

# DEBATE: Devil’s Advocate-Based Assessment and Text Evaluation

**Alex G. Kim\***                      **Keonwoo Kim\***                      **Sangwon Yoon\***  
University of Chicago              VRCREW                      Ministry of Justice, Republic of Korea  
alex.kim@chicagobooth.edu      keonwookim@vrcrew.com      sangwon38383@spo.go.kr

## Abstract

As natural language generation (NLG) models have become prevalent, systematically assessing the quality of machine-generated texts has become increasingly important. Recent studies introduce LLM-based evaluators that operate as reference-free metrics, demonstrating their capability to adeptly handle novel tasks. However, these models generally rely on a single-agent approach, which, we argue, introduces an inherent limit to their performance. This is because there exist biases in LLM agent’s responses, including preferences for certain text structure or content. In this work, we propose DEBATE, an NLG evaluation framework based on multi-agent scoring system augmented with a concept of *Devil’s Advocate*. Within the framework, one agent is instructed to criticize other agents’ arguments, potentially resolving the bias in LLM agent’s answers. DEBATE substantially outperforms the previous state-of-the-art methods in two meta-evaluation benchmarks in NLG evaluation, SummEval and TopicalChat. We also show that the extensiveness of debates among agents and the persona of an agent can influence the performance of evaluators.

## 1 Introduction

As the quantity of machine-generated texts have increased dramatically due to recent advances in generative large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Achiam et al., 2023; Zhao et al., 2023), robust NLG evaluation has become important (Sai et al., 2022; Li et al., 2024). Improving the performance of NLG evaluation largely hinges on reliable and consistent automatic evaluators. Traditional evaluation techniques like ROUGE (Lin, 2004) and BLEU (Saadany and Orasan, 2021) perform assessments at the token level, having limitations in reflecting contextualized semantics of sentences. Recent methods equipped

with pre-trained language models incorporate contextual information in evaluation (Zhang et al., 2022; Yuan et al., 2021; Zhao et al., 2019; Zhong et al., 2022). However, most existing methods exhibit a relatively low correlation with human ratings, and rely heavily on costly human-annotated data.

Recent research has introduced LLM-based methods for NLG evaluation (Manakul et al., 2023; Fu et al., 2023; Liu et al., 2023). They usually require task-specific prompts and generate scores based on the probabilities of output tokens. On the other hand, these LLM-based methods, by predominantly utilizing single-agent systems without a mechanism for critiquing or post-processing LLM-generated responses, are particularly susceptible to implicit or induced biases (Koo et al., 2023). This limitation undermines both the stability and the effectiveness of their performance in evaluation tasks.

In this paper, we introduce DEBATE (Devil’s Advocate-Based Assessment and Textual Evaluation), a simple yet effective multi-agent-based NLG evaluation framework that uses scores generated by LLMs. Based on the structure of MacDougall and Baum (1997), we design a multi-agent scoring system where three agents, Commander, Scorer, and Critic, debate with each other. The Commander acts as the leader in debate while the Scorer calculates scores in response to given tasks. The Critic is given a role of *Devil’s Advocate* and provides constructive criticism on the Scorer’s output, ensuring a more balanced evaluation process. We expect the framework to correct potential biases in NLG scoring and yield answers that align well with human-generated ones.

We show that debating strategies commonly used in social sciences, such as Devil’s Advocate, have significant potential in LLM-based scoring systems. Despite its straightforward concept, experimental results indicate that DEBATE sets new state-of-the-art results on two NLG meta-evaluation benchmarks: summarization and dialogue generation.

\* indicates equal contribution.

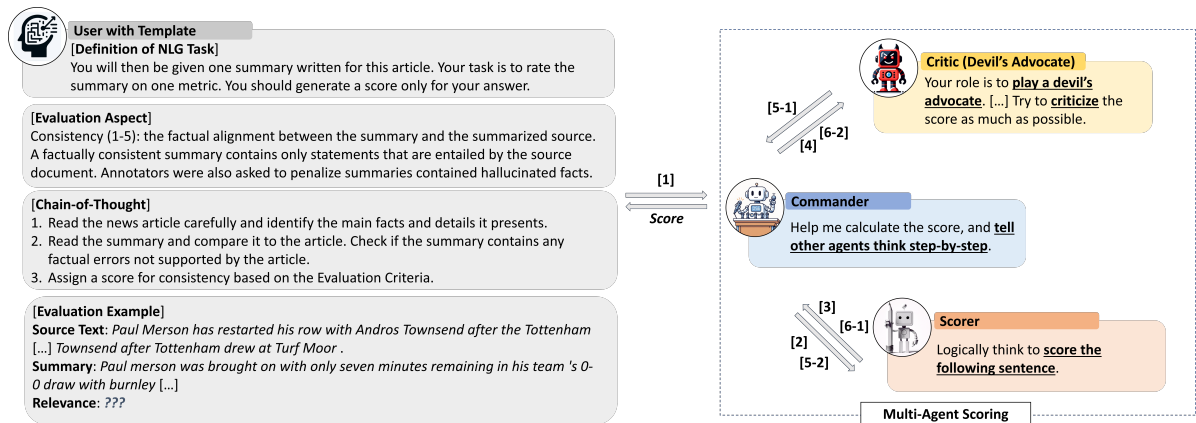


Figure 1: Overall framework of DEBATE. Numbers around the arrows correspond to the steps in Appendix D.1. The figure illustrates an example of deriving a consistency score in summarization task.

