Fusion-Eval: Integrating Assistant Evaluators with LLMs

Anonymous ACL submission

Abstract

Evaluating natural language systems poses significant challenges, particularly in the realms of natural language understanding and high-level reasoning. In this paper, we introduce "Fusion-Eval", an innovative approach that leverages Large Language Models (LLMs) to integrate insights from various assistant evaluators. Each of these evaluators specializes in assessing distinct aspects of responses. This unique strategy enables Fusion-Eval to function effectively across a diverse range of tasks and criteria, enhancing the effectiveness of existing evaluation methods. Fusion-Eval achieves a 0.962 system-level Kendall-Tau correlation with humans on SummEval and a 0.744 turn-level Spearman correlation on TopicalChat, which is significantly higher than baseline methods. These results highlight Fusion-Eval's significant potential in the realm of natural language system evaluation.

1 Introduction

Evaluating the performance of natural language generation models has significant challenges (Ouyang et al., 2022), particularly in terms of evaluation benchmarks and evaluation paradigms (Wang et al., 2023b). This study focuses on the latter one. Typically, the evaluation paradigms fall into three categories: human-based, automatic-metrics-based and model-based evaluations. Among these, human evaluations are regarded as the most reliable, yet they come with high costs and issues of scalability.

Automatic metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are prevalent in evaluations, relying on comparisons with a 'gold' standard reference. However, the creation of these gold references is a labor-intensive process. Moreover, in tasks involving content generation, the variety of potential correct responses can mean that comparisons to a single or limited number of references may not fully capture the quality of the generated content. Furthermore, studies such as Fabbri et al. (2021) have demonstrated that these automatic metrics often do not correlate well with human judgment.

Model-based evaluations aim to enhance the correlation with human judgment using neural networks fine-tuned on specific datasets. Neural evaluators like BLEURT (Sellam et al., 2020) and its variant SMART (Amplayo et al., 2022) show improved alignment with human assessments in various generative tasks. These models offer flexibility in evaluation methods. As source-dependent (reference-free) evaluators, they directly compare responses to the original content, such as articles in text summarization. As reference-dependent evaluators, they utilize a gold standard reference for more accurate assessment.

Recent advancements have seen the use of Large Language Models (LLMs) as reference-free evaluators in Natural Language Generation (NLG) tasks. Notably, studies by Fu et al. (2023); Wang et al. (2023a) have leveraged LLMs to rate candidate outputs based on their generation probability alone, eliminating the need for reference text comparisons. Additionally, Liu et al. (2023) have introduced a method where LLMs, guided by human-crafted evaluation criteria, score responses. Meta-evaluations indicate that these LLM-based evaluators reach a level of human correlation on par with medium-sized neural evaluators (Zhong et al., 2022). In light of these developments in evaluation paradigms, the following question arises:

> "Can Large Language Models (LLMs) devise an evaluation plan and integrate existing evaluators to achieve higher correlation with human judgments?"

In response to this question, we introduce *Fusion-Eval*, an innovative evaluation framework that integrates a variety of existing evaluators—termed *assistant evaluators*—to enhance correlation with human judgment. Fusion-Eval leverages LLMs not only for direct evaluation but also to adeptly fuse insights from these assistant evaluators. It is designed to work well with different tasks and criteria, maximizing the efficacy of the existing evaluators. Empirical tests conducted on SummEval (Fabbri et al., 2021) and TopicalChat (Mehri and Eskenazi, 2020) validate Fusion-Eval's proficiency in developing and executing an evaluation plan incorporating assistant evaluators. This approach achieves new state-of-the-art correlations with human judgment.

2 Method

Fusion-Eval is a prompt-based evaluation framework leveraging a Large Language Model (LLM) to fuse assistant evaluators, enhancing overall evaluation quality. This process has two primary steps:

2.1 Step 1: Creation of the Fusion-Eval Evaluation Prompt Template

The first step involves creating an evaluation prompt template. This template outlines the evaluation task, criteria, and the strategy for integrating assistant evaluators, along with placeholders for their scores and examples. Central to this template is the LLM-generated plan, which specifies how to strategically utilize assistant evaluators for each criterion, illustrating the LLM's capability to effectively combine diverse evaluators. This plan is then integrated into the template for subsequent execution in Step 2.

