

Collaborative Evaluation of Deepfake Text with Deliberation-Enhancing Dialogue Systems

Jooyoung Lee¹, Xiaochen Zhu², Georgi Karadzhov²
Tom Stafford³, Andreas Vlachos², Dongwon Lee¹

¹The Pennsylvania State University, USA, {jfl5838, dongwon}@psu.edu

²University of Cambridge, United Kingdom, {xz479, gmk34, av308}@cam.ac.uk

³University of Sheffield, United Kingdom, t.stafford@sheffield.ac.uk

Abstract

The proliferation of generative models has presented significant challenges in distinguishing authentic human-authored content from deepfake content. Collaborative human efforts, augmented by AI tools, present a promising solution. In this study, we explore the potential of DeepFakeDeLiBot, a deliberation-enhancing chatbot, to support groups in detecting deepfake text. Our findings reveal that group-based problem-solving significantly improves the accuracy of identifying machine-generated paragraphs compared to individual efforts. While engagement with DeepFakeDeLiBot does not yield substantial performance gains overall, it enhances group dynamics by fostering greater participant engagement, consensus building, and the frequency and diversity of reasoning-based utterances. Additionally, participants with higher perceived effectiveness of group collaboration exhibited performance benefits from DeepFakeDeLiBot. These findings underscore the potential of deliberative chatbots in fostering interactive and productive group dynamics while ensuring accuracy in collaborative deepfake text detection.

Datasets —

<https://github.com/Brit7777/icwsm26-deepfakedelibot>

Introduction

Large language models (LLMs) have transformed people’s writing practice with remarkably fluent and human-like generation capabilities. However, their use in real-world applications is hampered by issues such as preserving biases or stereotypes (Kotek, Dockum, and Sun 2023), spreading mis/disinformation (Lucas et al. 2023; Barman, Guo, and Conlan 2024), and facilitating plagiarism (Lee et al. 2023; Hutson 2024). To this end, researchers have recently been putting efforts into distinguishing deepfake texts (i.e., texts generated by machines) from human-authored texts. In particular, the focus has been on the curation of detection benchmarks (Uchendu et al. 2021; Li et al. 2024; Wang et al. 2024) and the automation of detection procedure (Venkatraman, Uchendu, and Lee 2024; Hu, Chen, and Ho 2023; Wang et al. 2023; Mitchell et al. 2023). Yet, these detectors can be easily fooled by simple paraphrasing (Krishna et al. 2024) and are not robust to unseen models and domains

(Weber-Wulff et al. 2023). These limitations necessitate exploring alternative strategies, such as integrating human-in-the-loop mechanisms, where human evaluators validate or supplement existing detectors.

Prior studies (Uchendu et al. 2021; Dou et al. 2022; Jakesch, Hancock, and Naaman 2023) have reported that humans alone struggle to differentiate deepfake texts, performing only slightly better than random guessing. Even with additional training with examples and instructions, the performance gain was limited (Clark et al. 2021). While much of the existing literature has focused on individual-based detection, the role of collaborative problem-solving in tackling deepfake text remains underexplored. This gap is particularly consequential given that detecting deepfake content is a cognitively demanding task that extends beyond surface-level recognition. It requires individuals to engage in sophisticated reasoning and critically evaluate ambiguous or potentially deceptive information. Uchendu et al. (2023) provide a first step by demonstrating that group collaboration can improve detection accuracy by 10 to 15% over individual performance. This naturally opens up a new question: *what makes group deliberation effective in such cognitively demanding, adversarial settings—and how might we further support it, particularly with the help of AI systems?*

Facilitating constructive and balanced discussion can be challenging due to several factors. For example, not all people are willing to actively participate in discussions. Additionally, some people may solely seek information consistent with their own perspectives, which can make it difficult for them to understand or respect others’ contrasting viewpoints (Stromer-Galley and Muhlberger 2009). Human moderators can play an important role in bridging the aforementioned gaps. Yet, due to the synchronous nature of online chat, moderators face a high managerial overhead in tasks like discussion stage management, opinion summarization, and consensus-building support. The research community has attempted to develop systems that can assist human moderation (Lee et al. 2020) or an artificial moderator that can completely replace humans’ involvement (Kim et al. 2021; Karadzhov 2024). Not limited to moderation, a system to support reasoned argumentation (Drapeau et al. 2016) and a consensus-building (Shin et al. 2022) have also been explored.