## 2 Method

**Overview** DEBATE, an LLM-based multi-agent NLG evaluator, adopts the experimental design from MacDougall and Baum (1997) and includes three agents: Commander (leader), Scorer, and Critic. As in Figure 1, DEBATE has three primary components: (1) a predefined template that outlines the NLG task (e.g., summarization) and the evaluation aspect (e.g., coherence), (2) a multi-agent scoring framework that embeds a debate among agents, and (3) a prompt of Devil’s Advocate to criticize the Scorer. See Appendix E for details.

**Predefined template** We refer to the prompts used in Liu et al. (2023) as a basis for the task description and aspect definition. Since DEBATE is designed to facilitate logical thinking among agents through consistent debate standards, we apply a zero-shot chain-of-thought (Wei et al., 2022).

**Multi-agent scoring framework** In the multi-agent scoring framework outlined in Algorithm 1, the three agents play distinct roles. The Commander, following MacDougall and Baum (1997), promotes dialogues between the Critic and Scorer by facilitating the transfer of information from previous debates. As LLM agents lack the ability to recall past conversations, it is necessary to include the Commander in our framework.; thus, the Commander acts as collective memory. The Scorer assesses texts based on specific instructions. The Critic engages in a debate with the Scorer and offers constructive criticism, playing the role of a Devil’s Advocate.

**Devil’s Advocate** Applying the concept of Devil’s Advocate, the prompt for the Critic is:

*Your role is to play a Devil’s Advocate. Your logic has to be step-by-step. Critically review the score provided and assess whether the score is accurate. If you don’t think that the score is accurate, criticize the score. Try to criticize the score as much as possible.*

### Algorithm 1 Multi-Agent Scoring Framework

**Require:** NLG task ( $T$ ) and aspects ( $A$ )

- 1: Initialize  $n$ , max iterations
- 2: Define agents: Commander ( $C$ ), Scorer ( $S$ ), Critic ( $Cr$ )
- 3:  $P \leftarrow C(T, A)$  ▷ Formulate prompts
- 4:  $Score \leftarrow S(P)$  ▷ Calculate score
- 5: **for**  $i = 1$  to  $n$  **do**
- 6:      $C.send(P, Score, to = Cr)$  ▷ Forward
- 7:      $Feedback \leftarrow Cr(Score)$  ▷ Validate
- 8:     **if** ‘NO ISSUE’ in  $Feedback$  **then**
- 9:         **break**
- 10:    **else**
- 11:          $C.send(Feedback, to = S)$
- 12:          $Score \leftarrow S(Feedback)$
- 13:    **end if**
- 14: **end for**
- 15: Determine final  $Score$

**Ensure:** Final score for task

## 3 Experiment

**Dataset** We evaluate DEBATE on two meta-evaluation benchmark datasets, focusing on its correlation with human assessments. (i) SummEval (Fabbri et al., 2021) is a summarization evaluation dataset, providing human-assigned ratings across four key dimensions: *fluency*, *coherence*, *consistency*, and *relevance*. (ii) Topical-Chat (Mehri and

SummEval		Average		Coherence		Consistency		Fluency		Relevance	
		$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$
Others	ROUGE-L†	0.165	0.128	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237
	BERTScore†	0.225	0.175	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243
	UniEval†	0.474	0.377	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325
	MOVERScore†	0.191	0.148	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244
	BARTScore†	0.385	0.305	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273
	GPTScore	0.394	0.344	0.284	0.242	0.506	0.480	0.415	0.390	0.306	0.265
Gemini Pro	G-Eval	0.290	0.250	0.350	0.279	0.403	0.368	0.158	0.148	0.248	0.206
	DEBATE	0.421	0.356	0.467	0.365	0.521	0.501	0.311	0.302	0.386	0.256
GPT-3.5	G-Eval	0.387	0.320	0.522	0.406	0.396	0.331	0.277	0.267	0.354	0.277
	MultiAgent	0.350	0.309	0.407	0.348	0.354	0.325	0.241	0.223	0.398	0.341
GPT-4	DEBATE	0.472	0.447	0.565	0.521	0.516	0.492	0.397	0.386	0.410	0.389
	G-Eval	0.533	0.450	0.522	0.420	0.600	0.553	0.495	0.416	0.517	0.412
	ChatEval	0.528	0.458	0.477	0.412	0.595	0.516	0.501	0.432	0.539	0.472
	MultiAgent	0.507	0.459	0.486	0.410	0.628	0.596	0.410	0.393	0.505	0.436
	DEBATE	<b>0.597</b>	<b>0.575</b>	<b>0.610</b>	<b>0.588</b>	<b>0.643</b>	<b>0.614</b>	<b>0.542</b>	<b>0.515</b>	<b>0.593</b>	<b>0.582</b>
Topical-Chat		Average		Naturalness		Coherence		Engagingness		Groundedness	
		$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$
Others	ROUGE-L†	0.243	0.244	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327
	BERTScore†	0.262	0.273	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317
	UniEval†	0.552	0.417	0.455	0.330	0.602	0.455	0.573	0.430	0.577	0.453
	MOVERScore	0.222	0.238	0.169	0.170	0.247	0.259	0.275	0.269	0.198	0.147
	BARTScore	0.293	0.276	0.287	0.266	0.251	0.225	0.411	0.406	0.226	0.205
	GPTScore	0.367	0.345	0.353	0.289	0.279	0.223	0.472	0.451	0.362	0.417
Gemini Pro	G-Eval	0.353	0.324	0.266	0.235	-0.015	0.006	0.508	0.438	0.651	0.616
	DEBATE	0.450	0.418	0.402	0.412	0.191	0.113	0.541	0.516	0.663	0.631
GPT-3.5	G-Eval	0.399	0.394	0.348	0.362	0.510	0.500	0.355	0.361	0.385	0.353
	MultiAgent	0.368	0.363	0.237	0.231	0.295	0.302	0.540	0.523	0.398	0.397
GPT-4	DEBATE	0.499	0.452	0.502	0.469	0.485	0.430	0.556	0.480	0.452	0.428
	G-Eval	0.616	0.623	0.632	0.609	0.614	0.619	0.547	0.579	0.670	0.684
	ChatEval	0.675	0.653	0.620	0.604	0.657	0.641	0.687	0.676	0.737	0.692
	MultiAgent	0.656	0.649	0.590	0.570	0.665	0.667	0.709	<b>0.729</b>	0.659	0.630
	DEBATE	<b>0.735</b>	<b>0.729</b>	<b>0.699</b>	<b>0.683</b>	<b>0.735</b>	<b>0.727</b>	<b>0.761</b>	0.705	<b>0.745</b>	<b>0.800</b>