Eliciting LLM's Evaluation Plan This describes the process of eliciting a strategic plan for integrating assistant evaluators from the planning LLM. The prompt clarifies the LLM's evaluator role and supplements it with relevant information. Evaluation criteria can be either explicitly specified or left for the LLM to generate. To align with the SummEval and TopicalChat benchmarks, specific evaluation criteria from these datasets were provided to the LLM. The LLM was also informed about various assistant evaluators and requested to create a plan. The planning LLM, in response, developed a detailed plan specifying how each assistant evaluator would be integrated into the evaluation process, ensuring a thorough assessment based on the defined criteria. The information given to the LLM for the SummEval (Fabbri et al., 2021) task is displayed below, with "<...>" indicating condensed sections.

```
You are an evaluation agent. I will give you one summary written for a news
        article . Please evaluate the quality of the summary. <...>
Three assistant evaluators are provided.
1. Natural Language Inference (NLI) provides the probability of the entailed
        relationship between source text (as premise). Its range is between 0−1,
        close to 1 indicates that the hypothesis is entailed by the premise.
2. BLEURT is an evaluation metric for Natural Language Generation. It takes a
        pair of sentences as input, a reference and a candidate, and it returns a
        score that indicates to what extent the candidate is fluent and conveys
        the meaning of the reference .
3. SUM_BLEURT is a variant of BLEURT which is fine−tuned on a summarization
        dataset. It treats the article as the reference and the summary as a
        candidate and it returns a score indicating to what extent the summary is
        coherent and conveys the meaning of the article .
Please share your understanding of the evaluation task and plan for using
        assistant evaluators , including criteria planning and steps. <...>
```

LLM's generated evaluation Plan The LLM's plan includes steps like reviewing sources and summaries and incorporating assistant evaluator scores, pinpointing optimal evaluators for each criterion. Tables 1 display the chosen assistant evaluators for different criteria. The final Fusion-Eval template, incorporating the LLM's plan, features placeholders for test cases and assistant evaluators' scores. The condensed strategic evaluation plan from the planning LLM is below. Full Fusion-Eval templates are available in Appendix A.1 for SummEval and A.2 for TopicalChat.

```
Evaluate a provided summary using criteria : Coherence, Consistency, Relevance,
        and Fluency.
Assistant Evaluators like NLI, BLEURT, and SUM_BLEURT, which give scores
        between below 0 and 1 (closer to 1 being better), will assist in this
        evaluation .
∗∗1. NLI (Natural Language Inference)∗∗:
This assistant evaluator provides a probability score indicating how much the
        summary (hypothesis) is entailed by the original news article (premise).
∗∗Usage∗∗:
− ∗∗Consistency Evaluation∗∗: A high entailment probability indicates that the
        summary is factually aligned with the source text . <...>
∗∗Plan Using Assistant Evaluators∗∗:
1. ∗∗Read the News Article and Summary∗∗: <...>
2. ∗∗Use NLI & BLEURT for Consistency∗∗: <...>
∗∗Criteria & Steps∗∗: <...>
2. ∗∗Consistency (1−5)∗∗:
   − Use NLI & BLEURT to get scores.
   − Read the article and summary.
   − Compare factual details .
   − Assign a consistency score based on factual alignment. <...>
∗∗Evaluation Summary (1−5)∗∗:
Consider the scores from each criterion and their importance. <...>
```

2.2 Step 2: Executing the Evaluation Prompt on Test Examples

In Step 2, the prepared evaluation prompt template is applied to each test example. This template is filled with the inputs, responses, and scores of assistant evaluators for each test case. The executing LLM then processes this filled prompt, yielding Fusion-Eval's final evaluation scores. The details are provided in the Experiment Section (Section 3.1).

3 Experiment

We conduct a meta-evaluation of Fusion-Eval, utilizing the SummEval (Fabbri et al., 2021) and TopicalChat (Mehri and Eskenazi, 2020) benchmarks.

	SummEval					TopicalChat				
	Coh	Con	Flu	Rel		Coh	Eng	Nat	Gro	Und
BLEURT		✓		✓	BLEURT				✓	
NLI		✓			PaLM2 Prob			✓		✓
SumBLEURT	✓			✓						

Table 1: LLM-Suggested Assistant Evaluator Alignment for SummEval and TopicalChat Criteria. The criteria include coherence (Coh), consistency (Con), fluency (Flu), relevance (Rel), engagingness (Eng), naturalness (Nat), groundedness (Gro), and understandability (Und).

