GPT-3, assuming that a generative pre-training model
333 will assign a higher probability of high-quality generated text following a given instruction and
334 context. Wang et al. [2023] conducted a preliminary survey of using ChatGPT as an NLG evaluator.
335 Kocmi and Federmann [2023] proposed to use GPT models for evaluating machine translation
336 tasks. Hada et al. [2023] investigate whether LLM-based evaluators can help scale up multilingual
337 evaluation. Liu et al. [2023b] introduced G-Eval, a novel framework using large language models
338 through a chain-of-thoughts (CoT) approach combined with a form-filling methodology to evaluate
339 natural language generation outputs. Chiang and Lee [2023] subsequently demonstrated that G-
340 Eval’s implementation of an automated CoT does not consistently align with human evaluations.
341 Furthermore, they highlighted the limitations of restricting LLMs to solely numeric evaluations within
342 G-Eval, prompting our investigation into generating evaluative outputs that include both critical
343 commentary and numerical ratings. Jiang et al. [2023] introduced TIGERScore, an innovative metric
344 designed to offer detailed error analysis (in contrast to the scores) for identifying specific inaccuracies
345 within generated texts, moving beyond mere scoring. This metric is underpinned by the use of
346 Llama-2, which was fine-tuned on a proprietary dataset. Shu et al. [2024] introduced FusionEval, a
347 novel evaluation framework that leverages auxiliary evaluators such as NLI, BLEURT, and SBLEURT
348 to analyze questions for assessment. This analysis is then conveyed to a primary large language
349 model, which assigns the final score. Liu et al. [2023a] developed X-Eval, a two-stage instruction
350 tuning framework designed to evaluate texts across both familiar and novel dimensions, tailored to the
351 specific needs of end-users. Chan et al. [2023] proposed ChatEval, a multi-agent referee system that
352 employs a unique method of autonomous debating among the same agents with different personas
353 to evaluate the quality of generated responses. While their approach shares similarities with our
354 work, it primarily relies on a debate and discussion methodology utilizing the same models under
355 varied personas. In contrast, our method employs diverse models acting as peers and area chairs and
356 incorporates a significantly richer evaluation schema.

357 **Reasoning using Multiple LLMs as Peers.** Chen et al. [2024] unveiled ReConcile, a multi-model,
358 multi-agent framework structured akin to a round table conference among various LLM agents. Their
359 findings suggest that LLMs exhibit enhanced reasoning capabilities when engaging in discussions and
360 reaching consensus. Xu et al. [2023] introduced a novel framework aimed at augmenting reasoning
361 abilities, drawing inspiration from the academic peer review process. This approach uniquely
362 emphasizes iterative improvement through feedback from peer evaluations, distinguishing it from our
363 methodology, which does not facilitate direct communication between peers but instead involves an
364 area chair reviewing all peer responses. Pham et al. [2023] advocated for the use of embeddings as a
365 communication medium within multi-agent frameworks to optimize reasoning. Conversely, Du et al.
366 [2023] focuses on using solutions from other peers to enhance an individual’s reasoning, employing

367 a repetitive improvement cycle. Lastly, Wang et al. [2024] proposes a strategy for selecting the
368 most coherent response from multiple reasoning chains, offering a different perspective on achieving
369 consensus and enhancing reasoning accuracy.

370 **D Scoring Function**

371 In their seminal work, Liu et al. [2023b] broached the subject of a post-evaluation scoring function
372 designed to alleviate inherent biases and discrepancies within scoring mechanisms. However, the
373 intricacies and the practical application of this scoring function remained undisclosed, echoing
374 the reservations posited by Chiang and Lee [2023]. In our approach, we similarly refrain from
375 integrating an unspecified scoring function into our schema. This decision stems from the aspiration
376 to ensure clarity and reproducibility in our methodology. Despite this, the potential benefits of
377 incorporating a scoring function cannot be understated, particularly in addressing two significant
378 challenges highlighted by Liu et al. [2023b]: the propensity of scoring outcomes to gravitate towards
379 a dominant value—thereby exhibiting low variability and a diminished correlation with human
380 assessments—and the constraints of Large Language Models (LLMs) in generating only integer
381 values for scores, precluding fractional evaluations and consequently leading to a proliferation of ties
382 that mask the nuanced differences among Natural Language Generation (NLG) outputs.

383 To confront these challenges, both Liu et al. [2023b] and Chiang and Lee [2023] have explored the
384 utilization of the “n” parameter in LLMs, notably OpenAI’s GPT-3.5. This parameter, which dictates
385 the quantity of generated outputs per given prompt, serves as a cornerstone in their strategy to yield
386 decimal scores. By calculating the average of these multiple outputs, they endeavored to engender
387 a scoring system characterized by enhanced variance and distribution more closely aligned with
388 human evaluative patterns. Although Liu et al. [2023b] alluded to the employment of log probabilities
389 within their scoring function, their implementation primarily leveraged a straightforward averaging
390 mechanism. This discrepancy is presumed to arise from the unavailability of a log probabilities
391 functionality in versions of GPT-3.5-turbo and subsequent iterations. In light of this limitation, our
392 framework adopts a simplistic averaging approach whenever the “n” value exceeds unity, thereby
393 ensuring consistency and uniformity in our evaluative processes.

394 Furthermore, the application of log probability within the scoring function emerges as a feasible
395 approach solely under the condition that the evaluation conforms to the methodology outlined in [Liu
396 et al., 2023b], focusing exclusively on the generation of scores. This technique is predicated on the
397 calculation of probabilities associated with the generation of specific outputs, offering a nuanced
398 metric for evaluation. However, this method’s relevance diminishes when the evaluative process
399 extends beyond mere scoring to encompass reasoning or the generation of critical commentary prior to
400 the assignment of a score. In such contexts, where evaluative narratives or qualitative feedback precede
401 quantitative scoring, the direct application of log probabilities becomes less pertinent. The essence of
402 incorporating critical commentary or explanatory feedback is to shed light on the rationale behind
403 the score, thus providing a comprehensive understanding of the evaluated output’s strengths and
404 weaknesses. In these scenarios, the scoring mechanism necessitates a more adaptable and interpretive
405 approach, one that transcends the straightforward application of mathematical probabilities and
406 ventures into the realm of qualitative assessment. Consequently, while log probabilities offer a
407 rigorous and mathematically grounded method for score calculation in certain instances, their utility
408 is contextually bound and may not align with evaluative frameworks that prioritize explanatory or
409 critical analysis alongside numerical scoring.