Building on this line of work, we integrate a deliberation-

enhancing dialogue agent into group discussions and investigate its role in the domain of deepfake text detection—a high-stakes task where reasoning and collaboration are critical. Rather than moderating or classifying content, our bot prompts reflection, encourages balanced participation, and supports group reasoning dynamics. Through this setup, we aim to address three research questions (RQs): (1) **RQ1**: How does a deliberation-enhancing bot affect the individuals’ performance of deepfake text detection?; (2) **RQ2**: How does the involvement of a deliberation-enhancing bot affect collaboration dynamics (e.g., engagement, even participation, consensus formation, probing dynamics and change of minds)?; (3) **RQ3**: What conditions make group collaboration with a deliberation-enhancing bot effective?.

In this work, we first curate a set of 14 articles, each consisting of three paragraphs: two written by humans and one generated by GPT-2 (Radford et al. 2019) or GPT-3.5. We also introduce DeepFakeDeLiBot, a deliberation-enhancing bot for deepfake text detection, which is built upon the dialogue system presented by Karadzhov (2024). DeepFakeDeLiBot (Karadzhov 2024) prompts users with questions that foster collaborative discussion *without* providing task-specific solutions or knowledge. This design allows us to isolate and analyze the effects of the deliberation process itself, free from interference by the bot offering correct answers. Our experiments are in two folds; the first stage is where 49 participants solve the detection tasks individually and the second stage is where participants form a group and solve the questions collectively. We employ a between-subjects design where 10 groups have DeepFakeDeLiBot involved and the remaining does not. We conduct a statistical comparative analysis of detection performance across three different setups (solo vs. group without DeepFakeDeLiBot vs. group with DeepFakeDeLiBot).

The results of our experiments consistently highlight the superior performance of group problem-solving compared to individual problem-solving. Although the improvement of groups that also interacted with DeepFakeDeLiBot was numerically higher, this improvement was not statistically significant. Groups interacting with DeepFakeDeLiBot exhibited more positive group dynamics, including higher engagement levels, more even participation, better consensus formation, and enhanced probing qualities. Our analysis suggests several conditions and contexts under which DeepFakeDeLiBot promotes performance gains via deliberation, examining this through the lens of participants’ backgrounds, group dynamics, and the bot’s interaction patterns.

To summarize, our contributions are as follows: (1) We present the first deliberation-enhancing conversational agent specifically designed for deepfake text detection; (2) Our statistical analysis reveals that groups outperform individuals in deepfake text detection. Additionally, groups interacting with DeepFakeDeLiBot demonstrated more positive group dynamics, suggesting that DeepFakeDeLiBot can enhance deliberation effectiveness without compromising detection performance; (3) We explore the effect of a variety of features and the involvement of DeepFakeDeLiBot on performance gain.

Related Work

Human Evaluation of Deepfake Text

As generative models become more capable of producing coherent, contextually appropriate, and convincing text, distinguishing between human and machine-generated content has become challenging. This has led to a growing interest in understanding how human evaluators assess the authenticity and credibility of deepfake text. According to Garbacea et al. (2019), evaluators could detect reviews generated by Word LSTM and GAN models with 66.61% accuracy. More recent works such as Ippolito et al. (2020) and Ippolito et al. (2020) re-evaluated humans’ detection performance on modern LLMs including GPT-2 and GPT-3 and found that their performance was slightly better than random guessing. The authors further attempted to train the evaluators by providing detailed instructions or walking through the task together, but there was a minor performance gain.

