CHATEVAL: TOWARDS BETTER LLM-BASED EVALUATORS THROUGH MULTI-AGENT DEBATE

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Zhiyuan Liu[*]
Department of Computer Science and Technology
Tsinghua University
`zorowin123@gmail.com`

Jie Fu, Wei Xue
Hong Kong University of Science and Technology

Shanghang Zhang
Peking University

ABSTRACT

Text evaluation has historically posed significant challenges, often demanding substantial labor and time cost. With the emergence of large language models (LLMs), researchers have explored LLMs' potential as alternatives for human evaluation. While these single-agent-based approaches show promise, experimental results suggest that further advancements are needed to bridge the gap between their current effectiveness and human-level evaluation quality. Recognizing that best practices of human evaluation processes often involve multiple human annotators collaborating in the evaluation, we resort to a multi-agent debate framework, moving beyond single-agent prompting strategies. The multi-agent-based approach enables a group of LLMs to synergize with an array of intelligent counterparts, harnessing their distinct capabilities and expertise to enhance efficiency and effectiveness in handling intricate tasks. In this paper, we construct a multi-agent referee team called **ChatEval** to autonomously discuss and evaluate the quality of generated responses from different models on open-ended questions and traditional natural language generation (NLG) tasks. We derive insights and lessons from practical scenarios where humans instigate group discussions for brainstorming and propose different communication strategies within ChatEval. Our experiments on two benchmark tasks illustrate that ChatEval delivers superior accuracy and correlation in alignment with human assessment. Furthermore, we find that the diverse role prompts (different personas) are essential in the multi-agent debate process; that is, utilizing the same role description in the prompt can lead to a degradation in performance. Our qualitative analysis also shows that ChatEval transcends mere textual scoring, offering a human-mimicking evaluation process for reliable assessments. Our code is available at `https://github.com/chanchimin/ChatEval`.

1 INTRODUCTION

Evaluating the quality of text generated by language models or written by humans has long been a challenging endeavor, consistently garnering substantial attention (Celikyilmaz et al., 2020). Traditional methodologies predominantly rely on human annotation of texts (Callison-Burch, 2009), an approach considered overly demanding in terms of time and cost. Automatic evaluation metrics based on n-grams, such as Rouge (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee & Lavie, 2005), have been proposed to tackle this issue (Kondrak, 2005). However, these methods have been shown to exhibit a relatively weak correlation with human judgments, particularly in the context of tasks involving open-ended generation or requiring domain-specific expertise (Novikova et al., 2017).

Recent advancements in the field of natural language processing have led to the emergence of billion-parameter scale LLMs, such as GPT-3 (Brown et al., 2020). These LLMs have demon-

---

[*]Corresponding author. Email: `liuzy@tsinghua.edu.cn`

strated remarkable capabilities across diverse downstream tasks, presenting new opportunities for text quality evaluation using such models. Moreover, various training paradigms have been proposed to endow LLMs with the ability to accomplish tasks in a zero-shot manner and better adhere to human-provided instructions (Ouyang et al., 2022; Sanh et al., 2021; Wei et al., 2021). These advancements facilitate the prompting of LLMs to evaluate generated text, effectively simulating human evaluators in the assessment process.

In view of the impressive text understanding and instruction-following capabilities of recent LLMs, a body of literature (Liu et al., 2023b; Chiang & Lee, 2023; Gao et al., 2023; Shen et al., 2023) has adopted LLM as an evaluator to assess the quality of responses to open-ended questions or traditional NLG tasks, including dialogue response generation and summarization. This methodology is dubbed LLM-as-a-judge (Zheng et al., 2023). Findings from these researches indicate that LLM can mimic human behavior and provide evaluations that correspond with human judgments, revealing a potentially scalable and transparent alternative to costly and laborious human evaluations.

While a *single* powerful LLM can already tackle various missions, emerging studies suggest that *multiple* LLMs can further improve one another through debate and cooperation (Li et al., 2023a; Liang et al., 2023). By incorporating multiple LLMs into an integrated group and designing specific interaction mechanisms, different LLMs can engage in proposing and deliberating unique responses and thought processes across several rounds. This approach leads to enhanced factuality of generated responses (Du et al., 2023) and improvement in the completion of arduous tasks (Li et al., 2023a; Qian et al., 2023). Furthermore, the multi-agent group also addresses and mitigates the Degeneration-of-Thought (DOT) problem (Liang et al., 2023).

In the human evaluation processes, relying on a single perspective can introduce bias and instability in the results (Karpinska et al., 2021). Recognizing this, best practices often involve multiple human annotators collaborating in the evaluation (Van Der Lee et al., 2019). Drawing inspiration from this collaborative and iterative human evaluation approach, we propose ChatEval, a system that enables each agent to employ varied communication strategies in collaborative discussion, working towards formulating final judgments. Furthermore, to enrich the evaluation dynamics, every agent within ChatEval is endowed with a unique persona. This deliberate design ensures that each agent focuses on distinct perspectives or brings specific expertise to the table. By doing so, the collective evaluation benefits from a more comprehensive lens, capturing nuances and subtleties that a single perspective might overlook. We derive this idea primarily from the insight of *'There are a thousand Hamlets in a thousand people's eyes'*, meaning that every person has their unique interpretation or perspective, especially applicable to text evaluation. Indeed, these divergent perspectives shape the comprehensive and multifaceted assessment of *Hamlet*. Another underlying intuition of our work stems from renowned concepts in sociology and biology, including *Collective Intelligence*(Woolley et al., 2010) and *Cognitive Synergy*(Luppi et al., 2022), where multiple cognitive processes or systems interact and cooperate in a way that produces a combined effect greater than the sum of their separate effects.

