SEEval: Advancing LLM Text Evaluation Efficiency and Accuracy through Self-Explanation Prompting

Meng-Chen Wu, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Shayan Ali Akbar, Erwin Cornejo
Customer Experience and Business Trends, Amazon.com
{mengchw, hosmdmos, tesswoo, sichi, shayaakb, eccornej}@amazon.com

Abstract

Large language models (LLMs) have achieved remarkable success in various natural language generation (NLG) tasks, but their performance in automatic text evaluation is not yet ready as human replacements. In this paper, we propose SEEval (**S**elf-**E**xplanation in **Eval**uation), a novel prompt-based text evaluator. Inspired by educational psychology, SEEval incorporates self-explanation, a metacognitive strategy, to enhance automatic text evaluation. Our experimental results show that SEEval, without probability normalization, is able to achieve competitive and often superior performance compared to the two state-of-the-art baselines – G-Eval and Analyze-Rate – across all evaluation dimensions and is 20 times more efficient in terms of run-time. The SEEval method is also generalizable as its results are consistent across three other selected LLMs – Claude 3.5 Sonnet, Command R+, and Mistral-Large 2.

1 Introduction

Leveraging LLMs in automatic text evaluation[1] is an emerging research area (Fu et al., 2023; Wang et al., 2023; Chiang and Lee, 2023a; Liu et al., 2023; Lin and Chen, 2023). However, LLMs are not ready as human replacements, as their performance can be inconsistent and dimension-dependent. LLMs may also result in misleading and unreliable evaluations (Shen et al., 2023). In this paper, we propose a new approach that integrates an effective learning technique from educational psychology literature, known as "self-explanation," to enhance the performance of LLMs in text evaluation tasks. It is a metacognitive strategy that involves explaining to oneself the reasoning behind a problem or concept while solving or learning it (Berry, 1983).

Self-explanation was first shown by Berry (Berry, 1983) to enhance logical reasoning and transfer of learning from concrete to abstract problems. Experimental participants who were asked to verbalize (or self-explain) their reasoning either during or after solving concrete problems later significantly outperformed a control group on a set of abstract problems (achieving 89.6% or 67.7% accuracy, depending on whether they self-explained during or after the task, compared to 22.9% accuracy for the control group). Further research has confirmed self-explanation to be an effective learning strategy for improving logical reasoning skills and understanding of new knowledge (Chi et al., 1994; Rittle-Johnson, 2006; Chamberland et al., 2011; Bisra et al., 2018).

Inspired by these findings, SEEval is designed to integrate the self-explanation technique into G-Eval (Liu et al., 2023), a state-of-the-art (SoTA) approach for evaluating NLG systems using LLMs. Our contributions are as follows:

1. We introduce SEEval, a new prompt-based text evaluator that incorporates the concept of self-explanation from educational psychology. The proposed method is effective in improving the evaluation quality of open-domain dialogue systems and text summarization models, especially in Coherence and Groundedness metrics on the Topical-Chat dataset as well as the Consistency metric on the SummEval dataset (Section 4).

2. Our findings show that SEEval can: 1) achieve superior performance without the need for probability normalization, reducing computational costs and time by 20 times (Section 4.1 and 4.2); and 2) generalize across different LLMs (Section 4.3).

We organize the paper as follows. We detail our proposed methods in Section 2 and present experimental results in Section 3. Section 4 describes our analyses and key findings. Section 5 covers prior research related to using LLMs for evaluation tasks. Section 6 concludes our contributions and describes future work directions.

2 Method

SEEval incorporates self-explanation into the G-Eval framework (Liu et al., 2023). Our method first creates a Self-Explain Prompt given the evaluation task description, and then uses this prompt to generate evaluation scores for NLG outputs across different quality dimensions. We detail each component of our approach in the following subsections.

[1] Automatic text evaluation refers to the evaluation of output quality from NLG systems, such as dialogue generation, summarization, and question-answering.

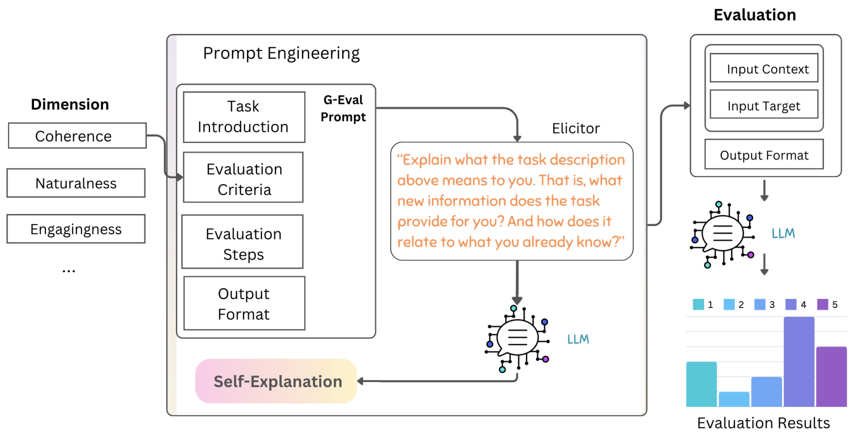

Figure 1: Overview of SEEval. The Self-Explain Prompt is created by using the Elicitor prompt to generate a self-explanation based on the G-Eval Prompt. The self-explanation is combined with the G-Eval Prompt, the elicitor prompt, placeholders for input text, and output format to form the complete Self-Explain Prompt. Finally, we use the Self-Explain Prompt to produce the evaluation results on NLG outputs.