While the majority of prior works have framed the task as a binary classification—determining whether an entire text is generated by humans or machines—Dugan et al. (2023) and Uchendu et al. (2023) are among the first to explore human detection of the transition point where authorship switches from human to machine. Specifically, Dugan et al. (2023) demonstrated that framing the deepfake text detection task as a Real or Fake Text (RoFT) detection game, introduced by Dugan et al. (2020), enables participants to achieve an accuracy of 72.3%. In contrast, Uchendu et al. (2023) focused on how collaborative decision-making impacts detection performance. Building on these findings, this study investigates the role of a dialogue agent specialized for driving effective group deliberation in enhancing detection performance and discussion quality.

AI-Assisted Group Deliberation

Research has shown that AI-powered dialogue agents are capable of supporting group decision-making and deliberation across domains without the need for human intervention (Sahab, Haqbeen, and Ito 2024; Kim et al. 2021). By leveraging advanced machine learning and natural language processing techniques, these agents seamlessly drive group discussions by encouraging participation, posing insightful questions, and summarizing key discussion outcomes (Agarwal, Shahid, and Vashistha 2024; Kim et al. 2020). One standout example is DeLiBot (Deliberation Enhancing Bot), developed by Karadzhov, Stafford, and Vlachos (2023). Unlike traditional systems, DeLiBot is built to foster constructive group deliberation through strategic probing, often delivered in the form of three (moderation/solution/reasoning) different types of probing questions. Specifically, the bot monitors the dialogue histories, and tracks group dynamics, conversational patterns, and participant interactions to identify optimal moments for intervention. Upon determining the need for intervention, DeLiBot responds with tailored prompts designed to stimulate deeper reflection and improve group performance. The authors suggest that groups engaging with the bot achieved better solutions collectively than individuals could on their own when solving the Wason card selection task. To the best of our knowledge, there is no prior

work that investigated how AI-assisted group deliberation affects deepfake detection performance.

Methodology

Deepfake Data Curation

Prior literature (e.g., Tulchinskii et al. (2024), Mitchell et al. (2023), Mireshghallah et al. (2024)) primarily focused on the detection of sentences or paragraphs solely composed by LLMs. Yet, in the real-world setting, it is more likely that humans employ the generative models to improve the quality of certain parts of their draft (Lingard 2023). Also, they may amend or replace portions of their written content and evade LLM-text detectors (Sadasivan et al. 2023). Hence, in this study, we perform synthetic data generation where the articles comprise two paragraphs authored by humans and one paragraph generated by the LLM. This design is grounded on Uchendu et al. (2023), and we started with their dataset. Specifically, the authors selected 50 human-written news articles (mostly from the politics domain) and randomly replaced one out of three paragraphs with artificial texts written by GPT-2. We selected news article, particularly in the political domain, because they represent a high-stakes setting where deepfake texts can cause significant societal harm, such as spreading misinformation, shaping public opinion, or destabilizing democratic processes (Barari, Lucas, and Munger 2025; Gosse and Burkell 2020). Moreover, political content tends to elicit stronger reasoning, skepticism, and activation of prior beliefs compared to more neutral domains like weather reports or product reviews (Taber and Lodge 2006), making it a valuable testbed for studying the dynamics of collective judgment.

GPT-2 was released in early 2019 and is estimated to be 100 times smaller than their newer models like GPT-3.5. We hypothesize that texts written by GPT-2 are prone to make writing errors and hence easier to detect from humans’ lens than those from larger models due to scaling laws in generation capabilities (Kaplan et al. 2020). To ensure that our experiment reflects the current state of LLMs and possesses a good balance in task difficulties, we attempt to regenerate half of the dataset using GPT-3.5 and include them in our experiment.¹