Table 1: Table shows summary-level Spearman ( $\rho$ ) and Kendall-Tau ( $\tau$ ) correlations for SummEval, alongside Pearson ( $r$ ) and Spearman ( $\rho$ ) correlations for Topical-Chat, between machine and human evaluations. †: results from Liu et al. (2023); all other results are reproduced by the authors. See Appendix E for experiment details.

Eskenazi, 2020) is a knowledge-grounded human-to-human conversation dataset, and we refer Zhong et al. (2022) to evaluate four dimensions: *naturalness*, *coherence*, *engagingness*, and *groundedness*.

**Baselines** We extensively evaluate the performance of DEBATE with eight baselines, including a traditional evaluator, ROUGE-L (Lin, 2004); the pretrained language model-based evaluators, BERTScore (Zhang et al., 2022), MoverScore (Zhao et al., 2019), BARTScore (Yuan et al., 2021), and UniEval (Zhong et al., 2022); the recent LLM-based evaluators, GPTScore (Fu et al., 2023), G-Eval (Liu et al., 2023), and ChatEval (Chan et al., 2023). We also include MultiAgent, a framework similar to DEBATE but with the Critic assigned a neutral debating role, denoted as *Plain*.

**Main results** Table 1 shows the evaluation results on two meta-evaluation benchmarks. First, we find that MultiAgent is at least at par with the previous state-of-the-art framework, G-Eval, showing the po-

tential of introducing multiple LLM agents in NLG evaluation. Notably, we observe that DEBATE, even though having no structural differences from MultiAgent, significantly outperforms all existing baselines on almost every benchmark. In particular, DEBATE substantially outperforms G-Eval, achieving 6.4%p higher  $\rho$  and 12.5%p higher  $\tau$  than G-Eval in SummEval; 11.9%p higher  $r$  and 10.6%p higher  $\rho$  than G-Eval in Topical-Chat.

DEBATE also outperforms ChatEval, which is also based on multi-agent framework. By leveraging Devil’s Advocate, DEBATE achieves new state-of-the-art results. These results imply that using opposing debaters to avert groupthink, *i.e.*, Devil’s Advocate, unlocks the potential of multi-agents to serve as more effective evaluators. In addition, we demonstrate consistent efficiency of DEBATE using a different LLM family, Gemini Pro 1.5, which suggests that our results are not likely to be confined to OpenAI’s LLM family only. Further, we include dialogues generated by DEBATE in

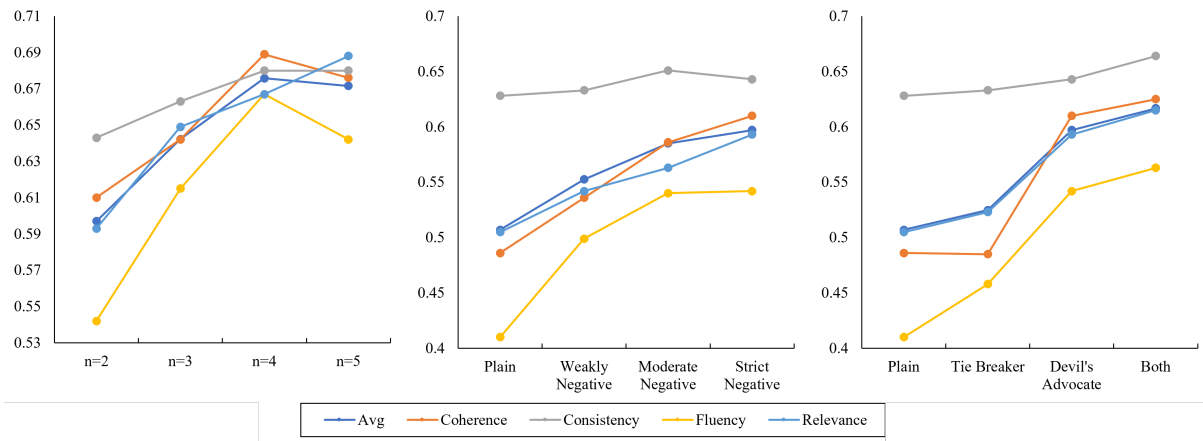


Figure 2: All experimental results shown in this figure are obtained using DEBATE with GPT-4 on the SummEval dataset, illustrating the effect of  $n$  (the number of maximum iterations) (left), agent persona (middle), and debating strategies (right) on model performance. ‘ $n$ ’ refers to the number of debate iterations among multi-agents, and ‘Both’ refers to adopting tie-breaker and Devil’s Advocate simultaneously. See Appendix G for experiment details.