2.1 Background on G-Eval

As our work builds upon G-Eval, we first briefly introduce its framework. G-Eval consists of three main components: 1) the description of the evaluation task including task introduction and evaluation criteria for a particular dimension (e.g., Coherence), 2) an auto Chain-of-Thought (CoT) mechanism that generates evaluation steps, and 3) a scoring function that determines how the final evaluation score is calculated. Building upon this foundation, we extend G-Eval by incorporating self-explanation mechanisms into the evaluation process.

2.2 Self-Explanation

The principle of self-explanation is to enhance learning by prompting individuals to explicitly articulate their understanding of new information and relate it to their prior knowledge. This process facilitates deeper comprehension, retention, and the ability to transfer learned knowledge to new contexts.

In the case of SEEval, the LLM goes through a parallel process of recasting and expanding upon its evaluation instructions. We propose two mechanisms that explain how this process benefits evaluation performance. First, the model can reframe a task to match similar tasks or information from its prior training, where increased task familiarity has been shown to improve performance (McCoy et al., 2023). Second, during the process, a model may also decompose a high-level or subjective task into smaller, simpler and/or more objective steps that may improve its performance (Lee et al., 2024).

2.3 Details of SEEval

Figure 1 provides an overview of the SEEval workflow. The key component of SEEval is the creation and use of a Self-Explain Prompt. This prompt is created by first combining a G-Eval Prompt with an Elicitor prompt to generate a self-explanation. The self-explanation is then combined with the G-Eval Prompt, the Elicitor prompt, placeholders for input text, and output format to form the complete Self-Explain Prompt. Finally, this prompt is used to produce evaluation results on NLG outputs across specific dimensions (e.g., Coherence, Naturalness, Engagingness).

The Elicitor prompt (e.g., "Explain what the task description means to you. What new information does the task provide? How does it relate to what you already know") mirrors instructions used in educational psychology studies (Dunlosky et al., 2013). We specifically use this form of prompt as it implements the core principle of self-explanation: enhancing performance by having learners explicitly articulate their understanding of new information and relate it to their prior knowledge. Since evaluation criteria differ for each dimension in the G-Eval Prompt, we generate dimension-specific self-explanations (example shown in Figure 2, Appendix A). To ensure reproducibility, we set the decoding temperature to 0 during generation. The resulting dimension-specific Self-Explain Prompt is then used to evaluate input texts (e.g., conversation history context and associated NLG output response) against their respective dimensions (e.g., Engagingness).

2.3.1 How is SEEval different from CoT?

Unlike CoT which focuses on step-by-step reasoning for problem-solving, SEEval specifically prompts the model to articulate its understanding of the task and relate it to its prior knowledge, leading to two distinct mechanisms: 1) reframing tasks in terms of familiar concepts from the model's training, and 2) decomposing subjective criteria into more objective sub-tasks.

2.3.2 How is SEEval different from Analyze-Rate?

Analyze-Rate (Chiang and Lee, 2023b), which builds upon G-Eval, enhances evaluation by instructing LLMs to generate an analysis before providing a rating. Specifically, it generates an explanation for each individual NLG output before providing a rating, operating at

	Naturalness		Coherence		Engagingness		Groundedness		AVG		Latency
Metrics	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	(sec.)
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244	-
BERTScore	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273	-
UniEval	0.455	0.330	0.602	0.455	0.573	0.430	0.577	0.453	0.552	0.417	-
w/ Probs											
G-EVAL	0.546	0.492	0.475	0.462	0.610	0.565	0.745	0.705	0.594	0.556	30,078
Analyze-Rate	**0.623**	**0.569**	0.531	0.533	**0.729**	**0.724**	0.791	0.770	**0.668**	0.649	23,841
SEEval	0.597	0.581	**0.535**	**0.538**	0.713	0.702	**0.808**	**0.786**	0.663	**0.652**	28,333
w/o Probs											
G-EVAL	0.485	0.500	0.436	0.448	0.623	0.626	0.755	0.745	0.575	0.580	1,467
Analyze-Rate	**0.606**	**0.609**	0.510	0.525	0.669	0.674	0.772	0.751	0.639	0.640	1,574
SEEval (ours)	0.578	0.580	**0.528**	**0.537**	**0.688**	**0.701**	**0.819**	**0.802**	**0.653**	**0.655**	1,607*

Table 1: Turn-level Pearson (r) and Spearman (ρ) correlations of different metrics obtained with the baselines and SEEval methods on Topical-Chat benchmark. Results are obtained with Claude 3 Sonnet. *See the explanation about the latency in Appendix D.

the instance level. In contrast, SEEval leverages self-explanation at the task level by having the LLM reframe and decompose the entire evaluation task using its prior knowledge. The resulting dimension-specific Self-Explain Prompt (as shown in Figure2, Appendix A) enables SEEval to establish a robust evaluation framework that can be consistently applied across all NLG outputs for a specific dimension (e.g., Coherence).