410 **E Hyperparameters**

411 Regarding the selection of hyperparameters for LLM agents, we adhered to default settings with
412 exceptions for ‘n’ and ‘temperature’. Echoing findings from [Chiang and Lee, 2023], we set the
413 temperature to 1 across all tasks to optimize NLG task evaluations. The ‘n’ parameter, dictating the
414 number of responses generated per prompt, played a crucial role in our methodology. Following the
415 precedent set by Liu et al. [2023b], who utilized $n = 20$ to average out scores from multiple responses,
416 we explored the impact of varying ‘n’ on evaluation outcomes. Preliminary experiments demonstrate
417 the influence of higher ‘n’ values on achieving more representative scores. All the experiments are
418 conducted on a A100 (80GB) GPU server.

419 **F Is this a General Purpose Framework?**

420 The ReFeR framework shows its efficacy for NLG assessment, utilizing Large Language Models
 421 (LLMs) as its cornerstone evaluative agents. The framework encourages a paradigm shift towards a
 422 more nuanced examination of NLG outputs, fostering a structured approach that emphasizes review,
 423 feedback, and reasoning processes. But the framework is by default modality independent, and we
 424 can extend the ReFeR framework’s applicability beyond its textual confines, aiming to encompass a
 425 broader spectrum of data modalities by using the capabilities of Multi-Modal Foundation Models
 426 (FMs) in these domains [Li et al., 2023]. Also, with the increased use of external knowledge and
 427 tool-usage [Schick et al., 2023] [Patil et al., 2023], in conjunction with LLMs, we can use these
 428 for improved peer evaluation and further feedback and reasoning. And these things can be added
 429 modularly without any change in the framework, just like we experiment with different peers.

430 The review, feedback, and reasoning modules make the ReFeR framework useful even beyond the
 431 evaluation of NLG content, making it an effective generator of instruction-tuning data for fine-tuning
 432 smaller models and an effective reasoning module for complex tasks.

433 Due to constraints of computing and time, we could not verify these results using multimodal LLMs
 434 and external tools. This remains a promising direction to extend our work. This adaptability and
 435 easy extendability not only broadens the framework’s applicability across diverse AI outputs but also
 436 highlights its evolutionary potential alongside technological progressions in the field of generative
 437 models. Thus, the ReFeR framework stands as a testament to the ongoing evolution in the evaluation
 438 of multimedia content, offering nuanced and multidimensional assessments that reflect the complexity
 439 and diversity of modern AI-generated outputs.

440 **G Ablations**

441 **Prompt Ablations**

442 To identify the most effective prompt for a task, we designed prompts aimed at achieving the highest
 443 correlation. We employed the same three models used in our main experiments (GPT-3.5 Turbo,
 444 Gemini-Pro, Mixtral-8x7B) and conducted ablation studies on the SummEval dataset by varying
 445 the prompts. We utilized two different prompt schemas for this experiment: Analyze-Rate and Eval
 446 Guidelines. Table 5 presents the average Spearman score for all three models, broken down by
 447 metric and averaged over two runs. It is evident that Analyze-Rate and Eval Guidelines perform
 448 very similarly, with Analyze-Rate being marginally better by 0.0001. When observing the average
 449 ρ of all three peers, we see that the Eval Guidelines prompt works better for GPT-3.5 and Mixtral,
 450 while the Analyze-Rate prompt is better for Gemini. This leads to the important observation that the
 451 best-performing prompt for one model may not be the best for another.

452 The results in Table 5 might suggest that using the Eval Guidelines prompt for both the peers and the
 453 Area Chair would yield the best results. However, to verify this, we conducted further experiments by
 454 permuting the Analyze-Rate and Eval Guidelines prompts for peers and the Area Chair.

Table 5: Prompt Ablation of Peers

Prompt	SummEval	Coherence	Consistency	Fluency	Relevance	Average
Analyze Rate	GPT-3.5	0.337	0.333	0.270	0.298	0.309
	Mixtral	0.291	0.383	0.365	0.237	0.319
	Gemini	0.362	0.323	0.220	0.228	0.283
	Average Peers	0.330	0.346	0.285	0.255	0.3039
Eval Guidelines	GPT-3.5	0.452	0.278	0.328	0.336	0.348
	Mixtral	0.308	0.311	0.350	0.366	0.334
	Gemini	0.241	0.258	0.201	0.217	0.229
	Average Peers	0.334	0.286	0.293	0.306	0.3038

Table 6: Prompt Ablation by Varying Both Peer Prompt and Area Chair Prompt on SummEval

Peer Prompt	AC Prompt	Coh	Con	Flu	Rel	Avg
Analyze Rate	Analyze Rate	0.463	0.404	0.380	0.535	0.445
Analyze Rate	Eval Guidelines	0.502	0.428	0.414	0.459	0.450
Eval Guidelines	Analyze Rate	0.459	0.403	0.377	0.474	0.428
Eval Guidelines	Eval Guidelines	0.480	0.392	0.350	0.463	0.421

455 Table 6 displays the outcomes when the prompts were permuted between the peers and the Area
 456 Chair. We found that the optimal combination was not using Analyze-Rate for both the peers and the
 457 Area Chair, but rather using Analyze-Rate for the peers and Eval Guidelines for the Area Chair on
 458 the SummEval dataset. Although the difference in average correlation across all four metrics is only
 459 0.005, a closer examination of the metric-wise differences reveals that Row 2 outperforms Row 1 in
 460 all metrics except for the Relevance metric. This indicates that refining the Eval Guidelines prompt
 461 for the Relevance metric could yield even better results. As previously mentioned, identifying the
 462 optimal prompt is always challenging. Therefore, we did not further explore improvements to the
 463 Eval Guidelines prompt, as this is not the primary focus of our paper.

464 Our current conclusion is that prompt modification can enhance scores, and we leave the task of
 465 finding the best method for determining the optimal prompt to future research. Another important
 466 observation is that just because a prompt works best for a model does not guarantee that using the
 467 same prompt in a framework setting would yield better results. This suggests that users might need
 468 to perform experiments on their downstream applications to determine the best working prompt for
 469 their use case.

470 Performance Ablations

471 To understand the overall percentage gains from different parts of the model, we conducted a
 472 performance ablation experiment where we added each component of the framework incrementally
 473 and observed the improvement in overall performance. Table 7 shows the results of this experiment.
 474 We first start with the base model, i.e., Single Peer (GPT-3.5 Turbo with $n = 1$). We check the
 475 average Spearman correlation across the four metrics of SummEval (coherence, consistency, fluency,
 476 relevance) for all the models listed and then calculate their relative percentage gain with respect to
 477 the base model (Row 1). For the second row, we increased the hyperparameter $n = 20$ and observed
 478 a gain of +21.29%, highlighting the importance of this hyperparameter. We suspect, this is the reason
 479 for the performance behind G-Eval and Analyze-Rate. Then we use all three peers individually and
 480 take the average of the peers, resulting in a decline in performance (1.6%) compared to the base
 481 model.