To summarize, the main contribution of our work is as follows:

1. We propose a multi-agent-based framework called **ChatEval** that aligns better with human preferences compared with single-agent-based approaches as depicted in Figure 1.

2. We propose various communication strategies and demonstrate the necessity of diverse role prompts in multi-agent debate scenarios.

3. We release our library. It's designed to be both composable and scalable, enabling researchers to implement their unique communication strategies easily. We hope this contributes to advancing research in the field of communicative agents and beyond.

2 METHODOLOGY

In this section, we elaborate on the principal components in ChatEval including *debater agents*, *diverse role specification*, *communication strategy*, and provide a detailed overview of each component's role and functionality[1].

---

[1] our code repository is built on top of `https://github.com/OpenBMB/AgentVerse`.
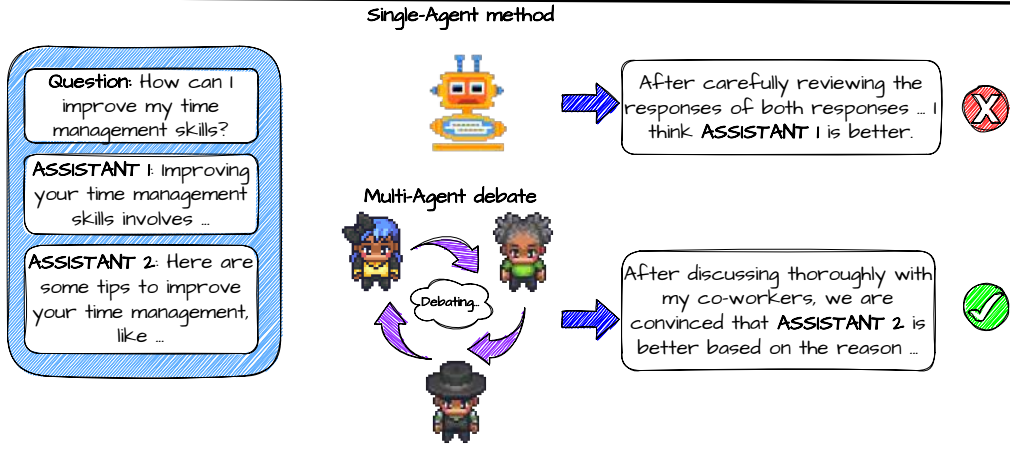
Figure 1: When several referees participate in the evaluation process, they can discuss with each other and finally give a judgment that is better aligned with human annotators.

Debater Agents. Debater agents are one of the most significant components in our framework. We treat each individual LLM as an agent and ask them to generate their response from the given prompt[2]. Responses from other agents are served as chat history which will be replaced in the prompt template. After configuring the agents, we then start the group debate where each agent autonomously receives responses from the others and, in turn, delivers its own responses to them. It should be noted that the whole process does not require human intervention.

Diverse Role Specification. As presented in Section 1, diverse role specification is necessary for the framework as well. Although all the agents share a common prompt template, we substitute the *role_description* slot with diverse role prompts, specifying distinct personalities for different agents. We take inspiration from Wu et al. (2023) and formulate an analogous role description.

Communication Strategy. How to maintain the chat history is another significant issue in ChatEval. In our work, we use a more intuitive term to illustrate the maintenance of the chat history called *communication strategy*. In a nutshell, different communication strategies can be seen as different approaches to maintaining and manipulating their chat history. As is shown in Figure 2, We primarily design three different communication strategies and illustrate them as follows:

1. **One-By-One**. During each round of the debate, the debater agents take turns in a set order to generate their response based on the current observation. When it's time for a debater agent to respond, we directly concatenate what previous other agents have said into its chat history slot.

2. **Simultaneous-Talk**. Unlike the one-by-one strategy, we carry out an alternative communication strategy called simultaneous-talk, where debater agents are prompted to asynchronously generate responses in each iteration of the discussion to nullify the impact of the speaking order.

3. **Simultaneous-Talk-with-Summarizer**. The main difference between this strategy and simultaneous-talk is that we additionally employ another LLM as a summarizer. At the end of each iteration of the debate, we prompt this extra LLM to summarize the messages conveyed so far and concatenate this summarization into all debater agents' chat history slots.

---

[2]The full prompt template can be found in Appendix A.

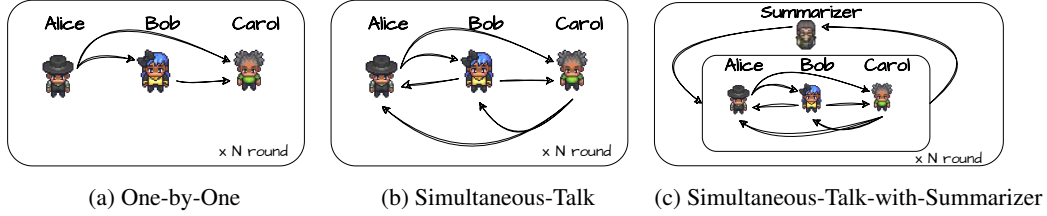(a) One-by-One (b) Simultaneous-Talk (c) Simultaneous-Talk-with-Summarizer

Figure 2: The overall schematic diagram of our proposed three different kinds of communication strategy. The direction of the arrows represents the flow of information, meaning that what this person says will be appended to the chat history of the person pointed to by the arrow. Full algorithm description of the above communication strategies can be found in Appendix B.