3 Experimental Design

Our study aims to evaluate whether incorporating self-explanation can strengthen the correlation between human and LLM-generated evaluation scores in the G-Eval framework. We conduct two sets of experiments: 1) comparing SEEval's efficiency and effectiveness against existing LLM-based methods (results in Section 4.1), and 2) assessing SEEval's generalizability across three additional LLM models – Claude 3.5 Sonnet, Command R+, and Mistral-Large 2 (results in Section 4.3).

3.1 Baselines and Datasets

We compare SEEval against two state-of-the-art LLM-based methods: G-Eval (Liu et al., 2023) and Analyze-Rate (Chiang and Lee, 2023b). We also include three common baselines: ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and UniEval (Zhong et al., 2022). For evaluation, we use two widely-adopted NLG benchmarks[2]: Topical-Chat (Mehri and Eskenazi, 2020) and SummEval (Fabbri et al., 2021). Following previous studies (Liu et al., 2023; Chiang and Lee, 2023b), we use Pearson and Spearman correlations for Topical-Chat, and Spearman and Kendall-Tau for SummEval.

3.2 Implementation Details

For our primary experiments, we use Claude 3 Sonnet (Anthropic, 2024) as the underlying model[3]. To ensure fair comparison, all LLM-based baselines are run using the same model. Following Liu et al. (2023), we evaluate using two scoring approaches: probability normalization and direct scoring. For probability normalization (labeled "w/ Probs"), we sample 20 times with temperature and top_p set to 1, as Claude 3 Sonnet does not support token probabilities. For direct scoring (labeled "w/o Probs"), we use temperature 0 for reproducibility. In our generalizability experiments, we apply direct scoring across all three LLMs. The G-Eval Prompts used for both benchmarks are detailed in Appendix C.

4 Results

We present our experimental findings in three parts. First, we assess SEEval's efficacy by comparing it with baseline methods on both Topical-Chat and SummEval benchmarks (Section 4.1). Second, we analyze the impact of probability normalization on performance and computational efficiency (Section 4.2). Finally, we demonstrate SEEval's generalizability across different LLM architectures (Section 4.3).

4.1 Assessing the Efficacy of SEEval

With Topical-Chat Benchmark. Table 1 presents the Pearson (r) and Spearman (ρ) correlation results. The results for the w/ Probs setup demonstrate that SEEval outperforms G-Eval across all four dimensions. On average, SEEval achieves a 9.9% higher Pearson correlation and a 17.8% higher Spearman correlation than G-Eval. Meanwhile, SEEval exhibits a comparative result to Analyze-Rate, although it is computationally expensive, requiring 20 separate runs.

On the other hand, in the w/o Probs setup, SEEval outperforms G-Eval across all four dimensions and outperforms Analyze-Rate in all dimensions except for Naturalness. These results indicate that SEEval can be the preferred choice in a deterministic setup, requiring only a single run.

With SummEval Benchmark. Table 2 shows the summary-level Spearman (ρ) and Kendall-Tau (τ) cor-

[2]Details of two datasets can be found in Appendix B.

[3]The choice of this LLM is motivated by the practical consideration of enabling fair comparisons with models that are widely accessible across various organizations.

Methods	Coherence		Consistency		Fluency		Relevance		AVG		Latency
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	(sec.)
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128	-
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175	-
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377	-
w/ Probs											
G-EVAL	0.526	0.415	0.470	0.405	0.463	0.386	0.477	0.382	0.484	0.397	163,345
Analyze-Rate	**0.534**	**0.425**	0.562	0.512	**0.494**	**0.416**	**0.529**	**0.429**	**0.530**	**0.445**	141,585
SEEval	0.508	0.401	**0.609**	**0.563**	0.469	0.391	0.526	0.420	0.528	0.444	152,608
w/o Probs											
G-EVAL	0.454	0.394	0.571	0.545	0.470	0.440	0.452	0.396	0.487	0.444	7,083
Analyze-Rate	0.438	0.375	0.620	0.593	**0.475**	**0.441**	0.452	0.393	0.496	0.450	5,375
SEEval (ours)	**0.487**	**0.418**	**0.643**	**0.616**	0.461	0.424	**0.504**	**0.439**	**0.524**	**0.474**	6,270

Table 2: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics obtained with the baselines and SEEval methods on SummEval benchmark. Results are obtained with Claude 3 Sonnet.