Table 7: Performance Ablation on SummEval dataset; AR stands for Analyze-Rate prompt, EG stands for Eval Guidelines prompt, AC stands for Area chair. n is the hyperparameter that tells the model how many responses to give for each prompt. Average ρ is the average spearman correlation across the 4 metrics for SummEval dataset. % gain is relative to 1st row.

Models	Avg ρ	% Gain	Reason
GPT-3.5 ($n=1$)	0.309	-	-
GPT-3.5 ($n=20$)	0.375	+21.29%	$n=20$
Average of Peers	0.304	-1.603%	3 Peers (No AC)
AR (Peers + AC)	0.365	+18.05%	3 Peers + AC
AR (Peers + AC ($n=20$))	0.445	+44.20%	$n=20$ for AC
AR (Peers) + EG (AC)	0.450	+45.84%	Full Framework

482 Next, we add the Area Chair into the framework with ($n = 1$, Analyze-Rate prompt for both Peers
 483 and Area Chair), and we see a percentage gain of +18.05% relative to the base model. This shows
 484 the importance of the Area Chair in helping the model reconcile all the evaluations and provide a
 485 better overall evaluation. We then further conduct ablation by checking with $n = 20$ for the Area
 486 Chair, which improves the gain to +44.2% relative to the base model. From Table 6, we see that Row
 487 2 performs best, and when used in our ablation, it helps us decide what prompts should be finally
 488 used in the current framework to achieve maximum gain. This ablation is represented in the last row
 489 of Table 7, which shows the overall percentage gain our framework brings.

490 H Critic Communication Strategies and Discussion

491 H.1 Our Proposed Communication Strategies

492 The Critic module serves as a crucial second step in our ReFeR framework. This module operates
 493 by submitting the initial evaluations—comprising both scores and commentary generated by peer
 494 agents—to a distinct critic LLM agent. This agent then undertakes the critical decision-making
 495 process regarding the necessity of re-evaluating the task at hand. Significantly, the critic LLM’s
 496 feedback is designed to enable peer agents to refine and enhance their evaluations. And in some
 497 strategies, the critic LLM is used to give feedback on the peer reviews to the Areachair. Consequently,
 498 this process gives rise to a fundamental question: How can effective communication between peer
 499 LLM agents and the critic LLM agent be established?

500 Hence, we propose multiple communication strategies as follows:

501 **(1) Individual Peer Evaluation:** In this strategy, the peer evaluations are given to the critic module
 502 separately. The critic agent is tasked with assessing these responses individually, determining the
 503 need for re-evaluation, and suggesting enhancements to bolster the evaluation process. Using this
 504 feedback from the critic, the peer LLMs do a re-evaluation of the sample.

Table 8: Spearman Correlation (ρ) and Kendall-Tau Correlation (τ) on SummEval. The best and second-best per column have been highlighted in bold and underlined, respectively.

	Model	Coherence		Consistency		Fluency		Relevance		Average	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ReFeR (Ours)	No Critic Phase	0.502	0.374	0.428	0.370	0.414	0.330	0.459	0.338	0.450	0.353
	Individual Peer Eval	0.500	0.381	0.331	0.305	<u>0.393</u>	0.309	0.433	0.306	0.414	0.326
	Collective Peer Eval	<u>0.505</u>	0.377	<u>0.381</u>	<u>0.321</u>	0.390	0.311	0.419	0.307	<u>0.424</u>	0.329
	Weighted Feedback	0.510	0.376	0.346	0.288	0.375	0.300	<u>0.435</u>	0.328	0.417	0.323
	Critic Comment Feedback	0.425	0.309	0.332	0.270	0.294	0.233	0.328	0.248	0.345	0.265

Table 9: Spearman Correlation (ρ) and Kendall-Tau Correlation (τ) on TopicalChat. Best and second-best per column have been highlighted with bold and underline respectively.

	Model	Coherence		Engagingness		Groundedness		Naturalness		Average	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ReFeR (Ours)	No Critic Phase	0.514	0.390	0.651	0.502	0.678	0.590	0.544	0.414	0.597	0.474
	Individual Peer Eval	0.527	0.400	0.650	0.499	0.628	0.546	0.556	0.426	0.590	0.468
	Collective Peer Eval	0.527	0.399	0.643	0.495	0.659	0.571	0.536	0.414	0.591	0.470
	Weighted Feedback	0.535	0.403	0.635	0.483	0.638	0.553	0.546	0.417	0.588	0.464
	Critic Comment Feedback	0.481	0.369	0.507	0.383	0.526	0.454	0.437	0.326	0.488	0.383

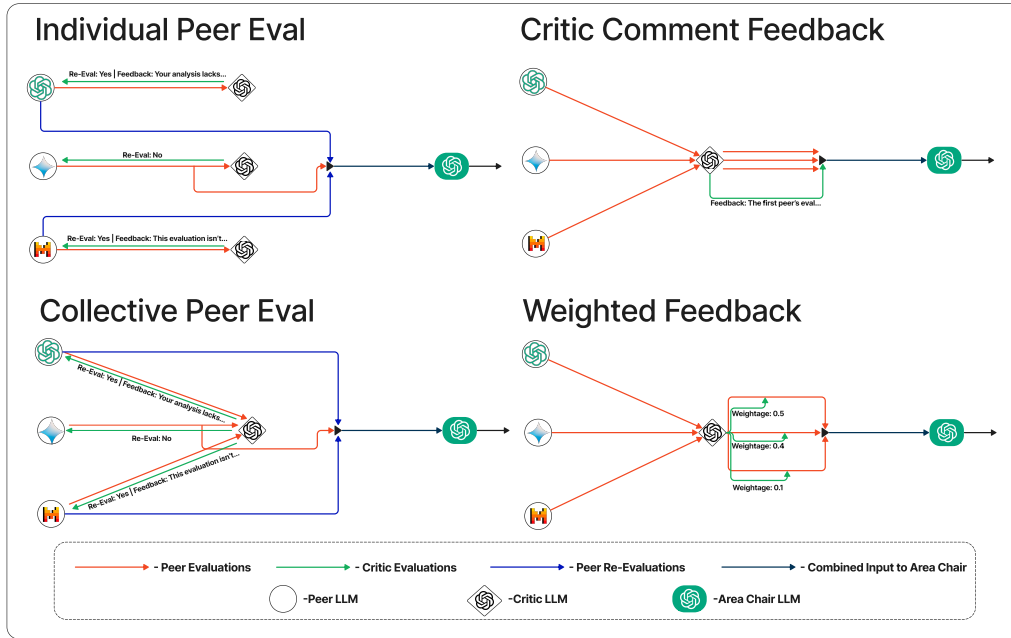


Figure 3: Critic Communication Strategies

505 **(2) Collective Peer Evaluation:** This strategy explores the efficacy of submitting all peer evaluations
 506 to the critic simultaneously. The goal was to ascertain how such an aggregate submission impacts the
 507 critique process. Feedback from the critic was structured in a JSON format, ensuring clear delineation
 508 of comments for each peer and facilitating targeted improvements. This strategy significantly reduces
 509 the number of calls made to the critic. This strategy is also a re-evaluation strategy similar to the
 510 previous strategy.