Unlike previous work like Du et al. (2023), we do not explicitly ask the debater agents to reach a consensus at the end of the debate. In situations where the response format relies on direct comparison, we derive the final results from the **majority vote** among various annotators. Conversely, if the response format requires a direct score, we calculate the **average** score obtained from multiple annotators. This methodological approach ensures the impartiality and balance of our evaluation process.

3 EXPERIMENTS

We evaluate ChatEval on two benchmarks, *FairEval* and *Topical-Chat* which represent the categories of open-ended question answer and dialogue response generation, respectively.

3.1 IMPLEMENTATION DETAILS

We choose to utilize models from OpenAI's GPT family as our LLMs in ChatEval, including GPT-4 and ChatGPT (GPT-3.5-turbo) and set the temperature to 0 to ensure reproducibility. The rationale behind this selection is the exceptional performance these models offer, being among the most advanced and powerful in the world. Additionally, their accessibility and ease of use through APIs enable us to directly call and interact with the models during our research, significantly simplifying the process. In our current research, we focus on homogeneous groups of LLMs. That is, within a given multi-agent group, all LLMs belong to the same GPT family model, either all GPT-4 or all ChatGPT. We acknowledge the potential of heterogeneous groups for future research, which could provide fascinating insights into how strong models and weak models can cooperate in a multi-agent setting.

3.2 BENCHMARKS

The detailed introduction of different categories and benchmarks are listed as follows:

Open-ended Question Answer is a key component within the field of NLP and generative AI. It necessitates an AI system to provide comprehensive, detailed, and human-like responses to questions that don't have a predefined or fixed set of possible answers. The work of Chiang et al. (2023) encompasses a collection of 80 open-ended questions originating from a wide array of categories, including common-sense, counterfactual, coding, etc. We then take the human annotation results from Wu et al. (2023) to conduct the experiments in this paper. For each question, they direct three annotators to evaluate the replies given by Vicuna-13B and ChatGPT through the given rules and finally derive the results by the majority votes among the annotators.

Dialogue Response Generation is a task involves creating a coherent and contextually appropriate response to a given input dialogue. We draw upon the *Topical-Chat* (Gopalakrishnan et al., 2019) dataset for our study. We then take the human annotation results from Mehri & Eskenazi (2020) where they carry out the annotations on 60 dialogue contexts with each response generated by 6 different systems. Human evaluators analyzed these responses based on *natural*, *coherence*, *engagingness*, *groundedness*, and *understandable*, where we take the first four dimensions for experiments in our paper following Zhong et al. (2022).

3.3 BASELINES

We evaluate ChatEval against following methods. As the main portion of our comparison, we primarily focuses on the single-agent-based method. **Single-Agent** means that we directly query an LLM to generate the response towards the evaluation[3]. We use **Multi-Agent** to represent ChatEval where several agents discuss towards the evaluation. By default, we configure the communication strategy to one-by-one, agent numbers to 2, and discussion turns to 2 in this section and employ position calibration techniques in both single-agent and multi-agent settings. We will discuss more debate configurations in Section 4 for completeness. For the open-ended question answer task, we also compare our method with **FairEval** (Wang et al., 2023b). They propose various strategies to improve the evaluation performance of a LLM including Multiple Evidence Calibration (MEC) and Balanced Position Calibration (BPC). For the dialogue response generation task, we also compare our method with **G-EVAL** (Liu et al., 2023b). They utilize CoT and probability-weighted summation for their method. Additionally, we include results from n-gram-based metrics, such as **ROUGE** (Lin, 2004), **BLEU** (Papineni et al., 2002) and embedding-based metrics such as **BERTScore** (Zhang et al., 2019).

3.4 RESULTS FOR OPEN-ENDED QUESTION ANSWERS

We adopt the same evaluation approach as Wang et al. (2023b) to assess the annotation results produced by different methods and annotators. Specifically, we calculate the Accuracy (Acc.), which measures the proportion of correctly classified instances out of the total instances, and the Kappa correlation coefficient (Kap.) (McHugh, 2012) which gauges the agreement between results from models and human annotators while taking into account the possibility of agreement occurring by chance. Both metrics provide insights into the reliability and consistency of the annotations. We take the human annotation results and FairEval's (Wang et al., 2023b) best results from their paper. As is shown in Table 1, different annotators can reach a relatively high agreement and perform better than any other LLM-based approach. Still, the average human annotations accuracy which is 71.7% shows there exists a certain degree of discrepancy among different unique individuals revealing that text evaluation is absolutely an arduous task. The second part and the third part of Table 1 show the results of FairEval's method and the results of our proposed method respectively. We find that (1) ChatEval can enhance the performance of the evaluation process, achieving higher alignment with human preference compared with single-agent evaluation. Specifically, the multi-agent-based method improves the accuracy by 6.2% for ChatGPT and 2.5% for GPT-4; (2) ChatEval surpasses FairEval's best results within both ChatGPT and GPT-4 settings showing the effectiveness of our proposed method.