Our primary goal in dataset curation is to ensure a balanced mix of *easy* and *challenging* questions in the final set. To gauge the difficulty levels of each article, we leverage three SOTA LLMs (GPT-3.5, LLAMA2-70B-chat, and Claude-2) as judges and feed the questions to the prompt. If all models considered a correct response to the question, we consider it to be an *easy* question. Out of the 50 questions in Uchendu et al. (2023)’s dataset, 16 questions were answered correctly by three LLM evaluators. We consider these 16 questions as *easy* questions. To construct *challenging* articles, We use GPT-3.5 to regenerate the remaining 32 questions. Here we follow the “fill-in-the-blank” prompting approach, where the model was asked to fill in the empty paragraph slot given the original article title and two human-

¹At the time of the generation, the most recent models including GPT-4 and GPT-4o were not released.

authored paragraphs. For instance, the prompt template to replace the second paragraph is illustrated below:

Given the title and two paragraphs of news articles, write Paragraph 2 on your own.

Article Title: $\${title}$
Paragraph 1: $\${paragraph_1}$
Paragraph 2:
Paragraph 3: $\${paragraph_3}$

Refer to Table 7 for generation examples. After the completion of data curation, we manually inspected GPT-3.5’s generation results and filtered out samples lacking consistency and coherence. This resulted in 28 questions out of 32 questions. Additionally, we automatically measured the coherence of three paragraphs as a whole by computing the paragraph-level cosine similarity after encoding the paragraphs with the T5 model (Raffel et al. 2020) and confirmed that all scores were above 0.8.

Implementation of DeepFakeDEliBot

In the original implementation of DeLiBot (Karadzhov 2024), probing utterances were retrieved from DeliData (Karadzhov, Stafford, and Vlachos 2023), a dataset of group discussions focused on the Wason Card Selection Task (Wason 1968). While DeliData provides general probing utterances, many are overly specific to this task, limiting their applicability to other domains. Additionally, its approach to generation, which involves replacing user and choice mentions in retrieved utterances using mask-filling, restricts flexibility and accuracy. To address these limitations and adapt DeLiBot to our domain, we collected and manually annotated domain-specific data and introduced an additional natural language generation component to produce probing utterances that better reflect the current conversation.

Inference. DeepFakeDeLiBot operates as a retrieval-based dialogue agent, maintaining a database of paired dialogue histories and probing utterances. During a conversation, it tracks the dialogue history and retrieves probing utterances through similarity-based retrieval. The retrieval context comprises the 5 utterances preceding the probing utterance, as specified by a hyperparameter. A Sentence-T5 language model computes embeddings for the current context, and cosine similarity is used to compare these embeddings with those in the database (Ni et al. 2021). DeepFakeDeLiBot selects the top 5 probing utterances with the most similar contexts and identifies the one expected to yield the greatest performance gain, estimated using an off-the-shelf Tactics-Strategy classifier. The selected utterance is then refined by a fine-tuned Flan-T5 (Longpre et al. 2023) Base language model (Chung et al. 2024), generating a probing utterance tailored to the conversation history with accurate participant and choice mentions.

Training. To adapt DeepFakeDeLiBot for the domain of DeepFake debunking, we expanded and refined the dataset of probing utterances, as summarized in Table 5. First, we