511 **(3) Weighted Feedback:** Here, the critic is requested to assign weights to each peer’s evaluation.
 512 These weighted assessments are intended for subsequent review by an area chair, offering a nuanced
 513 perspective on the relative merits of each evaluation. In this strategy, the weights are directly relayed
 514 to the Area Chair as feedback along with the peer evaluations.

515 **(4) Critic Feedback to AreaChair:** This strategy also involves the direct transmission of the critic’s
 516 commentary on the peer evaluations to the area chairs with no re-evaluation. This approach prioritizes
 517 the provision of expert insight at the decision-making apex of the evaluative hierarchy.

518 H.2 Discussion about the Communication Strategies

519 Table 8 compares different models based on coherence, consistency, fluency, and relevance on
520 SummEval. We make the following observations: (1) ReFeR with No Critic provides the best ρ and
521 τ . (2) In terms of ρ and τ , collective peer evaluation strategy is the second best. Individual peer
522 evaluation is worse than collective strategy, even though it incurs thrice the cost of its collective
523 counterpart and the critic gives feedback to the peers individually. (3) Strategies like the Critic
524 Comment Feedback Strategy (where the critic gives comments/feedback on the peer evaluations
525 to Area Chair) involve giving the model too much information through prompt which reduce its
526 performance significantly.

527 Table 9 compares different models based on various aspects of conversation quality like coherence,
528 engagingness, groundedness, and naturalness, and average scores on TopicalChat. We make the
529 following observations: (1) ReFeR with No Critic provides the best ρ and τ . (2) In terms of ρ and τ ,
530 ReFeR with No critic performs well for engagingness and groundedness, but ReFeR with Weighted
531 Feedback strategy is the best for coherence and Individual Peer Eval is best for Naturalness. (3)
532 ReFeR with Direct Feedback is the least performing ReFeR model since it obtains the lowest ρ , τ
533 values among the ReFeR models.

Article: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team's 0-0 draw with Burnley on Sunday...

Summary: Paul Merson has restarted his row with Andros Townsend. The Tottenham midfielder was brought on with only seven minutes remaining in his team's 0-0 draw with Burnley. Andros Townsend scores England's equaliser in their 1-1 friendly draw with Italy in Turin.

Human Annotated Scores:

- **Coherence:** 2.33

Peer 1's Evaluation:

Analysis: The summary includes the main topic and key points of the news article, but the sentences are not well-structured or organized. It seems to be a combination of different sentences from the source text without a clear flow or logical order.

Rating: 2

Peer 2's Evaluation:

Analysis: The summary lacks a clear and logical order in presenting the information, and the connection between the two events (Townsend's substitution and his goal for England) is not explicitly stated.

Rating: 3.0.

Peer 3's Evaluation:

Analysis: The summary lacks proper structure and does not provide clear transitions between sentences. The sentences appear to be a collection of disjointed facts without a coherent narrative.

Rating: 1

AC's Evaluation:

[**Analysis:** The summary accurately covers some key points from the article, including the restart of the row between Paul Merson and Andros Townsend and Townsend's goal for England against Italy. However, it fails to mention the specific incident with Townsend being brought on late in the game against Burnley. The summary lacks a clear structure and logical flow, as it jumps from one point to another without a cohesive progression. **Rating:** 2.5, ... (19 more responses like this)]

Final rating: 2.35

535 **J Case Study for Instruction Tuning Experiments**

Article: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team’s 0-0 draw with Burnley on Sunday...

Summary: Paul merson was brought on with only seven minutes remaining in his team ’s 0-0 draw with burnley . Andros townsend scored the tottenham midfielder in the 89th minute . Paul merson had another dig at andros townsend after his appearance . The midfielder had been brought on to the england squad last week . Click here for all the latest arsenal news news .,

Human Annotated Scores:

- **Coherence:** 1.33

Mistral-7B’s Evaluation (Non Finetuned):

Analysis: The given coherence example is about a summary of a news article. The summary is meant to be a concise summary that covers the main points of the news article. If each sentence of the summary is relevant to the main topic and key points mentioned in the news article, then the summary can have a high coherence rating. **Rating:** 4

Mistral-7B’s Evaluation (Finetuned):

Analysis: The summary seems to lack coherence as it is disjointed and does not link one idea into the next effectively. The central idea about Merson criticizing Townsend and Townsend scoring for England get mixed up in the middle, lacking a logical flow of information. Furthermore, the sentence ‘The midfielder had been brought on to the England squad last week’ seems out of place and does not correlate well with the rest of the content, hence, compromising overall coherence. **Rating:** 1.4

536 **K Example of Reasoning improved by Framework**

537 Example from GSM8k:

QUESTION: Solo has to read 4 pages from his Science textbook, 20 pages from his Social Studies textbook, 7 pages from his History textbook and 8 pages from his Geography textbook. Solo read 15 pages on Monday. If he has 4 more days to complete all his reading, how many pages does he need to read, on average, in one day?

GOLD ANSWER: 6 pages

PEER 1'S RESPONSE: To find the total number of pages Solo needs to read, we add up the pages from each textbook: 4 (Science) + 20 (Social Studies) + 7 (History) + 8 (Geography) = 39 pages. Solo needs to read, on average, 9.75 pages per day.

PEER 2'S RESPONSE: Total pages to read: 4 (Science) + 20 (Social Studies) + 7 (History) + 8 (Geography) = 40 pages. After subtracting 15 pages already read, Solo needs to read 25 more pages. He needs to read an average of 6.25 pages per day.

PEER 3'S RESPONSE: Total pages to read: 4 (Science) + 20 (Social Studies) + 7 (History) + 8 (Geography) = 39 pages. Solo needs to read, on average, 9.75 pages per day.