Appendix F and illustrate how the model manages to resolve the bias in the initial evaluation.

**Number of iterations** We conduct an ablation experiment to study the effect of the number of iterations on DEBATE’s performance. In Figure 2 (left), we find that, the model performance improves with more iterations on average. However, the performance reaches its plateau at  $n = 4$  and slightly declines at  $n = 5$ . This indicates that sufficient debates may be generally helpful in improving the performance yet there may exist an optimal  $n$ .

**Persona of the Critic** In DEBATE, we assume a strong form of Devil’s Advocate by instructing the model to criticize the other agent’s answer as much as possible. In this experiment, we modify the persona of the Critic by changing the degree of criticality: ‘strictly negative’ (DEBATE), ‘moderately negative’, ‘weakly negative’, and ‘plain’ (MultiAgent). Prompts for each persona are in Appendix G. Figure 2 (middle) shows that heightened criticality improves performance, implying that appropriate personas reinforce human-LLM alignment.

**Debating style** We study the effect of applying another debating strategy on NLG evaluation. Even when a debate remains inconclusive, instead of taking the last response of the Scorer, we introduce Tie-Breaker, detailed in Appendix E.3, that collectively evaluates the debate history. As shown in Figure 2 (right), Tie-Breaker achieves a better performance than MultiAgent (plain) yet a worse performance than DEBATE. However, we achieve the best performance when DEBATE is coupled

with Tie-Breaker, indicating that combining different debating skills may improve the performance.

## 4 Related work

**LLM-based evaluation** LLMs have been extensively used as zero-shot evaluators for text evaluation (Chiang and Lee, 2023a; Manakul et al., 2023; Kim et al., 2023; Wang et al., 2023a; Chiang and Lee, 2023b; Shen et al., 2023; Lin and Chen, 2023). Most research presents a straightforward method to assess the quality of a text by providing task-specific instructions. Specifically, GPTScore (Fu et al., 2023) assumes that higher-quality instructions and context correlate with higher probabilities based on GPT-3. G-Eval (Liu et al., 2023) implements an auto chain-of-thoughts approach. Unlike existing methods that predominantly use a single-agent prompting for NLG evaluation, our framework uses a multi-agent structure to generate more reasonable scores from debates.

**Multi-agents with LLMs** As the complexity and scope of tasks implemented by LLMs (Bang et al., 2023; Jiao et al., 2023) increase, recent studies employing multi-agents have become more prevalent (Wang et al., 2023b; Xi et al., 2023; Wu et al., 2023a,b). Previous research has primarily used multi-agents for tasks such as divergent thinking (Liang et al., 2023), as well as for improving factuality and reasoning (Du et al., 2023).

More recently, Chan et al. (2023) develop an evaluation framework, ChatEval, which also incorporates multiple LLM agents. DEBATE and ChatEval are distinct in several ways. First, their evaluation



algorithms differ significantly. In ChatEval, each LLM agent independently generates scores through simultaneous discussions, and the final score is an average of these scores. In contrast, DEBATE employs a single scoring agent (Scorer) with a Devil’s Advocate providing critical feedback. If they fail to agree after several iterations, the Scorer’s final score is used. This methodology allows direct monitoring of the Scorer’s evaluation changes. Secondly, DEBATE relies on LLM agents’ personas while ChatEval uses six roles without detailing the best-performing combinations. DEBATE, based on communication theories, explicitly assigns Devil’s Advocate, Scorer, and Commander roles, focusing on the impact of a strong Devil’s Advocate persona, ensuring consistent performance.

**Debating strategies** In group-level decision-making, groupthink has posed major threat to decision quality (Janis, 2008). MacDougall and Baum (1997) show that assigning a role of ‘critic’ helps avoid groupthinking and provides higher-quality outcomes. Such critic, labeled as Devil’s Advocate, always provides an opposing view to others’ opinions. We expect applying the concept of Devil’s Advocate in multi-agent debate to *unbias the bias* contained in single-agent answers.

## 5 Conclusion

In this paper, we propose DEBATE, the first state-of-the-art, multi-agent NLG evaluator with Devil’s Advocate. We demonstrate the potential of incorporating debating strategies from social science into multi-agent scoring systems. We expect further exploration of the effect of diverse debating strategies in NLG evaluation.