	Human Evaluation				
	Coh	Con	Flu	Rel	Overall
Reference-Based Metrics					
ROUGE-1	0.35	0.55	0.527	0.583	0.503
ROUGE-2	0.233	0.6	0.494	0.433	0.44
ROUGE-L	0.117	0.117	0.259	0.35	0.211
BLEU	0.217	0.05	0.326	0.383	0.244
CHRF	0.35	0.617	0.561	0.55	0.519
S1-CHRF	0.3	0.733	0.494	0.5	0.507
S2-CHRF	0.3	0.7	0.46	0.433	0.473
SL-CHRF	0.367	0.733	0.494	0.5	0.523
BERTScore	0.333	-0.03	0.142	0.2	0.161
MoverScore	0.217	-0.05	0.259	0.35	0.194
Source-dependent Metrics					
BARTScore	0.35	0.617	0.494	0.45	0.478
UniEval	0.683	0.75	0.661	0.667	0.728
DE-PaLM2	0.733	0.6	0.745	0.85	0.879
G-Eval (GPT-4)	0.733	0.583	0.778	0.883	0.912
Assistant Evaluators					
BLEURT	0.433	**0.767**	0.644	0.633	0.678
NLI	0.45	0.717	0.628	0.65	0.695
SumBLEURT	0.7	0.333	0.544	0.633	0.644
Fusion-Eval					
FE-PaLM2	**0.783**	**0.767**	0.778	**0.917**	**0.962**
FE-GPT-4	**0.783**	0.762	**0.812**	0.9	0.946

Table 2: System-level Kendall-Tau (τ) correlations of different evaluators to human judgements on SummEval benchmark. The assistant evaluators, BLEURT, NLI and SumBLEURT, treat the article as a premise and the summary as a hypothesis.

3.1 Experiment Setting

SummEval (Fabbri et al., 2021), a benchmark for text summarization evaluation, consists of 1600 data points. Each data point includes average ratings from three experts on a scale of 1 to 5, spanning four summary quality dimensions: coherence (Coh), consistency (Con), fluency (Flu) and relevance (Rel). The "Overall" score is derived as an average across these four dimensions. TopicalChat (Mehri and Eskenazi, 2020), a benchmark for evaluating knowledge-based dialogue response generation, includes 360 data points. It features human evaluations from three experts across six dimensions: coherence (Coh), engagingness (Eng), naturalness (Nat), groundedness (Gro), understandability (Und), and overall. Ratings for naturalness, coherence, and engagingness are on a scale from 1 to 3, while groundedness and understandability are scored between 0 and 1. The overall dimension is evaluated on a scale of 1 to 5. Each data point comprises a conversation history, a grounding fact, and a potential next-turn response. To measure the correlation between results generated by Fusion-Eval and human evaluations, we use Kendall-Tau scores for system-level analysis in SummEval (Fabbri et al., 2021), and Spearman scores for turn-level analysis in TopicalChat (Mehri and Eskenazi, 2020) to align with each benchmark's original scoring methodology.

In our experiments, PaLM2-Large (Anil et al., 2023) and GPT-4 (OpenAI, 2023) serve as the Large Language Models (LLMs) for execution, designated as FE-PaLM2 and FE-GPT-4, respectively. We integrate several assistant evaluators: NLI (Bowman et al., 2015), BLEURT (Sellam et al., 2020), and SumBLEURT—a BLEURT variant fine-tuned for human summarization evaluation (Clark et al., 2023). Additionally, we use the probability of PaLM (PaLM2 Prob) generating a re-

	Human Evaluation					
	Coh (1-3)	Eng (1-3)	Nat (1-3)	Gro (0-1)	Und (0-1)	Overall (1-5)
Source-dependent Metrics						
UniEval	0.613	0.605	0.514	0.575	0.468	0.663
DE-PaLM2	0.669	0.688	0.542	0.602	0.493	0.66
G-Eval (GPT-4)	0.605	0.691	0.565	0.551	-	-
Assistant Evaluators						
BLEURT	0.316	0.461	0.384	0.638	0.432	0.464
PaLM2 Prob	0.583	0.606	0.637	0.441	0.676	0.687
Fusion-Eval						
FE-PaLM2	**0.697**	0.728	0.651	**0.709**	0.632	0.764
FE-GPT-4	0.678	**0.747**	**0.691**	0.692	**0.687**	**0.774**

Table 3: Turn-level Spearman (ρ) correlations of different evaluators to human judgements on TopicalChat benchmark. BLEURT treats the fact and conversation as the premise and the response as the hypothesis. PaLM2 Prob represents the conditional probability of the response given the fact and conversation.