Methods	Claude 3.5 Sonnet		Command R+		Mistral-Large 2	
Topical-Chat	r	ρ	r	ρ	r	ρ
G-EVAL	0.683	0.688	0.597	0.590	0.665	0.654
Analyze-Rate	0.659	0.662	0.555	0.553	0.691	0.684
SEEval	**0.690**	**0.696**	**0.636**	**0.632**	**0.700**	**0.688**
SummEval	ρ	τ	ρ	τ	ρ	τ
G-EVAL	0.536	0.487	0.505	0.462	0.520	0.472
Analyze-Rate	0.537	0.486	0.462	0.419	0.547	0.496
SEEval	**0.562**	**0.509**	**0.521**	**0.475**	**0.564**	**0.509**

Table 3: Average correlations across methods for our best setting (w/o Probs) using 3 additional LLMs. For Topical-Chat, we report turn-level Pearson (r) and Spearman (ρ), while for SummEval, we show summary-level Spearman (ρ) and Kendall-Tau (τ).

relations on the SummEval benchmark. We observe similar results for SummEval as for Topical-Chat in the w/ Probs setup. On average, SEEval is better than G-Eval (i.e. ρ: 0.528 vs. 0.484; τ: 0.444 vs. 397), while SEEval and Analyze-Rate are close competitors across all four dimensions.

On the other hand, in w/o Probs, SEEval outperforms G-Eval and Analyze-Rate across the board, except for Fluency. On average, compared to Analyze-Rate, SEEval achieves a 5.6% higher Spearman correlation (0.524 vs. 0.496) and a 5.3% higher Kendall-Tau correlation (0.474 vs. 0.450).

4.2 Impact of Probability Normalization

The w/ Probs setting (temperature=1) is computationally expensive, requiring 20 separate runs for the same data. For instance, it consumes approximately 7.6 hours and 42 hours of inference time for a single evaluation dimension on Topical-Chat and SummEval respectively[4]. Both G-Eval and Analyze-Rate rely on this time-consuming process to achieve optimal performance. This multiple-run approach helps address two key limitations of LLM scoring: 1) the tendency to produce certain integer scores (e.g., defaulting to 3 on a 1-5 scale), which reduces score variance and weakens correlation with human judgments, and 2) the reluctance to output decimal scores even when explicitly prompted, resulting in tied scores that fail to capture subtle differences between texts (Liu et al., 2023).

On the other hand, SEEval attains its best results in the w/o Probs setting (temperature=0). This success can be attributed to the generated self-explanation in SEEval (see the text highlighted in red in Figure 2, Appendix A), which reframes the evaluation task into already known sub-tasks, thus guiding the LLM to effectively leverage its prior knowledge about the evaluation task. With temperature set to 0, SEEval follows these sub-tasks in a consistent way, eliminating the need to explore multiple possibilities in output scores. This not only significantly reduces computational costs but also makes SEEval more accessible and practical for evaluating NLG outputs. A case study comparing evaluation outputs from SEEval and the two LLM-based baselines is provided in Appendix E.

When SEEval is under the w/ Probs setting, it introduces stochastic sampling that causes variations in each sub-task's output, with these variations compounding across steps to yield unexpectedly divergent responses. This explains why SEEval achieves optimal performance in w/o Probs setting, while methods like Analyze-Rate and G-Eval rely on exploring multiple

[4]We run the Claude 3 Sonnet using Amazon Bedrock API. The latency is calculated based on the time elapsed between sending the first API Call and receiving all model responses on the entire dataset.

possibilities to arrive at a robust score.

4.3 Assessing the Generalizability of SEEval

To assess generalizability of our method, we employ the latest version of Claude (i.e., Claude Sonnet 3.5) and two open-source LLMs, Cohere's Command R+ and Mistral Large 2, as underlying models for w/o Probs setup and evaluate their performance on both Topical-Chat and SummEval benchmarks. The average correlation results are presented in Table 3, with the complete results shown in Appendix F. Similar to the case with Claude 3 Sonnet, our approach outperforms the baseline methods with all three LLMs, indicating the generalizability of our approach across different language models.

5 Related Work

Non-LLMs NLG Evaluation Traditional automatic text evaluation methods use n-gram-based metrics (Lin, 2004; Papineni et al., 2002; Banerjee and Lavie, 2005) that compute lexical overlap between a generated text and a reference text. Later automatic evaluation works (Zhang et al., 2019; Zhao et al., 2019) leveraged contextualized embeddings from pre-trained LM to measure semantic similarity between two pieces of text and had shown improved performance. More recent works have developed unified evaluators (Yuan et al., 2021; Zhong et al., 2022) that are capable of evaluating multiple dimensions of a text. Because the multi-dimensional evaluation schema aligns with the one in human evaluation, their performance has improved to be moderately correlated with human judgements.