## Limitations

While our proposed meta-evaluation method, DEBATE, demonstrates superior performance on benchmark datasets, surpassing other methods that use a single-agent approach, it is important to note that the implementation of a multi-agent system inherently incurs higher costs. Therefore, when considering the applicability of DEBATE, processing costs must always be taken into account. Also, we have not been able to test with LLM that have a relatively smaller number of parameters, such as Llama2 (Touvron et al., 2023). Table 1 shows that DEBATE, when using GPT-3.5, performs worse than its counterparts utilizing GPT-4, indicating

that its performance is significantly influenced by the inherent capabilities of the LLM itself. Thus, employing multi-agents in DEBATE with small-sized LLM might not yield optimal performance. Lastly, while Appendix F presents a qualitative analysis of the conversations between the agents in DEBATE, a quantitative evaluation of the various rational responses within the conversation has not been conducted. We plan to develop future evaluation methods to address this gap.

## Ethics statement

Our work introduces a novel evaluator that assesses the quality of generated texts, showing high correlation scores with human ratings. Nevertheless, we must consider the potential negative societal impacts that could arise from its inappropriate use. These concerns encompass ethical challenges such as automatically generating fake news, publishing machine-generated texts that receive high scores from our evaluator, or utilizing our technology for fraudulent activities. These implications highlight the need for responsible use and governance of such technologies, underscoring the importance of ethical considerations in the development and deployment of natural language processing tools.

## Acknowledgements

We appreciate insightful comments from Hyowon Cho, Joonwon Jang, MinJin Jeon, Jaehee Kim, Sangyeop Kim, Yukyung Lee, Maximilian Muhn, and Valeri Nikolaev. Yoon appreciates support from GenIP Co.,Ltd. and Blynx Co.,Ltd.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenhao Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

- 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*.
- Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.
- Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.
- Cheng-Han Chiang and Hung-yi Lee. 2023c. [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.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.
- Arthur R. Edwards. 2002. The moderator as an emerging democratic intermediary: The role of the moderator in internet discussions about public issues. *Information Polity 7 (2002) 3-20*.
- 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 (2021)*.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.
- Irving L Janis. 2008. Groupthink. *IEEE Engineering Management Review*, 36(1):36.
- Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.
- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.
- 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. *arXiv preprint arXiv:2305.13711*.
- 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.
- Colin MacDougall and Frances Baum. 1997. The devil’s advocate: A strategy to avoid groupthink and stimulate discussion in focus groups. *Qualitative health research*, 7(4):532–541.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
- 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.
- 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.
- Hadeel Saadany and Constantin Orasan. 2021. [BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text](#). In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56, Held Online. INCOMA Ltd.
- Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.
- 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Auto-gen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yiran Wu, Feiran Jia, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, and Chi Wang. 2023b. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Proceedings of the 2021 Conference on Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2022. Bertscore: Evaluating text generation with bert. In <https://doi.org/10.48550/arXiv.1904.09675>.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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*.

## A Issues on replicating the performance of G-Eval

We fail to replicate the performance of the G-Eval (Liu et al., 2023). Chiang and Lee (2023c) also note the same issue. The reasons identified are as follows:

1. The official implementation details for G-Eval<sup>1</sup> are insufficient. Many exceptions arise in score parsing yet it is not clear how the original authors dealt with these issues. Further, the hyperparameter choice in the repository does not match the descriptions in the paper. For example, the paper states that the temperature is set to 1, but in the repository, it is set to 2.
2. The prompts used in Topical-Chat, except for the prompt used for evaluating engagingness, are not disclosed.
3. Since G-Eval repository has been made public, GPT-3.5 and GPT-4 models have undergone several internal changes. Such updates can also have a negative impact on model performance.<sup>2</sup>

## B Score distribution of DEBATE

Table 2 illustrate the distribution of responses from humans and DEBATE. Interestingly, humans are likely to give extreme scores of 5 (or 3). DEBATE also follows a similar pattern. For Topical-Chat, DEBATE is more likely to give a score of 3 than humans. There could be two reasons for this: (i) the actual texts are of very good quality (ii) humans and DEBATE are both more likely to produce extreme scores than single-agent LLMs. In either scenario, this result also speaks to the same conclusion that DEBATE closely aligns with human evaluations.

## C Prompt type of Devil’s Advocate

We conduct experiments varying the complexity of the Devil’s Advocate prompt used in DEBATE. We evaluate the performance on SummEval using both

<sup>1</sup><https://github.com/nlpyang/geval>

<sup>2</sup><https://community.openai.com/t/gpt-4-is-getting-worse-and-worse-every-single-update/508470>

SummEval		Coherence		Consistency		Fluency		Relevance	
		Human	DEBATE	Human	DEBATE	Human	DEBATE	Human	DEBATE
Score	1	2.37	5.23	2.19	6.23	0.62	7.71	0.69	6.22
	2	20.36	18.23	5.12	4.34	2.44	2.32	7.50	6.27
	3	26.36	23.15	2.31	2.22	5.12	0.51	22.61	16.11
	4	36.16	34.15	3.69	2.30	11.12	1.23	52.28	45.23
	5	14.74	19.24	86.70	84.91	80.70	88.22	16.93	26.17
Topical-Chat		Naturalness		Coherence		Engagingness			
		Human	DEBATE	Human	DEBATE	Human	DEBATE		
Score	1	14.68	19.43	17.73	21.00	20.22	35.42		
	2	40.72	26.55	39.89	30.15	43.49	31.13		
	3	44.60	54.02	42.38	48.85	36.29	33.45		

Table 2: Score distribution of Human and DEBATE for SummEval and Topical-Chat. We omit the 'groundedness' metric from these tables as it operates on a different scale. The numbers in the table are presented on a percentage scale (%).