	FE-PaLM2				
	Coh	Con	Flu	Rel	Overall
BLEURT	0.583	0.867	0.733	0.65	0.717
NLI	0.6	0.783	0.75	0.667	0.733
SumBLEURT	0.75	0.467	0.633	0.717	0.683

Table 4: FE-PaLM2 and Assistant Evaluators System-level Kendall-Tau (τ) correlations on SummEval.

	FE-PaLM2					
	Coh	Eng	Nat	Gro	Und	Overall
BLEURT	0.524	0.558	0.59	0.662	0.622	0.67
PaLM2 Prob	0.711	0.784	0.808	0.588	0.711	0.792

Table 5: FE-PaLM2 and Assistant Evaluators Turn-level Spearman (ρ) correlations on TopicalChat.

	FE-GPT-4				
	Coh	Con	Flu	Rel	Overall
BLEURT	0.583	0.795	0.733	0.6	0.7
NLI	0.633	0.745	0.717	0.617	0.717
SumBLEURT	0.717	0.41	0.633	0.667	0.667

Table 6: FE-GPT-4 and Assistant Evaluators System-level Kendall-Tau (τ) correlations on SummEval.

	FE-GPT-4					
	Coh	Eng	Nat	Gro	Und	Overall
BLEURT	0.577	0.644	0.565	0.693	0.617	0.678
PaLM2 Prob	0.747	0.713	0.86	0.662	0.799	0.798

Table 7: FE-GPT-4 and Assistant Evaluators Turn-level Spearman (ρ) correlations on TopicalChat.

sponse based on prior conversation and context as an assistant evaluator, following methods in studies by Fu et al. (2023) and Wang et al. (2023a). For the execution of Fusion-Eval, the evaluation prompt template is filled with specific inputs, responses, and assistant evaluator scores for each test case. This complete prompt is then processed by the executing LLM, which generates a score for each evaluation dimension. The LLMs are configured to produce 8 predictions with temperatures of 0.5 for PaLM2 and 0.1 for GPT-4.

3.2 Baselines

For a thorough comparison, we meta-evaluated Fusion-Eval against a range of baseline methods on the SummEval benchmark. These baselines include ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), CHRF (Popović, 2015), SMART (Amplayo et al., 2022), BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), BARTScore (Yuan et al., 2021), UniEval (Zhong et al., 2022), and G-Eval (Liu et al., 2023). We derived scores for most baselines from the SMART paper (Amplayo et al., 2022), while for UniEval[1] and G-Eval[2], we calculated scores using their publicly available predictions. For the TopicalChat benchmark, we compared Fusion-Eval's performance with G-Eval (Liu et al., 2023) and UniEval (Zhong et al., 2022), utilizing scores from their respective publications. We also introduce DE-PaLM2 (Direct Evaluator PaLM2) as an ablation baseline. DE-PaLM2 uses the same approach as FE-PaLM2 but without including assistant evaluators and their scores in the template. This baseline provides insights into PaLM2's standalone performance on the SummEval and TopicalChat benchmarks.

[1] https://github.com/maszhongming/UniEval

[2] https://github.com/nlpyang/geval

3.3 Result Analysis

Tables 2 and 3 present the correlation of baselines, assistant evaluators, and Fusion-Eval with human judgment. Tables 4 and 5 illustrate the correlation of assistant evaluators with FE-PaLM2. Similarly, Tables 6 and 7 detail the correlation of assistant evaluators with FE-GPT-4.

Does Fusion-Eval achieve better correlation with human evaluation? Yes. As detailed in Tables 2 and 3, FE-PaLM2 and FE-GPT-4 outperforms all baselines and assistant evaluators. Notably, in the "Overall" dimension, FE-PaLM2 and FE-GPT-4 demonstrates superior alignment with human judgments, surpassing state-of-the-art methods. Moreover, FE-PaLM2 and FE-GPT-4 significantly improve LLM performance in weaker dimensions by incorporating assistant evaluators, as seen in SummEval's coherence and consistency, and TopicalChat's naturalness, groundedness, and understandability, especially when compared to direct LLM evaluation methods such as DE-PaLM2 and G-Eval.