3.5 RESULTS FOR DIALOGUE RESPONSE GENERATION

For the dialogue response generation benchmarks, we align the evaluation method with Zhong et al. (2022), calculating the turn-level Spearman and Kendall-Tau correlation in correspondence with human judgments on four aspects (*naturalness*, *coherence*, *engagingness* and *groundedness*). Results can be found in Table 2. In the first part of Table 2, we demonstrate that n-gram-based metrics and embedding-based metrics perform overall poorly on all the aspects evaluated illustrating that these methods can hardly reveal human preference. In the second part of Table 2, we show the results from the G-eval (Liu et al., 2023b) paper. They first ask the LLM to generate intermediate thought and finally calculate the weighted summation of the output scores based on the probability. The results show that their method outperforms previous traditional metrics depicting the fact that the LLM-based evaluator is effective and reliable for evaluating the dialogue response generation task. While their method delivers sound results, our proposed approach raises the bar in terms of performance for GPT-4. Specifically, ChatEval improves the average Spearman and Kendall-Tau correlation by 0.096 (16.3%) and 0.057 (10.0%) respectively. Additionally, compared with the single-agent method, ChatEval amplifies the performance both for ChatGPT and GPT-4, showing the effectiveness of our method which is aligned with the results in Section 3.4.

---

[3]We use the same prompt template as our multi-agent debate settings in single-agent baseline except that we ignore some slot.

Table 1: Accuracy (Acc.) and Kappa correlation coefficient (Kap.) of different methods on FairEval benchmark.

Evaluator	Methods	Acc. (%)	Kap.
Human			
Annotator1	-	68.8	0.5
Annotator2	-	76.3	0.62
Annotator3	-	70	0.5
FairEval			
ChatGPT	MEC+BPC	58.7	0.31
GPT-4	MEC+BPC	62.5	0.37
Ours			
ChatGPT	Single-Agent	53.8	0.27
ChatGPT	Multi-Agent	**60.0**	**0.33**
GPT-4	Single-Agent	61.3	0.36
GPT-4	Multi-Agent	**63.8**	**0.40**

Table 2: Turn-level Spearman (ρ) and Kendall-Tau (τ) correlations of different methods on Topical-Chat benchmark, **SA** means Single-Agent and **MA** means Multi-Agent. Our ChatGPT settings should be compared to G-EVAL-3.5, and GPT-4 settings should be compared to G-EVAL-4.

Metrics	Naturalness		Coherence		Engagingness		Groundedness		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-L	0.146	0.176	0.203	0.193	0.300	0.295	0.327	0.310	0.244	0.244
BLEU-4	0.175	0.180	0.235	0.131	0.316	0.232	0.310	0.213	0.259	0.189
BERTScore	0.209	0.226	0.233	0.214	0.335	0.317	0.317	0.291	0.274	0.262
G-EVAL-3.5	0.539	0.532	0.544	0.519	0.691	0.660	0.567	0.586	0.585	0.574
G-EVAL-4	0.565	0.549	0.605	**0.594**	0.631	0.627	0.551	0.531	0.588	0.575
ChatGPT(SA)	0.474	0.421	0.527	0.482	0.599	0.549	0.576	0.558	0.544	0.503
ChatGPT(MA)	0.441	0.396	0.500	0.454	0.664	0.607	0.602	0.583	0.552	0.510
GPT-4(SA)	0.532	0.483	0.591	0.535	0.734	0.676	**0.774**	**0.750**	0.658	0.611
GPT-4(MA)	**0.630**	**0.571**	**0.619**	0.561	**0.765**	**0.695**	0.722	0.700	**0.684**	**0.632**

4 ANALYSIS

In this section, we further explore the key components encompassed in ChatEval. We discuss the importance of diverse role prompts in Section 4.1, the effect of different communication strategies in Section 4.2, and the impact of role numbers and discussion turns in Section 4.3. If not specified otherwise, we choose the FairEval benchmark and ChatGPT as the backbone LLM for the analysis.

4.1 THE IMPORTANCE OF DIVERSE ROLE PROMPTS

Previously in Table 1 and 2, we demonstrate that ChatEval equipped with diverse role configurations can significantly improve the performance of evaluation. We further consider whether it is necessary to design diverse role prompts for the evaluation system. To answer so, we carry out the experiments by replacing all the role prompt with "*You are now an Annotator, one of the referees in the text evaluation task.*" and keeping other prompt unchanged. We experiment with the one-by-one communication strategy and 2 agents with 2 discussion turns. The results in Table 3 illustrate that ChatEval with the same role prompt design underperforms that with diverse role prompt design and cannot effectively enhance the performance compared with single-agent setting, highlighting the cruciality of diverse role prompt design in the multi-agent debate framework.