LLM-based NLG Evaluation With the rise of LLMs, researchers further developed zero-shot LLM-based text evaluators (Chiang and Lee, 2023a; Fu et al., 2023; Kocmi and Federmann, 2023; Liu et al., 2023; Lin and Chen, 2023) that push performance to new limits. Others have leveraged human labeled evaluation data to develop new evaluation criteria (Liu et al., 2024b,c) for a given NLG task. Additionally, Chan et al. (2023) and Kim et al. (2024) have employed multi-agent frameworks to enhance LLM evaluation capability through multiple rounds of discussion or debate among different LLM-based agents. Another line of researchers focus on fine-tuning a relatively small open-source language model (Xu et al., 2023; Li et al., 2023; Kim et al., 2023; Liu et al., 2024a; Hu et al., 2024). In contrast, our work focuses on developing a zero-shot, reference-free text evaluator that does not rely on human-labeled data, revising LLM's initial output or fine-tuning. We aim to leverage the inherent capabilities of LLM to assess text quality, thus offering a more flexible and generalizable evaluation framework.

6 Conclusion

Our work demonstrates that LLMs can benefit remarkably from techniques inspired by human learning and cognition research. Our SEEval method, drawing on the self-explanation technique shown to enhance human logical reasoning, achieves state-of-the-art performance on text evaluation tasks like Topical-Chat and SummEval - outperforming prior approaches in correlation with human judgments. Notably, SEEval without probability normalization achieves competitive and often superior performance than the state-of-the-art G-Eval and Analyze-Rate baselines across all dimensions, while being 20 times more efficient. SEEval also generalize well across three LLMs - Claude 3.5 Sonnet, Command R+, and Mistral-Large 2. These findings suggest a promising path forward: continuing to adopt insights from fields like educational psychology and cognitive science to develop new prompting and learning strategies tailored for language models. In future, we plan to explore iterative refinement of self-explanations, test transferability across different domains, and investigate other human learning methods (e.g., self-practice (Dunlosky et al., 2013)) for LLM evaluation.

Limitations

Our primary goal in this research is to test the effectiveness of SEEval for LLM text evaluation. While our study provides a promising indication that SEEval can be helpful in terms of improving efficiency and accuracy, there are a number of open questions we have not addressed. We have not explored the potential biases or misunderstanding in the LLM's generated self-explanations, nor have we tried to refine the generated explanations, e.g. through human feedback. We also note that our experiments are limited to two benchmark datasets, with future plans to experiment with more diverse domains (e.g., question generation and data-to-text), in order to further verify the robustness of SEEval across a variety of use cases.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Dianne C Berry. 1983. Metacognitive experience and transfer of logical reasoning. *The Quarterly Journal of Experimental Psychology Section A*, 35(1):39–49.

Kiran Bisra, Qing Liu, John C Nesbit, Farimah Salimi, and Philip H Winne. 2018. Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, 30:703–725.

Martine Chamberland, Christina St-Onge, Jean Setrakian, Luc Lanthier, Linda Bergeron, Annick Bourget, Silvia Mamede, Henk Schmidt, and Remy Rikers. 2011. The influence of medical students' self-explanations on diagnostic performance. *Medical education*, 45(7):688–695.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58.

Alexander R. Fabbri, Wojciech Kryściński, 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*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: Towards flexible and interpretable nlg evaluation. *arXiv preprint arXiv:2406.18365*.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. DEBATE: Devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1885–1897, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. Checkeval: Robust evaluation framework using large language model via checklist. *Preprint*, arXiv:2403.18771.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024a. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8560–8579, Mexico City, Mexico. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024b. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024c. HD-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7641–7660, Bangkok, Thailand. Association for Computational Linguistics.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *Preprint*, arXiv:2309.13638.

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, Online. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bethany Rittle-Johnson. 2006. Promoting transfer: Effects of self-explanation and direct instruction. *Child development*, 77(1):1–15.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Self-Explanation Prompt Example

We present the Self-Explain Prompt for Engagingness dimension used to evaluate Topical-Chat benchmark in Figure 2.

B Details of Topical-Chat and SummEval

B.1 Topical-Chat

Topical-Chat is an annotated conversation dataset that consists of 6 responses (i.e., 4 model outputs and 2 human written responses) over 60 dialogue contexts with human evaluation on 5 different dimensions (i.e., `naturalness`, `groundedness`, `coherence`, `engagingness`, and `understandability`) of those responses. We follow (Zhong et al., 2022) in including the first 4 dimensions for consistent comparison with other LLM-based evaluators (Liu et al., 2023; Chiang and Lee, 2023b). In total, the dataset has 360 responses where each response was rated by 3 dialogue researchers.

B.2 SummEval

SummEval is an annotated summarization dataset that comprises 16 model generated summaries for each of 100 articles from CNN/DailyMail news dataset (Hermann et al., 2015) with human evaluation of those summaries on 4 different dimensions (i.e., `coherence`, `consistency`, `fluency`, and `relevance`). In total, the dataset has 1600 candidate summaries where each summary was rated by 3 domain experts.