Does Fusion-Eval optimally integrate the assistant evaluators during execution? Likely. When looking at the correlation of assistant evaluators to FE-PaLM2 (as shown in Tables 4 and 5) and to FE-GPT-4 (as shown in Tables 6 and 7) together with the LLM's strategic plan (as shown in Tables 1), we notice that selected assistant evaluators consistently show higher correlation with FE-PaLM2 and FE-GPT-4. For example, in SummEval's coherence, SumBLEURT demonstrates a higher correlation than other evaluators. A similar trend is also observed in TopicalChat's naturalness and understandability. Additionally, none of the correlations between assistant evaluators and Fusion-Eval equals "1", suggesting that Fusion-Eval's approach uses assistant evaluators to supplement its judgment rather than relying entirely on them.

4 Conclusion

The paper presents Fusion-Eval, an innovative aggregator using Large Language Models (LLMs) for diverse evaluation tasks. It effectively integrates assistant evaluators according to specific criteria. Empirical results show Fusion-Eval achieves higher correlations with human judgments than baselines.

5 Limitation

The lengthy Fusion-Eval evaluation prompt may challenge LLMs, particularly those with limited context windows. These extensive prompts could exceed their processing capabilities. To address this, we are considering prompt decomposition for future exploration. This approach could make Fusion-Eval more adaptable and efficient for different LLM setups.

References

Reinald Kim Amplayo, Peter J Liu, Yao Zhao, and Shashi Narayan. 2022. Smart: sentences as basic units for text evaluation. *arXiv preprint arXiv:2208.01030*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. *arXiv preprint arXiv:2305.13194*.

Alexander R Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

A Appendix

A.1 Fusion-Eval Evaluation Prompt Template for SummEval (One Prompt Only in This Subsection - Do Not Be Surprised by Its Length)

Evaluate a provided summary using criteria : Coherence, Consistency, Relevance, and Fluency.

Assistant Evaluators like NLI, BLEURT, and SUM_BLEURT, which give scores between below 0 and 1 (closer to 1 being better), will assist in this evaluation.
1. NLI (Natural Language Inference):
This assistant evaluator provides a probability score indicating how much the summary (hypothesis) is entailed by the original news article (premise).
Usage:
− **Consistency Evaluation**: A high entailment probability indicates that the summary is factually aligned with the source text. Conversely, a low score might indicate discrepancies or hallucinated facts.

2. BLEURT:
This metric models human judgments. It gives a score indicating how closely the summary aligns with what human evaluators might consider a good summary given the source text.
Usage:
− **Relevance and Consistency Evaluation**: A high BLEURT score would suggest that the summary effectively captures the essential points of the source. A low score might indicate missing key points.

3. SUM_BLEURT (Summarization BLEURT):
Fine−tuned on a summarization dataset, this assistant evaluator offers a more targeted approach to measuring the quality of summaries in the context of human judgments.
Usage:
− **Relevance and Coherence Evaluation**: Like BLEURT, but given its specialization in summarization, SUM_BLEURT could offer more precise insights into the relevance and coherence of the summary in relation to the source text.

Plan Using Assistant Evaluators:
1. **Read the News Article and Summary**: Begin with a manual reading to form an initial impression.
2. **Use NLI & BLEURT for Consistency**: Check both scores. High scores from both assistant evaluators will reaffirm the consistency of the summary.
3. **Use BLEURT & SUM_BLEURT for Relevance**: Check scores from both assistant evaluators. High scores would suggest a good summary in terms of relevance.
4. **Use SUM_BLEURT for Coherence**: Check SUM_BLEURT score. High scores would suggest a good summary in terms of coherence.
5. **Manual Evaluation for Fluency**: The assistant evaluators don't directly address fluency. You'll evaluate grammar, punctuation, and sentence structure manually.
6. **Final Judgment**: The assistant evaluators' outputs will inform and validate your evaluations, but the ultimate judgment will be based on the provided criteria and steps, with the assistant evaluators serving as supplementary aids.

** Criteria & Steps**:
1. **Coherence (1−5)**:
   − Read the news article and the summary.
   − Compare the summary to the article for clarity and logical order.
   − Use SUM_BLEURT scores as supplementary insights for coherence.
   − Assign a coherence score based on organization and structure.

2. **Consistency (1−5)**:
   − Use NLI & BLEURT to get scores.
   − Read the article and summary.
   − Compare factual details.
   − Assign a consistency score based on factual alignment.