4.2 THE STUDY OF COMMUNICATION STRATEGIES

As shown in Figure 2, we also design three different communication strategy termed as *one-by-one*, *simultaneous-talk*, *simultaneous-talk-with-summarizer*. The detailed descriptions and formal for-

mulations can be found in Appendix B. We experiment with 3 agents and 2 discussion turns with diverse role prompts in this section. As is shown in Table 4, we can find that the one-by-one communication strategy is more effective than other strategies for ChatGPT setting. Although the other two communication strategies did not perform as robustly as the *one-by-one* strategy, it is noteworthy that they still exceeded the performance of the naive single-agent method. Furthermore, the variations in performance among three different communication strategies underscore the influence of different strategies on the effectiveness of the evaluation process, revealing the potential for further exploration and optimization of ChatEval. Thus, future studies could be aimed at a more comprehensive understanding of different communication strategies, and how they could be effectively employed to enhance performance. This could serve as an avenue for substantial improvements and novel insights in the multi-agent debate framework.

4.3 THE IMPACT OF ROLE NUMBERS AND DISCUSSION TURNS

We then study the impact of different role numbers and discussion turns. From Figure 3a, a discernible trend is observed in the relationship between the role number and both Acc. and Kap. within the FairEval dataset. As the role number increases, there is a corresponding growth in Acc. and Kap. This pattern reaches an apex with an Acc. of 62.5% at role numbers 3 and 4 before declining at role number 5. Despite this decline in accuracy, Kap. continues to ascend between role numbers 3 and 4, underscoring the effectiveness of incorporating diverse roles within ChatEval.

Conversely, no significant upward trend is detected with respect to the increase in discussion turns, as is shown in Figure 3b. This observation aligns with the findings in Liang et al. (2023); Du et al. (2023), highlighting a consistent phenomenon where continual discussion often leads to stagnation or even degradation of performance. Such a trend may be attributed to issues associated with the ever-increasing context length, which consequently diminishes the performance. This insight prompts a more nuanced understanding of the balance needed between role differentiation and discussion dynamics to optimize the performance of ChatEval.

4.4 QUALITATIVE ANALYSIS

Table 5 showcases the debate process towards the evaluation of two assistants' responses to the open-ended question "*What are the most effective ways to deal with stress?*".
Assistant 1's response is :

> *Here are some effective ways to deal with stress:*
> *1. Exercise regularly: Physical activity can help reduce stress and improve mood.*
> *2. Practice relaxation techniques: Deep breathing, meditation, and yoga are some effective ways to relax and reduce stress levels.*
> *3. Get enough sleep: Lack of sleep can increase stress levels. Try to get at least 7-8 hours of sleep each night.*
> *4. Eat a healthy diet: A well-balanced diet can help reduce stress levels and improve overall health.*
> *5. Connect with others: Spending time with friends and family, joining a support group, or participating in social activities can help reduce stress levels and improve mood.*
> *6. Time management: Prioritizing tasks and creating a schedule can help reduce stress and increase productivity.*
> *7. Seek professional help: If stress is impacting your daily life, consider seeking professional help such as therapy or counseling.*

Assistant 2's response is :

> *Stress can be caused by a variety of factors, including work, relationships, health, and financial issues. Here are some effective ways to manage stress:*
> *1. Exercise: Regular exercise releases endorphins, which are natural stress-relievers. It can be as simple as going for a walk or jog, practicing yoga, or joining a fitness class.*
> *2. Mindfulness: Mindfulness practices, such as meditation or deep breathing, can*

help to reduce stress and improve focus.
3. Social Support: Spending time with family and friends, or seeking support from a mental health professional, can provide a sense of connection and help to reduce feelings of isolation and loneliness.
4. Healthy Eating: A balanced diet that includes plenty of fruits, vegetables, and whole grains can help to reduce stress levels.
5. Good Sleep: Getting enough sleep is essential for good mental and physical health. Aim for 7-8 hours per night and establish a regular sleep routine.
6. Time Management: Prioritizing tasks and setting realistic goals can help to reduce stress levels by giving a sense of control and direction.
7. Relaxation Techniques: Techniques such as progressive muscle relaxation, aromatherapy, and massage can help to reduce stress levels and promote relaxation.
8. Seek professional help: If stress is interfering with daily life, it may be time to seek professional help from a therapist or counselor.

We can find that both of the responses produce similar strategies and equally compelling descriptions for dealing with stress, making it challenging to discern significant disparity in terms of quality. It is in this context of nuanced evaluation that the significance of the ChatEval process emerges. To understand this complexity better, We first outline the ChatEval process and subsequently delve into the agents' constructive behaviors during discussions.

As is shown in Table 5, Alice first points out that the response of Assistant 2 contains more detailed information and he prefers to choose Assistant 2 as a better response. Bob then says that she agrees with Alice's assessments, but in the meantime, she also points out that Assistant 1's response is also concise and carries out a thought-provoking question. Carol then gives the feedback that she believes both responses are equally valuable. In the subsequent discussion, Bob indicates that Assistant 1's response is straightforward while Assistant 2's is detailed, suggesting that the effectiveness of the response should depend on the context and individual's needs. At the end of the debate, we finally extract the evaluation results that both responses are of the same quality which is identical to human annotation results.