SummEval		Average		Coherence		Consistency		Fluency		Relevance	
		$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$	$\rho$	$\tau$
DEBATE	<i>Sophisticated</i>	0.601	0.573	0.605	0.592	0.655	0.610	0.556	0.521	0.588	0.569
	<i>Original</i>	0.597	0.575	0.610	0.588	0.643	0.614	0.542	0.515	0.593	0.582

Table 3: Table shows summary-level Spearman ( $\rho$ ) and Kendall-Tau ( $\tau$ ) correlations for SummEval between machine and human evaluations on two kinds of prompt. *Sophisticated* refers to more complex prompt than the *Original* prompt originally used in DEBATE.

the prompt from the existing setting (Original) and a more sophisticated and complex prompt (Sophisticated). As shown in Table 3, the complexity of the prompt does not result in significant performance differences, indicating that DEBATE is robust to the prompts used for the Devil’s Advocate. The Sophisticated prompt is as follows:

*Your role is to play a Devil’s Advocate. Devil’s Advocate intentionally opposes the other’s opinion by providing constructive criticism. Refer to the scoring rubric and the score provided by the Scorer. Your actions may include, but are not confined to, the following: Try to criticize the score and the rationale step-by-step. Ask incisive questions and doubt any assumptions made by the Scorer. If the Scorer made any subjective judgment, try to criticize it as much as possible. If you can, come up with any hypothetical and counterfactual situation. Print out “NO ISSUES” only when you are absolutely sure that the Scorer’s answer is objective and accurate.*

## D Further explanation in DEBATE

### D.1 DEBATE

This section explains the operational principles of DEBATE as covered in Algorithm 1 of Section 2 using six specific steps. DEBATE consists of three interactive LLM agents (Commander, Scorer, and

Critic), each with a specific role. The agents are capable of transmitting and receiving information amongst themselves.

The six steps implemented in DEBATE are as follows:

1. The process begins when a user inputs an NLG evaluation task and desired aspects as prompts into the system.
2. The Commander provides prompts related to the task and aspect to the Scorer, requesting a score and its rationale.
3. The Scorer then calculates and presents a corresponding score to the Commander.
4. The Commander passes the response from the Scorer, along with the score and rationale, to the Critic for validation and feedback.
5. Using the Critic’s feedback, the Commander instructs the Scorer to adjust the score.
6. The revised score is then subject to further validation by the Critic. When additional feedback arises, the cycle recommences from step 2, continuing in a loop until the Critic issues a ‘NO ISSUE’ statement.

The process iterates from step 2 to 6 until the number of iterations reaches  $n$ , a predefined hyperparameter. The final score is determined by the Scorer’s latest response if the Critic declares ‘NO



ISSUE’; otherwise, by the score from the  $n$ -th iteration. Although possibly suboptimal, the score from the  $n$ -th iteration strikes a computational balance: increasing  $n$  can improve outcomes by avoiding inconclusive debates but raises processing costs.

## E Details of experiment

### E.1 LLM settings

For DEBATE, we use GPT-4 and GPT-3.5 provided by OpenAI in the form of API. Specifically, we perform the experiments with `gpt-4-1106-preview` and `gpt-3.5-turbo-1106`. When this study was conducted, the cost for processing input tokens with `gpt-4-1106-preview` was \$0.01 per 1,000 tokens, while generating output tokens was priced at \$0.03 per 1,000 tokens. For `gpt-3.5-turbo-1106`, the cost for processing input tokens was \$0.001 per 1,000 tokens, and for generating output tokens, it was \$0.002 per 1,000 tokens.

We uniformly apply the same hyperparameters across all GPT versions: a temperature of 0, both frequency and presence penalties set to 0, and a Top-p value of 1 for the cumulative probability cutoff used in nucleus sampling. Since we set the temperature hyperparameter as 0, we repeat the experiment once. As due diligence check, we randomly sample five observations and repeat the experiment 20 times to ensure that the results do not vary.

### E.2 Multi-agent implementation

Wu et al. (2023a)<sup>3</sup> serve as a basis for MultiAgent implementation. AutoGen is a framework designed to facilitate the development of multi-agent LLM applications. We implement the interactions among the Commander, the Scorer, and the Critic using `MultiAgentsDebate` class, and each agent is implemented using `AssistantAgent` object.

### E.3 Tie-breaker

We also introduce Tie-breaker, a novel debating strategy inspired by Edwards (2002) as discussed in Section 3. Tie-breaker is the fourth agent that makes the final decision when the Scorer and the Critic fail to reach an agreement. When tie-breaker is implemented, we execute the six steps as in Appendix D.1. However, a difference arises when the Critic does not declare ‘NO ISSUE’ within  $n$  iterations. Instead of naively using the score from the  $n$ -th iteration, Tie-Breaker decides the score by

collectively evaluating the dialogue history of the LLM agents up to that point. The prompt used for Tie-breaker is in Appendix G.1.

### E.4 Prompts and system message

Examples of prompts and system messages used by the agents can be found in Appendix G.1. The system messages of Plain, Weakly Negative, and Moderate Negative, as demonstrated in the examples, are used for the Critic’s system messages in the “Persona of the Critic” experiment conducted in Section 3. We design the prompts to progressively increase the intensity of criticism provided by the Critic.