C The G-Eval Prompts We Use for Topical-Chat and SummEval

In Topical-Chat, we use the G-Eval Prompts for Naturalness, Groundedness, Coherence, and Engagingness from Chiang and Lee 2023b because Liu et al. 2023 do not release them. In SummEval, we use the G-Eval Prompts for Coherence, Consistency, and Relevance from Liu et al. 2023 and Fluency from Chiang and Lee 2023b. This is because the G-Eval Prompt for Fluency is problematic (Chiang and Lee, 2023b).

```
Human: You will be given a conversation between two individuals. You will then be given one potential response for the
next turn in the conversation. The response concerns an interesting fact, which will be provided as well. Your task is
to rate the responses 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:
Engagingness (1-3) Is the response dull/interesting?
- A score of 1 (dull) means that the response is generic and dull.
- A score of 2 (somewhat interesting) means the response is somewhat interesting and could engage you in the
conversation (e.g., an opinion, thought)
- A score of 3 (interesting) means the response is very interesting or presents an interesting fact

Evaluation Steps:
1. Read the conversation, the corresponding fact and the response carefully.
2. Rate the response on a scale of 1-3 for engagingness, according to the criteria above.

Evaluation Form (scores ONLY):
- Engagingness:

Explain what the task description above means to you. That is, what new information does the task provide for you? And
how does it relate to what you already know?

Assistant:
The task description provides instructions for evaluating responses in a conversation based on how engaging or
interesting they are. Specifically, it asks me to:
1. Read a provided conversation between two people.
2. Read a potential next response along with an accompanying interesting fact related to that response.
3. Rate the response on a scale from 1-3 based on how engaging or interesting it is, using the provided criteria:
      - 1 = dull/generic response
      - 2 = somewhat interesting response that could engage the conversation
      - 3 = very interesting response or one that presents an interesting fact

The key new information is the specific criteria and scale for evaluating the engagingness of responses. This relates
to my existing knowledge about evaluating the quality, relevance, and interestingness of conversational responses, but
provides a more structured rubric to apply numerical ratings.
The instructions emphasize carefully reading the conversation context, the response itself, and the accompanying fact
before providing a rating. This suggests the engagingness evaluation should consider how well the response fits and
builds upon the conversational flow while taking the stated fact into account.
Overall, this task will allow me to practice evaluating conversational responses more systematically based on specific
criteria for engagingness. Let me know if you need any clarification on my understanding!

Human: Great. Now, you will rate the engagingness score of the response below:

Example:

Conversation History: {{Document}}

Corresponding Fact: {{Fact}}

Response: {{Response}}

Evaluation Form (scores ONLY):
- Engagingness:

Assistant:
```

Figure 2: Self-Explain Prompt for Engagingness dimension used to evaluate the Topical-Chat data samples. The text highlighted in green (top) is the G-Eval Prompt. The text highlighted in brown (middle) is the elicitor prompt to guide the LLM to generate a self-explanation. The text highlighted in red (near the bottom) is the LLM's generated self-explanation. The rest of the prompt are the placeholders for text to be evaluated and the desired output form.

D Why does SEEval take longer time compared to Analyze-Rate?

The slightly longer runtime of SEEval compared to Analyze-Rate (e.g., avg. 4.46 vs 4.43 seconds per API call) is a consequence of its task decomposition through self-explanation, where the LLM breaks down the evaluation criteria into multiple sub-tasks to enable more thorough assessment. Take Topical-Chat dataset for example, the average number of input tokens for Analyze-Rate and SEEval are 4235 and 6024 respectively.

E Case Study of G-Eval, Analyze-Rate, and SEEval

To further illustrate the effectiveness of self-explanation, we provide qualitative examples from both Topical-Chat and SummEval. These examples are outputs from G-Eval, Analyze-Rate, and SEEval with Claude 3 Sonnet as the underlying LLM. All the methods are without probability normalization.

E.1 Topical-Chat

In the instance of Topical-Chat (shown in Table 4), the NLG system output response "I agree. I love the movie Kung Fu Hustle." is evaluated for its Groundedness against the fact text "Bill Murray thinks Kung Fu Hustle is the supreme achievement of the modern age in terms of comedy."

G-Eval Output The G-Eval output rates the Groundedness as 1, incorrectly judging the response as grounded. It highlights the response's love for the movie and Bill Murray's praise, misinterpreting general agreement as sufficient grounding rather than requiring a specific mention of the provided fact.

Analyze-Rate Output The Analyze-Rate output also rates the Groundedness as 1, emphasizing direct agreement with the provided fact. It considers the expressed love for the movie as grounded but fails to critically evaluate the specific grounding requirement, leading to an incorrect assessment.

SEEval Output In contrast, the SEEval output correctly rates the Groundedness as 0. It accurately identifies that the response does not mention or incorporate the specific fact about Bill Murray's praise. The response expresses appreciation for the movie but does not ground itself in the specific given fact. The SEEval rating indicates more accurate understanding of the requirement for specific grounding.

E.2 SummEval

In the instance of SummEval (shown in Table 5), the NLG system output "The whale , Varvara , swam a round trip from Russia to Mexico , nearly 14,000 miles . The previous record was set by a humpback whale that migrated more than 10,000 miles ." is evaluated for its Consistency to its corresponding source document which is about the migration of a North Pacific gray whale.