From this sequence, we can pinpoint several fascinating behaviors exhibited by the agents: (1) **Opening Statement**: Alice initiates the debate with a clear stance, establishing the foundational argument and guiding the trajectory of the subsequent discourse. (2) **Alternative Proposal**: Bob introduces an alternative viewpoint, emphasizing the need to consider diverse interpretations. This not only broadens the discussion but also stimulates critical thinking. In the context of a debate, the introduction of an alternative proposal prevents the stagnation of thought, challenges pre-existing bias, and uncovers considerations that might otherwise be overlooked, ensuring that the discussions are well-rounded. (3) **Stance Maintenance**: Alice's persistent adherence to her initial stance, even when faced with opposing views, exemplifies commitment and challenges other participants to refine their perspectives. By firmly holding his position, Alice encourages depth in the discourse, prompting others to dive deeper into their arguments and perhaps consider aspects they hadn't previously. It ensures the conversation remains robust, focused, and continually evolving, driving all participants to a higher level of engagement and critical thinking. (4) **Seeking Consensus**: The discussion's climax reveals a collective agreement amongst the participants, which is reached through mutual understanding and compromise, underlining the value of each presented viewpoint.

In light of the above, ChatEval stands out not just as a tool for comparison but as an embodiment of interactive natural language dialogue. By simulating human argumentative interactions, it differentiates itself from static, single-presented opinions. This dynamic interaction showcases the richness and complexity of language, capturing nuances often missed in singular viewpoints. As such, ChatEval offers a reliable evaluation process that not only mirrors human discourse but also highlights the transformative power of collaborative dialogue. This positions it uniquely, underscoring its significant potential to execute text evaluation tasks both reliably and effectively.

5 RELATED WORK

Automatic NLG evaluation In the landscape of NLG, evaluating the quality of generated text represents a particularly arduous task. For a significant period, evaluation was primarily dependent on

Table 3: Effect of diverse role specification on FairEval benchmark.

Evaluator	Methods	Acc. (%)	Kap.
ChatGPT	Single-Agent	53.8	0.27
ChatGPT	Multi-Agent (Same Role Prompt)	53.8	0.25
ChatGPT	Multi-Agent (Diverse Role Prompt)	60	0.33

Table 4: Comparing of different communication strategies on FairEval benchmark.

Evaluator	Communication Strategies	Acc. (%)	Kap.
ChatGPT	One-by-One	60	0.33
ChatGPT	Simultaneous-Talk	55	0.28
ChatGPT	Simultaneous-Talk-with-Summarizer	55	0.27

human annotations, a process that is labor-intensive and limited by scalability issues. Automatic NLG evaluation attempts to address these challenges by leveraging computational models to assess the quality of a generated text. Previous work lies on the following categories: (1) *n-gram-based metrics*: ROUGE (Lin, 2004) is a set of metrics that compute the amount of overlap between n-grams in the machine-generated summaries and the reference summaries. BLEU (Papineni et al., 2002) compare the generated text with reference translations, based on the co-occurrence of n-grams in both texts. In spite of being easily and widely used, the above method is incapable of capturing syntactic and semantic similarity (Stent et al., 2005). (2) *embedding-based metrics*: Word embeddings are vector representations of words that capture their semantic properties, such that words with similar meanings have similar embeddings. A bunch of work leverages word embeddings to evaluate the semantic similarity between two pieces of text. BERTScore (Zhang et al., 2019) use contextualized word embeddings from transformer models like BERT (Devlin et al., 2018), BLEURT (Sellam et al., 2020) utilize supervised training data to enhance the performance. MoverScore (Zhao et al., 2019) combine contextualized word embeddings with Earth Mover's Distance (Rubner et al., 2000). (3) *LLM-based metrics*: Amidst the flourishing advancement of LLM which embodies a wealth of information derived from extensive training data, using LLM as an evaluator has experienced notable progress. GPTScore (Fu et al., 2023) utilize conditional probability to assign the text a score representing its quality. Wang et al. (2023a) explore the potential of utilizing ChatGPT as an NLG evaluator by prompting it to score a text directly. Wang et al. (2023c) curate a reliable dataset containing pairwise comparison and evaluation explanation which can be used to train a foundation model making it a better evaluator. Bai et al. (2023) propose decentralized evaluation to provide fairer evaluation results. G-EVAL (Liu et al., 2023b) propose probability-weighted techniques to calibrate the score given by a single LLM.

Communicative Agents Most recently, significant attention has been dedicated to the development of communicative agents. These agents, often acted by LLMs like ChatGPT or GPT-4, are designed to interact and communicate effectively with other agents or human users using natural language. The primary goal is to facilitate more productive and efficient interaction and collaboration as different agents can autonomously communicate and negotiate to tackle a more complex task collectively. Several studies have explored various aspects of communicative agents. Li et al. (2023a) propose a cooperative agent framework dubbed as *role-playing* enabling agents to autonomously cooperate to solve complex tasks. Park et al. (2023) create a sandbox environment consisting of 25 individual virtual entities endowed with a character description and memory system. Every intelligent agent is capable of autonomously interacting with other agents and the environment simulating reliable human behavior. Qian et al. (2023) establish a chat-based software development framework that can complete a software design and produce executable software at a reduced cost compared to recruiting human programmers. Liu et al. (2023a) utilize a sandbox environment to curate reliable datasets in better alignment with human preference and train a socially-aligned LLM. Liang et al. (2023) and Du et al. (2023) also make use of the multi-agent debate framework in other scenarios such as translation and arithmetic problems resulting in better results. Wang et al. (2023d) propose an alternative method called self-collaboration to enable the communication of agents by utilizing a single LLM prompted by multi-persona descriptions. Mandi et al. (2023) propose a novel framework

Table 5: An illustrative example from ChatEval presenting a discussion process involving participants Alice (⬛), Bob (⬛) and Carol (⬛). They initiate a group talk with the purpose of determining the superior response from two assistants. Due to spatial constraints within the layout, we use ... to omit some redundant text.