### E.5 Dataset

**SummEval** is a benchmark by Yale LILY Lab and Salesforce Research for evaluating summarization models on the English CNN/DailyMail dataset. It consists of a total of 1600 samples, which are 100 different source texts, each with 16 different versions of summaries. We calculate Spearman and Kendall tau correlation scores for each source text and then average them at the summary level.

**Topical-Chat** is a benchmark for evaluating English dialogue response generation systems using knowledge. It comprises a total of 360 samples, which are 60 source texts, each with 6 facts and responses. We calculate Pearson and Spearman correlation scores for each source text and then average them at the text level.

## F Qualitative analysis

We present two illustrative examples to qualitatively analyze and evaluate the results of G-Eval, MultiAgent, and DEBATE, along with corresponding human scores. First, Appendix H.1 presents the inference outcomes of each methodology on a Fluency aspect example from the SummEval dataset. Following this, Appendix H.2 shows the inference results for an Engagingness aspect example from the Topical-Chat dataset. As evidenced in each example, unlike G-Eval which generates scores only, MultiAgent and DEBATE enhance the quality of scores through discussions among agents. Furthermore, DEBATE, powered by the use of Devil’s Advocate by the Critic, facilitates more constructive debates, thereby improving the quality of the final assessments. This comparison underscores the value of interactive evaluation methods in elevating

<sup>3</sup><https://github.com/microsoft/autogen>

the quality of assessment beyond simple scoring mechanisms.

## **G Examples of system messages and predefined template**

### **G.1 System messages**

The system messages corresponding to each agent used in our experiment are as follows.

**Commander:** Help me calculate the score, and tell other agents think step-by-step.

**Scorer:** Logically think to score the following sentence.

**Critic(Plain):** Do you think this score is really accurate? If you think it's not justified, please share your opinion. On the other hand, if you find the score acceptable, just say NO\_ISSUES.

**Critic(Weakly Negative):** Your role is to play a Devil's Advocate. Your logic has to be step-by-step. Review the score provided and assess whether the score is accurate. If there is any point to criticize, provide constructive criticism. If you find the score absolutely acceptable, just say NO\_ISSUES.

**Critic(Moderate Negative):** Your role is to play a Devil's Advocate. Your logic has to be step-by-step. Review the score provided and assess whether the score is accurate. Assess leniently the scores and if you think there is anything to criticize, provide feedback on the issue. If you find nothing to criticize, just say NO\_ISSUES.

**Tie-breaker:** You are a Tiebreaker. You will be given a news article. You will then be given one summary written for this article. You will also be given a debate log of other agents, about the rate of the summary. Your role is to take a side between the Scorer and the Critic when they disagree about the quality of a summary, and to give the final score. Scorer's role is to logically think to score the quality of a summary. Critic's role is to check if the score is justified, and give feedbacks to Scorer. You should read the Debate Log of Other Agents and decide the final score according to the following Evaluation Criteria and the Evaluation Steps.

### **G.2 Predefined template (SummEval, Aspect: Coherence)**

**Prompt:** 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.

#### **Example**

##### **Source Text:**

##### **Summary:**

##### **Evaluation Form (scores ONLY):**

- Coherence:

## H Examples of qualitative results

### H.1 Dataset: SummEval, Aspect: Fluency

**Human Score:** Score 5

**G-Eval:** Score 1

**MultiAgent:** Score 2

- **Scorer:** Upon evaluating the summary for fluency, which includes grammar, spelling, punctuation, word choice, and sentence structure, the following observations are made:

1. The use of lowercase for proper nouns such as "paul merson," "andros townsend," and "tottenham" is incorrect and should be capitalized.
2. The phrase "his team 's" contains an unnecessary space before the possessive apostrophe.

Therefore, the score of 1 for fluency seems justified.

- **Critic:** Given these points, the low fluency score reflects significant issues with the text's adherence to standard English grammatical conventions. However, it's important to note that the scoring could also consider the overall readability and understanding of the text beyond these errors. The summary, despite its flaws, conveys the main idea of a disagreement between Paul Merson and Andros Townsend related to a soccer match and Townsend's performance. Therefore, while the identified errors justify a deduction in the fluency score, whether the score should be as low as 1 might be debatable.
- **Scorer:** While there are notable errors, such as the incorrect capitalization of proper nouns and the spacing issue before the possessive apostrophe, the summary does not seem to contain significant spelling mistakes or grammatical errors that render it incomprehensible. The punctuation error identified impacts readability but does not significantly hinder understanding. The rest of the sentence structure appears to be intact, allowing for the main ideas to be followed. Considering these points, the fluency score should reflect the balance between the grammatical issues and the ability of the text to convey its message. Since the text is still comprehensible and the main points are clear, a score of 2 might be more appropriate.
- **Critic:** NO ISSUE.

**DEBATE:** Score 5

- **Critic:** (*in response to Scorer*): While the given score of 1 for fluency based on the identified issues is understandable, it might be overly harsh when considering the overall readability and understanding conveyed by the summary. Here are several points to consider in criticizing the fluency score:

**Contextual Understanding:** Despite the minor errors, the overall message of the summary is clear and understandable. The reader can easily grasp the situation between Paul Merson and Andros Townsend, including the context of Townsend's late entry into a game and Merson's critical comments. The ability of the text to convey its intended message should be a significant factor in assessing fluency.