AREA CHAIR'S RESPONSE: Total pages to read: 4 + 20 + 7 + 8 = 39 pages. After subtracting 15 pages already read, Solo needs to read 24 more pages. He needs to read an average of 6 pages per day. Peer 1 and Peer 3 provided the correct calculations but made a mistake in their final answers by not subtracting the number of pages Solo read on Monday from the total number of pages. They both stated that Solo needs to read, on average, 9.75 pages per day, which is incorrect. Peer 2 made a calculation mistake. My solution corrects these errors and provides the accurate calculation for the average number of pages Solo needs to read per day.

538

539 L Sample Prompts for SummEval dataset

Peer Prompt (Analyze Rate):

You will be given one summary written for a news article.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

540

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

Evaluation Form (Answer by starting with "Analysis:" to analyze the given example regarding the evaluation criteria as concise as possible, and then give the numeric rating on the next line by "Rating:):

- Coherence:

541

Peer Prompt (Eval Guidelines) :

You will be given one summary written for a news article.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

4. Scoring Guidelines:

Score = 5: The summary fully captures all key points of the article with an accurate and logical flow, without any significant omissions or irrelevant information.

4 ≤ Score < 5: Most key points are included with a generally logical sequence, albeit with minor omissions or slight inclusions of less relevant information.

3 ≤ Score < 4: Some key points are present, but others are missing, and the flow has noticeable gaps or jumps, including some irrelevant details.

2 ≤ Score < 3: Several key points are missed, and the flow is disjointed with significant omissions or inaccuracies, and noticeable irrelevant content.

1 ≤ Score < 2: Fails to represent the article accurately, lacks coherence and logical flow, with major elements missing or misrepresented, and significant irrelevant details.

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

Evaluation Form (Answer by starting with "Analysis:" to analyze the given example regarding the evaluation criteria as concise as possible, and then give the numeric rating on the next line by "Rating:):

- Coherence:

542

Peer Prompt (Re-Evaluation):

You will be given one summary written for a news article.

Your task is to re-evaluate the summary based on your previous evaluation, which will also be provided. Please consider the Critic Comment on your initial evaluation when re-evaluating.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

Initial Evaluation: {{initial_eval}}

Critic Comment: {{Critic Response}}

Re-Evaluation Instructions:

18

1. Analysis: Based on the critic's comment provided, re-evaluate the summary for coherence.
2. Rating: Provide a numeric rating for coherence based on your revised evaluation.

543

544

Critic Prompt: *(For Individual Peer Eval Strategy)*

You will be provided with a news article summary and the initial evaluation from a large language model (LLM), referred to as the assistant's evaluation. The assistant's evaluation includes a brief analysis by the assistant and a rating given by the assistant.

Your task is to correct one aspect of the assistant's evaluation based on a specific metric and provide feedback to the LLM in the form of a critic comment. Additionally, you need to determine whether there is a need for re-evaluation.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) - *The collective quality of all sentences. The summary should be well-structured and well-organized, not just a heap of related information, but building from sentence to a coherent body of information about a topic.*

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

Assistant's Evaluation: {{Peer Response}}

Evaluation Form:

-Critic Comment: Provide concise feedback to the assistant regarding the evaluation.

-Re-Evaluation: Yes/No, based on whether you believe there is a need for re-evaluation.

545

Critic Prompt: *(For Collective Peer Eval Strategy)*

546

You will be provided with a news article summary and the initial evaluation from three large language models (LLMs), referred to as the assistant's evaluation. The assistant's evaluation includes a brief analysis by the assistant and a rating given by the assistant.

Your task is to correct one aspect of each assistant's evaluation based on a specific metric and provide feedback to the LLM in the form of a critic comment. Additionally, you need to determine whether there is a need for re-evaluation for each assistant.

Please carefully review and understand these instructions. Keep this document open for reference while reviewing.

Evaluation Criteria:

Coherence (1-5) - The collective quality of all sentences. The summary should be well-structured and well-organized, not just a heap of related information, but building from sentence to a coherent body of information about a topic.

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

First Assistant's Evaluation: {{Peer Response}}

Second Assistant's Evaluation: {{Peer Response2}}

Third Assistant's Evaluation: {{Peer Response3}}

Evaluation Form:

-Critic Comment: Provide concise feedback to the assistant regarding the evaluation.

-Re-Evaluation: Yes/No, based on whether you believe there is a need for re-evaluation.

Provide your feedback for each assistant in the following format:

```
{
  "evaluators":
  {
    "evaluator": "Assistant 1",
    "critic_comment": "Your feedback for Assistant 1's evaluation ",
    "re_evaluation": "Yes/No"
  },
  {
    "evaluator": "Assistant 2",
    "critic_comment": "Your feedback for Assistant 2's evaluation ",
    "re_evaluation": "Yes/No"
  },
  {
    "evaluator": "Assistant 3",
    "critic_comment": "Your feedback for Assistant 3's evaluation ",
    "re_evaluation": "Yes/No"
  }
}
```

Please provide the critic comments and re-evaluation decisions for each assistant model as requested.

Critic Prompt: *(For Weighted Feedback Strategy)*

You will be provided with a news article summary and the initial evaluation from three large language models (LLMs), referred to as the assistant's evaluation. The assistant's evaluation includes a brief analysis by the assistant and a rating given by the assistant.

Your task is to correct one aspect of each assistant's evaluation based on a specific metric and provide feedback to the LLM in the form of a critic comment. Additionally, you need to provide a weightage for the assistant's evaluation.

Please carefully review and understand these instructions. Keep this document open for reference while reviewing.

Evaluation Criteria:

Coherence (1-5) - *The collective quality of all sentences. The summary should be well-structured and well-organized, not just a heap of related information, but building from sentence to a coherent body of information about a topic.*

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

First Assistant's Evaluation: {{Peer Response}}

Second Assistant's Evaluation: {{Peer Response2}}

Third Assistant's Evaluation: {{Peer Response3}}

Evaluation Form:

-Critic Comment: Provide concise feedback to the assistant regarding the evaluation.

-Weightage: Provide a weightage for the assistant's evaluation, indicating the quality of the evaluation. Use a scale of 0-1, where 0 is the lowest and 1 is the highest.

Provide your feedback for each assistant in the following format:

```
{
  "evaluators":
  {
    "evaluator": "Assistant 1",
    "critic_comment": "Your feedback for Assistant 1's evaluation",
    "weightage": "Weightage value (0-1)"
  },
  {
    "evaluator": "Assistant 2",
    "critic_comment": "Your feedback for Assistant 2's evaluation",
    "weightage": "Weightage value (0-1)"
  },
  {
    "evaluator": "Assistant 3",
    "critic_comment": "Your feedback for Assistant 3's evaluation",
    "weightage": "Weightage value (0-1)"
  }
}
```

Please provide the critic comments and weightage for each assistant model as requested.