3. **Relevance (1−5)**:
   − Use BLEURT & SUM_BLEURT to get alignment scores with human−like judgments.
   − Read both the article and summary.
   − Identify main points and coverage in the summary.
   − Assign a relevance score based on content importance and absence of redundancies.

4. **Fluency (1−5)**:
   − Evaluate the summary manually for grammar, punctuation, and sentence structure.
   − Assign a fluency score based on readability.

Evaluation Summary (1−5):
Consider the scores from each criterion and their importance.
– Derive an average score, ensuring the final score ranges between 1−5.
– Provide overall comments on the summary.
– Highlight strengths and areas needing improvement.

Input Template:
Source:
[Provide the source text here]

Answer:
[Provide the summary text here]

NLI Score (Source as Premise and Answer as Hypothesis):
[Provide NLI entailment probability score]

BLEURT Score (Source as Premise and Answer as Hypothesis):
[Provide BLEURT score]

SUM_BLEURT Score (Source as Premise and Answer as Hypothesis):
[Provide SUM_BLEURT score]

Output Template:
Criterias ' Scores and Explanations:

Coherence
Score: [Your evaluation] Explanation: [Your explanation on evaluation]

Consistency
Score: [Your evaluation] Explanation: [Your explanation on evaluation]

Relevance
Score: [Your evaluation] Explanation :[Your explanation on evaluation]

Fluency
Score: [Your evaluation] Explanation: [Your explanation on evaluation]

Evaluation Summary:
Overall Score: [Your evaluation]
Explanation: [Your explanation on evaluation]

Input Example:
Source:
[[source]]

Answer:
[[summary]]

NLI Score (Source as Premise and Answer as Hypothesis):
[[nli_score_source_answer]]

BLEURT Score (Source as Premise and Answer as Hypothesis):
[[bleurt_score_source_answer]]

SUM_BLEURT Score (Source as Premise and Answer as Hypothesis):
[[sum_bleurt_score_source_answer]]

Evaluation (please follow Output Template and provide the evaluation result):<< eval_result >>

A.2 Fusion-Eval Evaluation Prompt Template for TopicalChat (One Prompt Only in This Subsection - Do Not Be Surprised by Its Length)

You will be given a conversation between two individuals, followed by a potential response for the next turn in the conversation, which includes an interesting fact. Your task is to rate the responses on six metrics: Coherence, Engagingness, Naturalness, Groundedness, Understandability, and Overall Quality.

Assistant Evaluators' Descriptions and Usage:
1. LM_PROB (Language Model Probability):
- **Functionality**: LM_PROB provides a probability score, ranging from 0 to 1, indicating the likelihood that a given response would be generated by a language model, given the preceding conversation and fact.
- **Score Range**: 0 (least likely) to 1 (most likely).
- **Usage**:
  - **Naturalness Evaluation**: A higher probability score suggests that the response is more likely to occur naturally in human conversation, indicating greater naturalness.
  - **Understandability Evaluation**: Similarly, a higher probability can also imply that the response is more understandable within the given context, as it is more aligned with expected language patterns.

2. BLEURT:
- **Functionality**: BLEURT evaluates the quality of text generation by comparing the generated text (response) to a reference (conversation and fact). Its score range is 0 to 1, where higher scores indicate better alignment and quality.
- **Score Range**: 0 (poor alignment) to 1 (excellent alignment).
- **Usage**:
  - **Groundedness Evaluation**: A high BLEURT score indicates that the response accurately and relevantly utilizes the given fact, showing strong groundedness in the context of the conversation.

Plan Using Tools for Conversation Response Evaluation:
1. **Read the Conversation, Fact, and Response**: Begin with a careful reading of the provided materials to form an initial qualitative impression of the response in the context of the conversation and fact.
2. **Use LM_PROB for Naturalness and Understandability Evaluation**:
   - Apply LM_PROB to determine the probability that the response would be generated by a language model in the given context.
   - High probability scores from LM_PROB will indicate greater naturalness and understandability, as the response aligns well with expected language patterns.
3. **Use BLEURT for Groundedness Evaluation**:
   - Employ BLEURT to assess how accurately and relevantly the response utilizes the given fact in the context of the conversation.
   - A high score from BLEURT suggests that the response is well-grounded in the provided fact, demonstrating accuracy and relevance.
4. **Final Judgment and Integration of Tool Outputs**:
   - Integrate the outputs from the tools with your initial qualitative assessment.
   - The tools' outputs will provide quantitative support and validation for your evaluations in each metric.
   - Make the final judgment based on a holistic view, considering both the tool outputs and the original evaluation criteria for each metric.
   - Remember that the ultimate judgment should align with the predefined criteria and evaluation steps, with the tools serving as important but supplementary aids in the decision-making process.