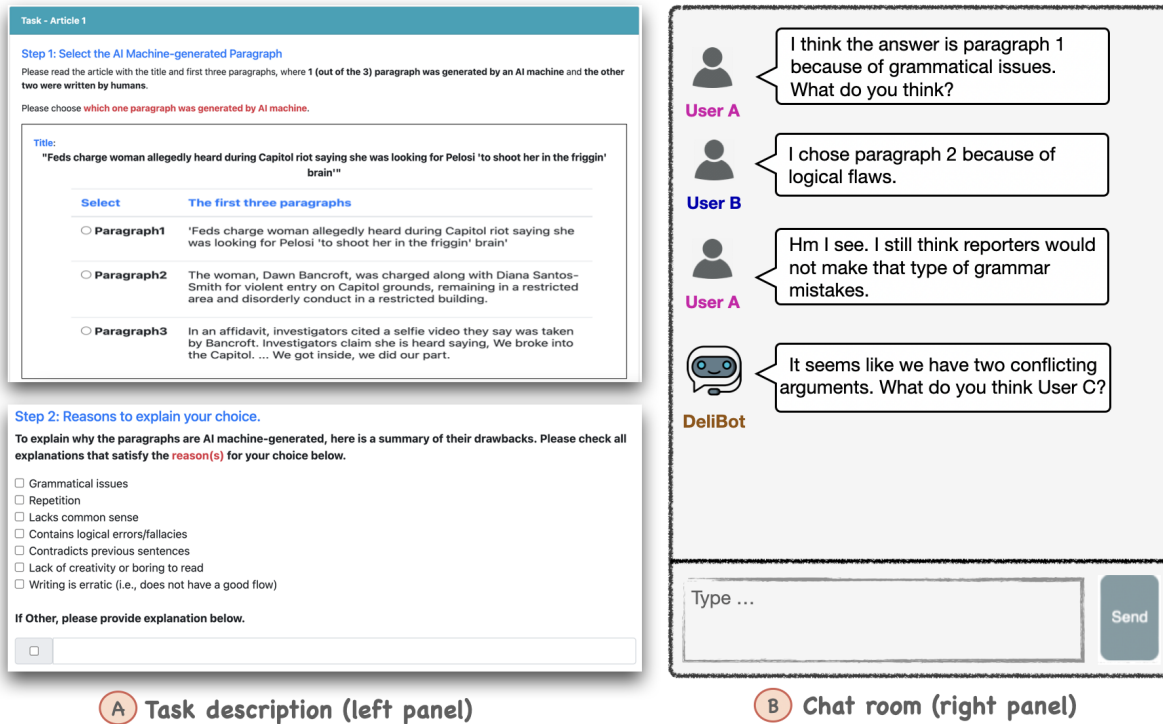


Figure 1: User interface example.

In-Context Learning Example
<p>Example Context: Lion: I chose 3, but looking over it now to see why I chose that Zebra: Actually 3 doesn't flow, reasonable effort to win election as many votes as I could Lion: I see why I chose 3. I didn't see where it connected to the 1st 2 paragraphs. Zebra: I have to say, in my experience with ChatGPT, I don't see this type of error Dolphin: I believe it is paragraph 2 Zebra: I now think 3 is incorrect as it is not objectively relevant to topic and other paragraphs</p> <p>Example Retrieved Probing: Dolphin, can you please explain a little ?</p> <p>Example Modified Probing: Dolphin, can you please explain why you believe it is paragraph 2?</p>

Table 1: An in-context learning example for GPT-3.5-Turbo to generate synthetic dataset.

manually annotated 5 transcribed group conversations related to the DeepFake task, identifying and extracting relevant probing utterances. These were combined with a filtered subset of DeliData to create an initial dataset. Next, we conducted 10 pilot studies using DeepFakeDeLiBot with this dataset, manually annotating the resulting dialogues to further enrich the context data. Additionally, we leveraged GPT-3.5-Turbo to paraphrase retrieved utterances, augmenting the dataset with more domain-specific probing utterances.

To ensure that retrieved utterances are contextually ap-

propriate, including accurate references to specific users and choices, we fine-tuned a Flan-T5 Base model. This process was framed as a sequence-to-sequence task: given the dialogue context and a retrieved probing utterance, the model generates a modified probing utterance with correct references. For training data creation, we employed 3-shot in-context learning with GPT-3.5-Turbo to generate synthetic pairs of context and retrieved utterances. An example of a manually annotated in-context learning demonstration is provided in Table 1. The fine-tuned Flan-T5 model was trained over three epochs, achieving strong performance in

generating contextually appropriate probing utterances.