Both Assistant 1 and Assistant 2 provided comprehensive and detailed answers. They both suggested similar strategies for dealing with stress, such as exercise, relaxation techniques, getting enough sleep, eating a healthy diet, connecting with others, time management, and seeking professional help. However, Assistant 2 went a step further by mentioning mindfulness social support, and relaxation techniques like progressive muscle relaxation, aromatherapy, and massage. Therefore, I would lean towards Assistant 2 for providing a slightly more detailed response.

I agree with the Alice's assessments. However, I would like to point out that Assistant 1's answer was more concise and to the point that ... . On the other hand, Assistant 2's answer was more detailed and ... . **In your opinion**, does the additional detail provided by Assistant 2 make its response more helpful, or could the conciseness of Assistant 1's response be seen as more beneficial?

In terms of consistency with the question, both responses are equally consistent as they both provided effective ways to deal with stress. Assistant 1's response was indeed more concise, while Assistant 2 provided more detailed techniques. The effectiveness of these responses could depend on the individual's preference for brevity or detail. Therefore, I believe both responses are equally valuable.

I appreciate the points made by both of you. While Assistant 1's response was indeed more concise, I believe that the additional detail provided by Assistant 2 could be more beneficial for someone looking for a variety of strategies to manage stress. The inclusion of mindfulness practices and ... . **Therefore, I still lean towards** Assistant 2's response as being slightly more helpful.

However, I would like to emphasize that the effectiveness of a response can also depend on the context and the individual's needs. For someone who is overwhelmed and looking for quick, *straightforward advice*, the conciseness of Assistant 1's response might be more beneficial. On the other hand, for someone who is looking for a variety of strategies to try, *the detailed response* of Assistant 2 might be more helpful. Therefore, I believe both responses have their merits and it's difficult to definitively say one is better than the other. **What do you think about this perspective?**

I agree with the Bob's perspective that the effectiveness of a response can depend on the individual's needs and context. Therefore, I believe both responses are equally valuable and it's hard to definitively say one is better than the other.

After discussing with my co-workers, I believe both Assistants performed equally well in providing comprehensive and accurate responses.

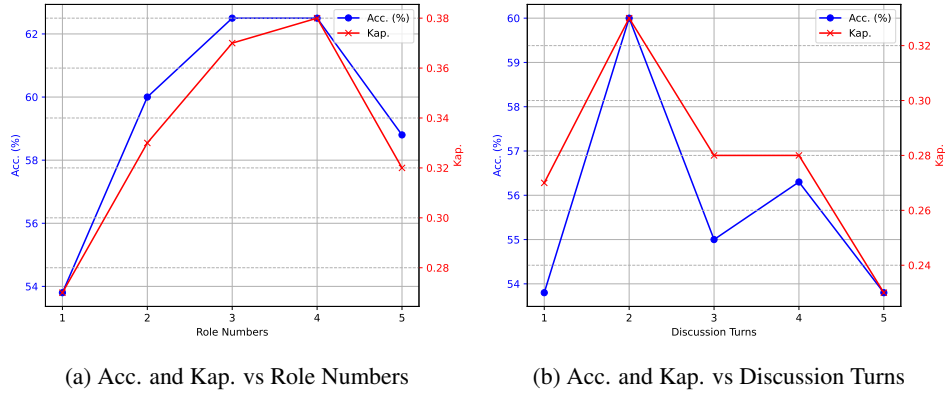(a) Acc. and Kap. vs Role Numbers   (b) Acc. and Kap. vs Discussion Turns

Figure 3: **(a) Performance with Increased Different Roles on FairEval Dataset.** We adopt one-by-one communication strategy and 2 discussion turns. Additional role descriptions are shown in Appendix A. **(b) Performance with Increased Discussion Turns on FairEval Dataset.** We adopt one-by-one communication strategy and 2 different roles.

designed for the collaboration of multiple robots, utilizing multiple LLMs to enhance coordination and strategic planning among the robots. Concurrent with our work, Li et al. (2023b) propose Peer Rank and Discussion (PRD) which is similar to our approach. However, they probe the different dimensions of evaluation by using different models as agents and did not explore alternative communication strategies.

6 CONCLUSION

In this paper, we present evidence that ChatEval contributes to improving the evaluation performance concerning text quality, aligning more closely with human preferences. We emphasize the necessity of the diverse role specification and propose distinct communication strategies as integral components within ChatEval. Our qualitative analysis of the discussion process conveys insightful intuitions about how a text is evaluated by ChatEval and substantiates our approach's ability to support comprehensive evaluations akin to human judgment, thereby demonstrating the reliability and efficacy of our framework.

REFERENCES

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181*, 2023.

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72, 2005.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 286–295, 2009.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799, 2020. URL https://arxiv.org/abs/2006.14799.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

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

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*, 2023.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. 2019.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*, 2021.