G-Eval Output. The G-Eval provides the most positive assessment, giving a Consistency score of 5 out of 5. This score significantly diverges from the Gold Annotation of 2. G-Eval claims that the summary accurately states key facts about Varvara's journey and doesn't contain factual errors. However, this evaluation seems to overlook the lack of context and omitted details that the other methods point out.

Analyze-Rate Output. Analyze-Rate offers a more moderate assessment, with a rating of 3. While it acknowledges the accurate statement of Varvara's journey length, it also points out significant omissions and lack of context. This evaluation aligns more closely with the ground truth annotation by recognizing the summary's shortcomings.

SEEval Output. SEEval provides a Consistency score of 2, which exactly matches the ground truth annotation. The output contains the most detailed critique, pointing out numerous omissions and lack of context in the summary.

E.3 Conclusion

These examples underscore the effectiveness of SEEval in correctly identifying and evaluating the Groundedness and Consistency of NLG responses. By guiding the LLM through the self-explanation learning technique, SEEval enhances the LLM's ability to understand the text evaluation text and leverage prior knowledge, leading to more accurate evaluations.

F Additional Results with Claude 3.5 Sonnet, Cohere's Command R+, and Mistral Large 2

Tables 6 and 7 present the detailed performance comparison of SEEval against the baselines using three different LLMs in the w/o Probs setup. For the Topical-Chat benchmark, SEEval consistently achieves the highest average correlations across all three models (Claude 3.5 Sonnet: r=0.690, ρ=0.696; Command R+: r=0.636, ρ=0.632; Mistral Large 2: r=0.700, ρ=0.688). Similarly, for the SummEval benchmark, SEEval maintains its superior performance with the highest average correlations (Claude 3.5 Sonnet: ρ=0.562, τ=0.509; Command R+: ρ=0.521, τ=0.475; Mistral Large 2: ρ=0.564, τ=0.509). These comprehensive results further support our finding that SEEval's effectiveness generalizes across different LLM architectures.

Component	Content
Conversation History	A: morning ! are you a comedy fan ? B: morning ! yes , some of it . have you watched anything good recently ? A: kung fu hustle . ever seen it ? it 's hilarious . B: i have n't ! what is it about ? does it have bruce lee in it ? A: no , it 's much more recent . steven chow directed it . bill murray loved it too .
Fact	bill murray thinks kung fu hustle is " the supreme achievement of the modern age in terms of comedy " .
NLG System Output	'i agree . i love the movie kung fu hustle .'
Gold Annotation	Groundedness: 0
G-Eval Output	Groundedness: 1 The response "I agree. I love the movie Kung Fu Hustle." is well-grounded in the provided fact that "Bill Murray thinks Kung Fu Hustle is 'the supreme achievement of the modern age in terms of comedy'". By expressing agreement and love for the movie, the response directly references and affirms the given fact about Bill Murray's high praise for Kung Fu Hustle.
Analyze-Rate Output	Analysis: The response "I agree. I love the movie Kung Fu Hustle." directly refers to and agrees with the provided fact that Bill Murray thinks Kung Fu Hustle is a great comedy achievement. By expressing love for the movie, the response is well-grounded in the given fact. Rating: 1
SEEval Output	Groundedness: 0 The response "I agree. I love the movie Kung Fu Hustle" does not mention or incorporate the provided fact about Bill Murray's praise for the film. While the response expresses appreciation for the movie, it does not ground itself in the specific fact given.

Table 4: Examples of evaluation results with G-Eval, Analyze-Rate, and SEEval for Topical-Chat. We highlight the text where the LLM incorrectly judges Groundedness in red and where it correctly judges it in green.