Area Chair Prompt: *(For No Critic, Individual and Collective Strategies)*

You will be given one summary written for a news article and you will also receive initial evaluations from three large language models, referred to as the assistants' evaluations.

Your task is to rate the summary on one metric.

Please read the instructions and criteria below carefully and use them as a guide in your evaluation.

Evaluation Criteria:

Coherence (1-5) - Assess the structural and organizational quality of the summary. It should present information logically and clearly, relating to the main topic of the news article. Consider if the summary is well-structured, if it progresses logically from point to point, and if it effectively encapsulates the key points of the article.

Evaluation Guidelines:

1. Read the news article to understand the main topic and key points.
2. Review the summary. Analyze if it accurately and logically covers the main points of the article.
3. Rate the summary's coherence on a scale of 1 to 5, with 1 being the least coherent and 5 being the most coherent.

4. Scoring Guidelines:

Score = 5: The summary fully captures all key points of the article with an accurate and logical flow, without any significant omissions or irrelevant information.

4 ≤ Score < 5: Most key points are included with a generally logical sequence, albeit with minor omissions or slight inclusions of less relevant information.

3 ≤ Score < 4: Some key points are present, but others are missing, and the flow has noticeable gaps or jumps, including some irrelevant details.

2 ≤ Score < 3: Several key points are missed, and the flow is disjointed with significant omissions or inaccuracies, and noticeable irrelevant content.

1 ≤ Score < 2: Fails to represent the article accurately, lacks coherence and logical flow, with major elements missing or misrepresented, and significant irrelevant details.

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

First Assistant's Evaluation: {{Peer_response1}}

Second Assistant's Evaluation: {{Peer_response2}}

Third Assistant's Evaluation: {{Peer_response3}}

Evaluation Form (Please provide your analysis and rating as follows):

- Analysis: [Your detailed analysis here, focusing on the structural and logical flow of the summary in relation to the source text.] - Rating: [Your coherence rating here on a scale from 1 to 5.]

Area Chair Prompt: *(For Critic Comment Feedback Strategy)*

You will be given one summary written for a news article and you will also receive initial evaluations from three large language models, referred to as the assistants' evaluations. You will also receive critic comments by another LLM for each of these assistant's evaluations. Consider these in your evaluation.

Your task is to rate the summary on one metric.

Please read the instructions and criteria below carefully and use them as a guide in your evaluation.

(Evaluation Criteria and Evaluation Guidelines same as above)

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

First Assistant's Evaluation: {{Peer_response1}} Critic Comment : {{Critic_Comment1}}

Second Assistant's Evaluation: {{Peer_response2}} Critic Comment : {{Critic_Comment2}}

Third Assistant's Evaluation: {{Peer_response3}} Critic Comment : {{Critic_Comment3}}

Evaluation Form (Please provide your analysis and rating as follows):

- Analysis: [Your detailed analysis here, focusing on the structural and logical flow of the summary in relation to the source text.] - Rating: [Your coherence rating here on a scale from 1 to 5.]

548

549

550

Area Chair Prompt: *(For Weighted Feedback Strategy)*

You will be given one summary written for a news article and you will also receive initial evaluations from three large language models, referred to as the assistants' evaluations. You will also receive weightage's for evaluations by another critic LLM for each of these assistant's evaluations. Consider these in your evaluation.

Your task is to rate the summary on one metric.

Please read the instructions and criteria below carefully and use them as a guide in your evaluation. (Evaluation Criteria and Evaluation Guidelines same as above)

551

Example:

Source Text: {{Full Article}}

Summary: {{Summary of Article}}

First Assistant's Evaluation: {{Peer_response1}} Weightage : {{weightage1}}

Second Assistant's Evaluation: {{Peer_response2}} Weightage : {{weightage2}}

Third Assistant's Evaluation: {{Peer_response3}} Weightage : {{weightage3}}

Evaluation Form (Please provide your analysis and rating as follows):

- Analysis: [Your detailed analysis here, focusing on the structural and logical flow of the summary in relation to the source text.] - Rating: [Your coherence rating here on a scale from 1 to 5.]

552

M Sample Prompts for GSM8k Dataset

Peer Prompt:

You will be provided with a problem that requires logical reasoning, mathematical calculation, or both. Your task is to solve the problem accurately, providing not just the answer but also a clear explanation of the steps taken to reach that answer. It is crucial to understand the problem thoroughly and apply the correct principles or formulas to solve it.

Instructions:

553

- 1. Read the problem statement carefully, ensuring you understand all the details and what is required for the solution.*
- 2. Work through the problem logically and methodically, explaining your reasoning and the steps you take to solve the problem.*
- 3. Provide the final answer clearly, specifying it as a numerical value or a specific explanation as required by the problem statement.*

Problem Statement:

{{Problem}}

Evaluation Form:

- Analysis: Start with "Analysis:" to provide a concise and structured explanation of the steps and reasoning used to solve the problem. Ensure your analysis is clear and follows a logical sequence. - Final Answer: Clearly state the final answer only (number) on the line after your analysis.

Area Chair Prompt:

You will be provided with a problem that requires logical reasoning, mathematical calculation, or both. Alongside the problem, you will also see solutions provided by three other Language Models (LLMs). Your task is to analyze these solutions, identify any inaccuracies or inconsistencies, and then solve the problem accurately yourself. This requires not just providing the correct answer but also a clear explanation of the steps taken to reach that answer, especially in light of the previous solutions provided. It is crucial to understand the problem thoroughly and apply the correct principles or formulas to solve it correctly.

Instrucions:

1. *Read the problem statement carefully, ensuring you understand all the details and what is required for the solution.*
2. *Review the solutions provided by the three other LLMs. Note any common mistakes or discrepancies in their reasoning or calculations.*
3. *Work through the problem logically and methodically, explaining your reasoning and the steps you take to solve the problem.*
4. *Provide the final answer clearly, specifying it as a numerical value or a specific explanation as required by the problem statement.*

Problem Statement:

{{Problem}}

Solution by other LLMs:

LLM1 Answer: {{Peer_response1}}

LLM2 Answer: {{Peer_response2}}

LLM3 Answer: {{Peer_response3}}

Evaluation Form:

- Analysis: Start with "Analysis:" to provide a concise and structured explanation of the steps and reasoning used to solve the problem. Ensure your analysis is clear and follows a logical sequence. -

- Final Answer: Clearly state the final answer only (number) on the line after your analysis.

554

555 **NeurIPS Paper Checklist**

556 The checklist is designed to encourage best practices for responsible machine learning research,
557 addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove
558 the checklist: **The papers not including the checklist will be desk rejected.** The checklist should
559 follow the references and follow the (optional) supplemental material. The checklist does NOT count
560 towards the page limit.