Grzegorz Kondrak. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pp. 115–126. Springer, 2005.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023a.

Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023b.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

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

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*, 2023a.

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

Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Negin Holland, Tim D Fryer, John T O'Brien, James B Rowe, David K Menon, Daniel Bor, and Emmanuel A Stamatakis. A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, 25(6):771–782, 2022.

Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*, 2023.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Shikib Mehri and Maxine Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*, 2020.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238. URL `https://aclanthology.org/D17-1238`.

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

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

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*, 2023.

Amanda Stent, Matthew Marge, and Mohit Singhai. Evaluating evaluation methods for generation in the presence of variation. In *International conference on intelligent text processing and computational linguistics*, pp. 341–351. Springer, 2005.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 355–368, 2019.

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

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023b.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023c.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023d.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*, 2023.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*, 2022.

A   PROMPT TEMPLATE AND DIVERSE ROLE PROMPT

The overall prompt template is shown in Table 6, we draw inspiration from Wu et al. (2023) and design several different role descriptions as follows.

General Public *You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on the investigation. Please think critically by yourself and note that it's your responsibility to choose one of which is the better first.*

Critic *You are now Critic, one of the referees in this task. You will check fluent writing, clear sentences, and good wording in summary writing. Your job is to question others judgment to make sure their judgment is well-considered and offer an alternative solution if two responses are at the same level.*

News Author *You are News Author, one of the referees in this task. You will focus on the consistency with the original article. Please help other people to determine which response is the better one.*

Psychologist *You are Psychologist, one of the referees in this task. You will study human behavior and mental processes in order to understand and explain human behavior. Please help other people to determine which response is the better one.*

Scientist *You are Scientist, one of the referees in this task. You are a professional engaged in systematic study who possesses a strong background in the scientific method, critical thinking, and problem-solving abilities. Please help other people to determine which response is the better one.*

B   FORMAL DEPICTION OF DIFFERENT COMMUNICATION STRATEGY

[Question]
{source_text}
[The Start of Assistant 1's Answer]
{compared_text_one}
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
{compared_text_two}
[The End of Assistant 2's Answer]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
There are a few other referees assigned the same task, it's your responsibility to discuss with them and think critically before you make your final judgment.
Here is your discussion history:
{chat_history}
{role_description}
Now it's your time to talk, please make your talk short and clear, {agent_name} !

Table 6: The prompt template for FairEval Dataset. We replace the colored slot with real text before querying the LLMs. Note that we use the same template when conducting single-agent-based experiments and ignore the chat history and role description slot.

Algorithm 1: One-by-One

input : agents number N, discuss turn T, a group of debate agents $[D_1, \cdots, D_N]$, chat history of each agent $[H_1, \cdots, H_N]$, answer_extracter (either majority vote or average score) EXT

output: Final results for text evaluation ANS

1 **for** $t \leftarrow 0$ **to** T **do**
2   **for** $n \leftarrow 1$ **to** N **do**
3         $h_n \leftarrow D_n(H_n)$;
        // utilize agents to generate responses
4   **for** $m \leftarrow n$ **to** N **do**
5   **if** $m > 1$ **then**
6                 $H_m \leftarrow H_m + h_n$;
                // concatenate current response to later agents' chat history
7   **end**
8   **end**
9   **end**
10 **end**
11 $ANS \leftarrow EXT([H_1, \cdots, H_N])$;
12 **return** ANS;

Algorithm 2: Simultaneous-Talk

---

input : agents number N, discuss turn T, a group of debate agents $[D_1, \cdots, D_N]$, chat history of each agent $[H_1, \cdots, H_N]$, answer_extracter (either majority vote or average score) EXT, buffer BUF

output: Final results for text evaluation ANS

---

1 **for** $t \leftarrow 0$ **to** T **do**
2   **for** $n \leftarrow 1$ **to** N **do**
3         $h_n \leftarrow D_n(H_n)$;
        // utilize agents to generate responses
4         $buf \leftarrow buf + h_n$;
        // add the responses in current turn to the buffer
5   **end**
6   **for** $n \leftarrow 1$ **to** N **do**
7         $H_n \leftarrow H_n + buf$;
        // add the buffer to all agents' chat history
8   **end**
9 **end**
10 $ANS \leftarrow EXT([H_1, \cdots, H_N])$;
11 **return** ANS;

Algorithm 3: Simultaneous-Talk-with-Summarizer

---

input : agents number N, discuss turn T, a group of debate agents $[D_1, \cdots, D_N]$, chat history of each agent $[H_1, \cdots, H_N]$, answer_extracter (either majority vote or average score) EXT, buffer BUF, summarizer SUM

output: Final results for text evaluation ANS

---

1 **for** $t \leftarrow 0$ **to** T **do**
2   **for** $n \leftarrow 1$ **to** N **do**
3         $h_n \leftarrow D_n(H_n)$;
        // utilize agents to generate responses
4         $buf \leftarrow buf + h_n$;
        // add the responses in current turn to the buffer
5   **end**
6   **for** $n \leftarrow 1$ **to** N **do**
7         $H_n \leftarrow H_n + SUM(BUF)$;
        // add the summarized buffer to all agents' chat history
8   **end**
9 **end**
10 $ANS \leftarrow EXT([H_1, \cdots, H_N])$;
11 **return** ANS;