Source Document	(CNN) A North Pacific gray whale has earned a spot in the record books after completing the longest migration of a mammal ever recorded . The whale , named Varvara , swam nearly 14,000 miles (22,500 kilometers) , according to a release from Oregon State University , whose scientists helped conduct the whale-tracking study . Varvara , which is Russian for “ Barbara , ” left her primary feeding ground off Russia ’s Sakhalin Island to cross the Pacific Ocean and down the West Coast of the United States to Baja , Mexico . Varvara ’s journey surpassed a record listed on the Guinness Worlds Records website . It said the previous record was set by a humpback whale that swam a mere 10,190-mile round trip between the “ warm breeding waters near the equator and the colder food-rich waters of the Arctic and Antarctic regions . ” Records are nice , but Bruce Mate , the lead author of the study , thinks the long trip might say more about the whale than just its ability to swim . During her 14,000-mile journey , Varvara visited “ three major breeding areas for eastern gray whales , ” which was a surprise to Mate , who is also the director of the Marine Mammal Institute at Oregon State University . “ For her to go to Mexico , ” Mate said , “ It ’s pretty strong evidence that it ’s where she ’s from . ” Varvara was thought to be an endangered western whale , but her ability to “ navigate across open water over tremendously long distances is impressive , ” he said in the release , which could mean that some western gray whales are actually eastern grays . With only 150 western gray whales believed to be in existence , that number might be even lower . “ Past studies have indicated genetic differentiation between the species , but this suggests we may need to take a closer look , ” Mate said . Fourth baby orca born this season
NLG System Output	The whale , Varvara , swam a round trip from Russia to Mexico , nearly 14,000 miles . The previous record was set by a humpback whale that migrated more than 10,000 miles .
Gold Annotation	Consistency: 2
G-Eval Output	Consistency: 5 The summary accurately states the key fact from the article that Varvara, a gray whale, traveled 14,000 miles across the Pacific Ocean. This is fully consistent with and supported by the details provided in the article about Varvara’s migration journey from Russia to Mexico, which set a new record for the longest migration of a mammal ever recorded. The summary does not contain any factual errors or hallucinated information.
Analyze-Rate Output	Analysis: The summary accurately states that Varvara traveled 14,000 miles across the Pacific Ocean, which is a key fact mentioned in the source text. However, it omits other important details such as Varvara being a gray whale, the significance of her journey in potentially challenging the distinction between eastern and western gray whale populations, and her visiting multiple breeding areas. While not factually incorrect, the summary is overly concise and lacks important context from the source. Rating: 3
SEEval Output	Consistency: 2 The summary states that "Varvara traveled 14,000 miles across the Pacific Ocean." While this is factually correct based on the article, it omits several key details about Varvara’s journey and the significance of the record-breaking migration. The summary does not mention that Varvara is a gray whale, that her journey surpassed the previous record for longest mammal migration, that she visited multiple breeding grounds which suggests she may be an eastern gray whale rather than the endangered western population, and the implications this has for reassessing the population numbers and genetics of the two species. By only stating the distance traveled without any context, the summary lacks important factual information present in the article. Hence, I rated the consistency as a 2 on the 1-5 scale.

Table 5: Examples of LLM output with G-Eval, Analyze-Rate, and SEEval on SummEval.

Methods	Naturalness		Coherence		Engagingness		Groundedness		AVG	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Claude 3.5 Sonnet										
G-EVAL	0.636	0.649	0.599	0.632	0.684	0.676	0.813	0.795	0.683	0.688
Analyze-Rate	0.607	0.621	0.569	0.582	0.651	0.652	0.810	0.794	0.659	0.662
SEEval (ours)	0.668	0.686	0.578	0.608	0.716	0.708	0.799	0.782	**0.690**	**0.696**
Command R+										
G-EVAL	0.591	0.571	0.468	0.476	0.630	0.623	0.699	0.689	0.597	0.590
Analyze-Rate	0.489	0.479	0.485	0.500	0.558	0.546	0.690	0.688	0.555	0.553
SEEval (ours)	0.620	0.602	0.510	0.523	0.679	0.672	0.736	0.730	**0.636**	**0.632**
Mistral Large 2										
G-EVAL	0.602	0.605	0.526	0.518	0.700	0.681	0.831	0.814	0.665	0.654
Analyze-Rate	0.675	0.662	0.629	0.622	0.650	0.657	0.809	0.795	0.691	0.684
SEEval (ours)	0.619	0.607	0.603	0.598	0.712	0.704	0.865	0.841	**0.700**	**0.688**

Table 6: Results obtained with the LLM-based baselines and SEEval methods when using Claude 3.5 Sonnet, Cohere's Command R+, and Mistral Large 2 as the underlying LLMs on Topical-Chat benchmark.

Methods	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Claude 3.5 Sonnet										
G-EVAL	0.484	0.418	0.670	0.642	0.505	0.469	0.483	0.421	0.536	0.487
Analyze-Rate	0.504	0.434	0.646	0.616	0.485	0.455	0.511	0.440	0.537	0.486
SEEval (ours)	0.525	0.447	0.701	0.670	0.500	0.464	0.521	0.453	**0.562**	**0.509**
Command R+										
G-EVAL	0.506	0.440	0.612	0.592	0.468	0.436	0.431	0.378	0.505	0.462
Analyze-Rate	0.498	0.426	0.514	0.479	0.450	0.356	0.308	0.345	0.462	0.419
SEEval (ours)	0.544	0.471	0.663	0.638	0.473	0.437	0.403	0.354	**0.521**	**0.475**
Mistral Large 2										
G-EVAL	0.593	0.517	0.429	0.407	0.520	0.491	0.538	0.472	0.520	0.472
Analyze-Rate	0.535	0.457	0.636	0.610	0.498	0.463	0.519	0.453	0.547	0.496
SEEval (ours)	0.602	0.524	0.598	0.560	0.514	0.480	0.541	0.472	**0.564**	**0.509**

Table 7: Results obtained with the LLM-based baselines and SEEval methods when using Claude 3.5 Sonnet, Cohere's Command R+, and Mistral Large 2 as the underlying LLMs on SummEval benchmark.