561 Please read the checklist guidelines carefully for information on how to answer these questions. For
562 each question in the checklist:

- 563 • You should answer [Yes] , [No] , or [NA] .
- 564 • [NA] means either that the question is Not Applicable for that particular paper or the
565 relevant information is Not Available.
- 566 • Please provide a short (1–2 sentence) justification right after your answer (even for NA).

567 **The checklist answers are an integral part of your paper submission.** They are visible to the
568 reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it
569 (after eventual revisions) with the final version of your paper, and its final version will be published
570 with the paper.

571 The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation.
572 While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a
573 proper justification is given (e.g., "error bars are not reported because it would be too computationally
574 expensive" or "we were unable to find the license for the dataset we used"). In general, answering
575 "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we
576 acknowledge that the true answer is often more nuanced, so please just use your best judgment and
577 write a justification to elaborate. All supporting evidence can appear either in the main paper or the
578 supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification
579 please point to the section(s) where related material for the question can be found.

580 IMPORTANT, please:

- 581 • **Delete this instruction block, but keep the section heading “NeurIPS paper checklist”,**
- 582 • **Keep the checklist subsection headings, questions/answers and guidelines below.**
- 583 • **Do not modify the questions and only use the provided macros for your answers.**

584 1. **Claims**

585 Question: Do the main claims made in the abstract and introduction accurately reflect the
586 paper’s contributions and scope?

587 Answer: "[Yes]"

588 Justification: Our abstract and introduction only talks about the contribution of our paper in
589 brief.

590 Guidelines:

- 591 • The answer NA means that the abstract and introduction do not include the claims
592 made in the paper.
- 593 • The abstract and/or introduction should clearly state the claims made, including the
594 contributions made in the paper and important assumptions and limitations. A No or
595 NA answer to this question will not be perceived well by the reviewers.
- 596 • The claims made should match theoretical and experimental results, and reflect how
597 much the results can be expected to generalize to other settings.
- 598 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
599 are not attained by the paper.

600 2. **Limitations**

601 Question: Does the paper discuss the limitations of the work performed by the authors?

602 Answer: [Yes]

603 Justification: We clearly define the current limitations and potential future works to solve
604 them.

605 Guidelines:

- 606 • The answer NA means that the paper has no limitation while the answer No means that
607 the paper has limitations, but those are not discussed in the paper.
- 608 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 609 • The paper should point out any strong assumptions and how robust the results are to
610 violations of these assumptions (e.g., independence assumptions, noiseless settings,
611 model well-specification, asymptotic approximations only holding locally). The authors
612 should reflect on how these assumptions might be violated in practice and what the
613 implications would be.
- 614 • The authors should reflect on the scope of the claims made, e.g., if the approach was
615 only tested on a few datasets or with a few runs. In general, empirical results often
616 depend on implicit assumptions, which should be articulated.
- 617 • The authors should reflect on the factors that influence the performance of the approach.
618 For example, a facial recognition algorithm may perform poorly when image resolution
619 is low or images are taken in low lighting. Or a speech-to-text system might not be
620 used reliably to provide closed captions for online lectures because it fails to handle
621 technical jargon.
- 622 • The authors should discuss the computational efficiency of the proposed algorithms
623 and how they scale with dataset size.
- 624 • If applicable, the authors should discuss possible limitations of their approach to
625 address problems of privacy and fairness.
- 626 • While the authors might fear that complete honesty about limitations might be used by
627 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
628 limitations that aren't acknowledged in the paper. The authors should use their best
629 judgment and recognize that individual actions in favor of transparency play an impor-
630 tant role in developing norms that preserve the integrity of the community. Reviewers
631 will be specifically instructed to not penalize honesty concerning limitations.

632 3. Theory Assumptions and Proofs

633 Question: For each theoretical result, does the paper provide the full set of assumptions and
634 a complete (and correct) proof?

635 Answer: [NA]

636 Justification: The paper does not include theoretical results.

637 Guidelines:

- 638 • The answer NA means that the paper does not include theoretical results.
- 639 • All the theorems, formulas, and proofs in the paper should be numbered and cross-
640 referenced.
- 641 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 642 • The proofs can either appear in the main paper or the supplemental material, but if
643 they appear in the supplemental material, the authors are encouraged to provide a short
644 proof sketch to provide intuition.
- 645 • Inversely, any informal proof provided in the core of the paper should be complemented
646 by formal proofs provided in appendix or supplemental material.
- 647 • Theorems and Lemmas that the proof relies upon should be properly referenced.

648 4. Experimental Result Reproducibility

649 Question: Does the paper fully disclose all the information needed to reproduce the main ex-
650 perimental results of the paper to the extent that it affects the main claims and/or conclusions
651 of the paper (regardless of whether the code and data are provided or not)?

652 Answer: [Yes]

653 Justification: Yes we present a straight forward plug and play framework with all the details
654 to reproduce the results in the paper.

655 Guidelines:

- 656 • The answer NA means that the paper does not include experiments.
- 657 • If the paper includes experiments, a No answer to this question will not be perceived
- 658 well by the reviewers: Making the paper reproducible is important, regardless of
- 659 whether the code and data are provided or not.
- 660 • If the contribution is a dataset and/or model, the authors should describe the steps taken
- 661 to make their results reproducible or verifiable.
- 662 • Depending on the contribution, reproducibility can be accomplished in various ways.
- 663 For example, if the contribution is a novel architecture, describing the architecture fully
- 664 might suffice, or if the contribution is a specific model and empirical evaluation, it may
- 665 be necessary to either make it possible for others to replicate the model with the same
- 666 dataset, or provide access to the model. In general, releasing code and data is often
- 667 one good way to accomplish this, but reproducibility can also be provided via detailed
- 668 instructions for how to replicate the results, access to a hosted model (e.g., in the case
- 669 of a large language model), releasing of a model checkpoint, or other means that are
- 670 appropriate to the research performed.
- 671 • While NeurIPS does not require releasing code, the conference does require all submis-
- 672 sions to provide some reasonable avenue for reproducibility, which may depend on the
- 673 nature of the contribution. For example
 - 674 (a) If the contribution is primarily a new algorithm, the paper should make it clear how
 - 675 to reproduce that algorithm.
 - 676 (b) If the contribution is primarily a new model architecture, the paper should describe
 - 677 the architecture clearly and fully.
 - 678 (c) If the contribution is a new model (e.g., a large language model), then there should
 - 679 either be a way to access this model for reproducing the results or a way to reproduce
 - 680 the model (e.g., with an open-source dataset or instructions for how to construct
 - 681 the dataset).
 - 682 (d) We recognize that reproducibility may be tricky in some cases, in which case
 - 683 authors are welcome to describe the particular way they provide for reproducibility.
 - 684 In the case of closed-source models, it may be that access to the model is limited in
 - 685 some way (e.g., to registered users), but it should be possible for other researchers
 - 686 to have some path to reproducing or verifying the results.