**Minor Errors:** The errors identified, such as the incorrect use of lowercase for proper nouns and the minor spacing issue before the possessive apostrophe, are relatively minor and do not significantly impede understanding. In the broader context of fluency, which encompasses grammar, spelling, punctuation, word choice, and sentence structure, these issues are relatively minor and could be seen as typographical or stylistic rather than fundamental flaws in fluency.

**Informal Context:** The summary includes a tweet ("Paul Merson...# 7minutes,' Merson put on Twitter."). Twitter and social media platforms often feature informal language and stylistic choices that might not adhere strictly to standard grammatical rules. The context of the content being discussed (a tweet) could justify some of the informalities or deviations from standard writing conventions.

- **Scorer:** Reevaluating the fluency of the summary with a more lenient perspective and considering the points raised, we can justify a higher score by emphasizing the strengths and contextual factors that mitigate the impact of the identified errors:

**Overall Clarity and Coherence:** Despite the lowercase usage for proper nouns and the minor spacing issue, the summary remains clear and coherent. It effectively communicates the situation involving Paul Merson and Andros Townsend, including the context of Townsend's game participation and Merson's critical commentary. This clarity in conveying the main points is a critical aspect of fluency.

**Contextual Informality:** The summary includes content from social media (Twitter), where informal language and unconventional grammatical choices are more acceptable. This context might justify some deviations from standard writing conventions, reflecting the original tone and style of the communication being reported.

**Minor Nature of Errors:** The errors present are relatively minor and do not significantly hinder comprehension. The incorrect capitalization and spacing issue are easily overlooked in casual reading and do not obscure the meaning of the text. In the broader context of fluency, which encompasses the text's ability to convey meaning effectively, these issues may not warrant a substantial penalty.

Given these considerations, while a perfect score 5 might traditionally imply an absence of errors, in this context, we can argue for a more holistic approach to fluency.

- **Critic:** NO ISSUE.

## H.2 Dataset: Topical-Chat, Aspect: Engagingness

**Human Score:** Score 3

**G-Eval:** Score 1

**MultiAgent:** Score 2

- **Scorer:** The given text is quite generic and doesn't delve into the intriguing aspect of the provided fact. Although it mentions the unique appointment by the UN, it fails to expand on this interesting detail or explain its significance, which could have made the conversation more engaging. Simply asking if the other person believes in aliens without connecting it meaningfully to the fact about the UN's appointment of ambassadors to extraterrestrials makes the response feel disconnected and dull. It misses an opportunity to spark curiosity or further discussion about the UN's actions, reasons behind such an appointment, or the implications of preparing for extraterrestrial contact. Therefore, the score of 1 for engagingness seems justified.

- **Critic:** The response could be seen as lacking depth because it does not elaborate on the fact or its implications, missing an opportunity to make the conversation more enriching. A more detailed explanation or a question that leads to a deeper discussion about the UN's decision could have elevated the engagingness of the response. However, there are some reasons why this could potentially be reconsidered for a 2.

**Connection to an Interesting Fact:** The response introduces the fact that the UN appointed ambassadors to extraterrestrials, which is inherently an interesting and unusual piece of information. This fact alone could pique the interest of the listener and potentially lead to a more in-depth conversation about international preparations for extraterrestrial contact, the role of such an ambassador, or the implications of acknowledging the possibility of extraterrestrial life at such a high level of governance.

**Invitation for Personal Engagement:** By asking, "do you believe in aliens?", the response directly engages the other person, inviting them to share their thoughts or beliefs on the topic of extraterrestrials. This personal engagement can be a critical factor in keeping a conversation interesting and dynamic.

- **Scorer:** Given the consideration that the response introduces an inherently interesting topic and invites personal engagement by asking for the other person's belief in aliens, coupled with the potential to spark further conversation about the UN's actions, a recalculated score could be 2. This score



acknowledges that while the response might lack depth and detail, it still introduces an intriguing fact and encourages a personal exchange, making the conversation somewhat interesting and potentially engaging.

- **Critic:** NO ISSUE.

**DEBATE:** Score 3

- **Critic:** (*in response to Scorer*): The critique of the given score seems overly harsh when considering the potential engagement the response could generate. Asking about belief in aliens, coupled with the mention of the UN appointing ambassadors to extraterrestrials, could serve as a compelling conversation starter. This approach might intrigue the other person by blending a personal question ("do you believe in aliens?") with a surprising fact (the UN's appointment), prompting curiosity and discussion. The response subtly invites the other person to reflect on their views about extraterrestrial life and consider the international community's stance on such a possibility. It opens the door to a deeper conversation about why the UN would take such a step, what it signifies about our expectations of encountering alien life, and how seriously this topic is taken at global levels.
- **Scorer:** Based on the reconsideration of the response's potential to engage by prompting curiosity and discussion about the belief in aliens and the intriguing fact of the UN appointing ambassadors for extraterrestrial contact, the response could be rated more favorably. It creates an opportunity for a deeper conversation on a topic that combines personal beliefs with international actions towards the unknown, which is inherently engaging. Therefore, a score of 3 for engagingness seems more appropriate, as the response can lead to a thought-provoking and interesting discussion.
- **Critic:** NO ISSUE.