** Criteria & Steps**:
1. **Coherence (1-3, Any Floating Value)**:
   - Read the conversation, fact, and response to assess the logical flow and continuity.
   - Evaluate how well the response connects with and continues the conversation.
   - Assign a Coherence score, ranging from 1 to 3, based on the response's organization and logical integration into the conversation.

2. **Engagingness (1-3, Any Floating Value)**:
   - Review the conversation, fact, and response to determine the level of interest or intrigue.
   - Assess how the response contributes to the conversation's value and captivates interest.
   - Assign an Engagingness score, ranging from 1 to 3, based on the response's ability to captivate and add value to the conversation.

3. **Naturalness (1-3, Any Floating Value)**:
   - Read the conversation, fact, and response to gauge the natural fit of the response within the conversation's context.
   - Evaluate the tone, formality, and conversational flow to determine how naturally the response fits.
   - Use LM_PROB to supplement the evaluation, considering the likelihood of such a response in the given context.
   - Assign a Naturalness score, ranging from 1 to 3, focusing on how naturally the response fits into the conversation.

4. **Groundedness (0-1, Any Floating Value)**:
   - Examine the conversation, fact, and response to evaluate how well the response utilizes the given fact.
   - Assess the accuracy and relevance of the fact in the response.
   - Utilize BLEURT to provide supplementary insights into how accurately the response is grounded in the given

```
        fact .
    − Assign a Groundedness score, ranging from 0 to 1, based on the effective and accurate incorporation of the
        fact in the response.

5. ∗∗ Understandability (0−1, Any Floating Value)∗∗:
    − Review the conversation , fact , and response to assess the clarity and comprehension of the response.
    − Focus on how clearly and easily the response can be understood within the context of the preceding
        conversation .
    − Apply LM_PROB for additional data on the understandability of the response.
    − Assign an Understandability score, ranging from 0 to 1, based on the response's clarity and ease of
        comprehension in context.

6. ∗∗Overall Quality (1−5, Any Floating Value)∗∗:
    − Review the scores and insights from the previous criteria , including data from assistant evaluators .
    − Consider how the aspects of Coherence, Engagingness, Naturalness , Groundedness, and Understandability
        collectively contribute to the overall impression of the response.
    − Assign an Overall Quality score, ranging from 1 to 5, based on a holistic assessment of the response's
        strengths and weaknesses.
    − Provide a summary explanation for the overall quality rating , highlighting key factors and insights that
        influenced the judgment.


∗∗Input Template∗∗:
Conversation:
[Provide the conversation text here]

Fact:
[Provide the fact text here]

Response:
[Provide the response text here]

LM_PROB Score (Response in Context of Conversation and Fact):
[Provide LM_PROB probability score]

BLEURT Score (Response with Conversation and Fact as Reference):
[Provide BLEURT score]


∗∗Output Template∗∗:
 Criteria Scores and Explanations:

Coherence
Score: [Your evaluation ] Explanation: [Your explanation on evaluation ]

Engagingness
Score: [Your evaluation ] Explanation: [Your explanation on evaluation ]

Naturalness
Score: [Your evaluation ] Explanation: [Your explanation on evaluation ]

Groundedness
Score: [Your evaluation ] Explanation: [Your explanation on evaluation ]

 Understandability
Score: [Your evaluation ] Explanation: [Your explanation on evaluation ]

Evaluation Summary:
Overall Score: [Your evaluation ] Explanation: [Your comprehensive explanation on the overall evaluation ,
      integrating aspects from each criterion ]


∗∗Input Example∗∗:
Conversation:
[[ conversation ]]

Fact:
[[ fact ]]

Response:
```

[[response]]

LM_PROB Score (Response in Context of Conversation and Fact):
[[lm_prob_score]]

BLEURT Score (Response with Conversation and Fact as Reference):
[[bleurt_score]]

Evaluation (please follow Output Template and provide the evaluation result):<< eval_result >>