687 5. Open access to data and code

688 Question: Does the paper provide open access to the data and code, with sufficient instruc-

689 tions to faithfully reproduce the main experimental results, as described in supplemental

690 material?

691 Answer: [No]

692 Justification: We have our code ready but since workshop submission does not allow

693 supplementary material we cannot submit them.

694 Guidelines:

- 695 • The answer NA means that paper does not include experiments requiring code.
- 696 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/](https://nips.cc/public/guides/CodeSubmissionPolicy)
- 697 [public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 698 • While we encourage the release of code and data, we understand that this might not be
- 699 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
- 700 including code, unless this is central to the contribution (e.g., for a new open-source
- 701 benchmark).
- 702 • The instructions should contain the exact command and environment needed to run to
- 703 reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 704 [//nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 705 • The authors should provide instructions on data access and preparation, including how
- 706 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 707 • The authors should provide scripts to reproduce all experimental results for the new
- 708 proposed method and baselines. If only a subset of experiments are reproducible, they
- 709 should state which ones are omitted from the script and why.

- 710
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- 711
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.
- 712
- 713

714 6. Experimental Setting/Details

715 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
716 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
717 results?

718 Answer: [Yes]

719 Justification: We have given all experimental details along with hyperparameters in the
720 Experiments section and in Appendix.

721 Guidelines:

- 722 • The answer NA means that the paper does not include experiments.
- 723 • The experimental setting should be presented in the core of the paper to a level of detail
724 that is necessary to appreciate the results and make sense of them.
- 725 • The full details can be provided either with the code, in appendix, or as supplemental
726 material.

727 7. Experiment Statistical Significance

728 Question: Does the paper report error bars suitably and correctly defined or other appropriate
729 information about the statistical significance of the experiments?

730 Answer: [Yes]

731 Justification: We have given all details in the Results section of the paper.

732 Guidelines:

- 733 • The answer NA means that the paper does not include experiments.
- 734 • The authors should answer "Yes" if the results are accompanied by error bars, confi-
735 dence intervals, or statistical significance tests, at least for the experiments that support
736 the main claims of the paper.
- 737 • The factors of variability that the error bars are capturing should be clearly stated (for
738 example, train/test split, initialization, random drawing of some parameter, or overall
739 run with given experimental conditions).
- 740 • The method for calculating the error bars should be explained (closed form formula,
741 call to a library function, bootstrap, etc.)
- 742 • The assumptions made should be given (e.g., Normally distributed errors).
- 743 • It should be clear whether the error bar is the standard deviation or the standard error
744 of the mean.
- 745 • It is OK to report 1-sigma error bars, but one should state it. The authors should
746 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis
747 of Normality of errors is not verified.
- 748 • For asymmetric distributions, the authors should be careful not to show in tables or
749 figures symmetric error bars that would yield results that are out of range (e.g. negative
750 error rates).
- 751 • If error bars are reported in tables or plots, The authors should explain in the text how
752 they were calculated and reference the corresponding figures or tables in the text.

753 8. Experiments Compute Resources

754 Question: For each experiment, does the paper provide sufficient information on the com-
755 puter resources (type of compute workers, memory, time of execution) needed to reproduce
756 the experiments?

757 Answer: [Yes]

758 Justification: We have mentioned in detail the requirements in Experiments section of the
759 paper.

760 Guidelines:

- 761
- The answer NA means that the paper does not include experiments.
 - 762 • The paper should indicate the type of compute workers CPU or GPU, internal cluster,
763 or cloud provider, including relevant memory and storage.
 - 764 • The paper should provide the amount of compute required for each of the individual
765 experimental runs as well as estimate the total compute.
 - 766 • The paper should disclose whether the full research project required more compute
767 than the experiments reported in the paper (e.g., preliminary or failed experiments that
768 didn't make it into the paper).

769 9. Code Of Ethics

770 Question: Does the research conducted in the paper conform, in every respect, with the
771 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

772 Answer: [Yes]

773 Justification: We have reviewed the NeurIPS Code of Ethics and abide by it in the paper.

774 Guidelines:

- 775 • The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- 776 • If the authors answer No, they should explain the special circumstances that require a
777 deviation from the Code of Ethics.
- 778 • The authors should make sure to preserve anonymity (e.g., if there is a special consid-
779 eration due to laws or regulations in their jurisdiction).

780 10. Broader Impacts

781 Question: Does the paper discuss both potential positive societal impacts and negative
782 societal impacts of the work performed?

783 Answer: [Yes]

784 Justification: We discuss in detail both positive and negative societal impacts of our paper in
785 Appendix A

786 Guidelines:

- 787 • The answer NA means that there is no societal impact of the work performed.
- 788 • If the authors answer NA or No, they should explain why their work has no societal
789 impact or why the paper does not address societal impact.
- 790 • Examples of negative societal impacts include potential malicious or unintended uses
791 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations
792 (e.g., deployment of technologies that could make decisions that unfairly impact specific
793 groups), privacy considerations, and security considerations.
- 794 • The conference expects that many papers will be foundational research and not tied
795 to particular applications, let alone deployments. However, if there is a direct path to
796 any negative applications, the authors should point it out. For example, it is legitimate
797 to point out that an improvement in the quality of generative models could be used to
798 generate deepfakes for disinformation. On the other hand, it is not needed to point out
799 that a generic algorithm for optimizing neural networks could enable people to train
800 models that generate Deepfakes faster.
- 801 • The authors should consider possible harms that could arise when the technology is
802 being used as intended and functioning correctly, harms that could arise when the
803 technology is being used as intended but gives incorrect results, and harms following
804 from (intentional or unintentional) misuse of the technology.
- 805 • If there are negative societal impacts, the authors could also discuss possible mitigation
806 strategies (e.g., gated release of models, providing defenses in addition to attacks,
807 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from
808 feedback over time, improving the efficiency and accessibility of ML).

809 11. Safeguards

810 Question: Does the paper describe safeguards that have been put in place for responsible
811 release of data or models that have a high risk for misuse (e.g., pretrained language models,
812 image generators, or scraped datasets)?

813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

Answer: [NA]

Justification: Our paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have utilized fully open-sourced public datasets for all experiments in our paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release our framework code and instruction-tuning data as soon as the peer-review process is over.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.