Human Study Design

Participant Recruitment. We used Upwork², one of the largest freelance websites that has skilled freelancers in diverse domains such as writing, design, and web development. To promote our experiment on Upwork, we registered as a client and posted our research objectives and task descriptions. We explicitly mentioned that this posting is for research purposes and attached the consent form to the job posting for review. In addition, the following requirements were highlighted in the post: (1) participants should be at least 18 years old and (2) participants should be fluent in English. All freelancers could view our job posting, and those who were willing to participate were asked to return a questionnaire. It included three questions we crafted to understand their backgrounds better: (1) What is the highest level of degree you have completed in school?; (2) Did you major in English or English Literature?; and (3) Describe your recent experience with similar projects. Once we reached a reasonable number of applications, we verified participants’ eligibility by checking their self-reported age, language, and education in the profile. We also examined their desired hourly wage, as it substantially varied (ranging from \$15 to \$100) depending on expertise and experience. To limit the influence of desired payment differences on the experiment results, we only hired participants requesting \$30-\$35 per hour, which resulted in a total of 49 participants. Lastly, we collected the signed consent form and activated the contracts. The contracts were required by default in the Upwork platform to guarantee that freelancers and clients agree upon clients’ requested pay rate and that clients compensate freelancers based on submitted hours through the Upwork system. Assuming an average task duration of three hours per participant, the estimated total compensation for all participants was \$7,350 USD.

Experiment Design. Our study consists of two stages with two pre- and post-study surveys. Each stage of our experiment contains 14³ articles. In the first stage, we ask 49 participants to solve the deepfake detection task on their own. Upon completion, they are redirected to the pre-study survey link where we ask questions related to their backgrounds and self-perceived performance (see Table 6). The second stage, on the other hand, gathers groups of randomly selected individuals (consisting of two to three people) and asks them to discuss their answers to the same questions from the first experiment. This resulted in 20 groups. We then randomly assigned half of the groups (n=10) to engage with DeepFakeDeLiBot. Among 49 participants, we have 25 participants in the DeepFakeDeLiBot setup, and the remaining 24 participants were asked to solve the questions without DeepFakeDeLiBot.

To support the synchronous discussion, we included a chat service in our web interface. All members are assigned

²<https://www.upwork.com>

³To avoid participant cognitive fatigue, we reduced the total number of questions by sampling 7 questions from GPT-2 generated articles and 7 questions from GPT-3.5 generated articles.

Setting	Mean Accuracy	<i>p</i>
Individual vs. Group (w/o DeepFakeDeLiBot)	45.83% vs. 54.76% (8.93% ↑)	0.0004
Individual vs. Group (w. DeepFakeDeLiBot)	48.86% vs. 57.43% (8.67% ↑)	0.0013
Group (w/o DeepFakeDeLiBot) vs. Group (w. DeepFakeDeLiBot)	54.76% vs. 57.43% (2.67% ↑)	0.482

Table 2: T-test results for individual detection performance.

anonymized user names and can freely discuss their choices and reasoning. We explicitly informed participants that they were free to submit their own individual responses, regardless of whether the group reached a consensus. Lastly, participants were prompted to complete the post-study survey that inquires about participants’ experiences with group collaboration and interaction with DeepFakeDeLiBot (see Table 6).

Results

RQ1: Deepfake Detection Performance Comparison

To perform the analysis in a fine-grained manner, we measure the correctness of their submitted responses at the article level and use it for statistical testing. Let’s say C_i is the correctness of the response for the i -th article where $C_i \in \{0, 1\}$, with 1 indicating a correct response and 0 indicating an incorrect response. For a given user, the responses are represented as $[C_1, C_2, C_3, \dots, C_{14}]$. We then compute the statistical power of their mean differences before and after group collaboration through a paired t-test. When comparing groups with or without DeepFakeDeLiBot, we instead perform the unpaired t-test since the two populations (groups without DeepFakeDeLiBot vs. groups with DeepFakeDeLiBot) are independent.

Table 2 reports mean performance differences and their statistical significance. The results show that collaborative problem-solving consistently achieves higher performance in deepfake text detection than independent problem-solving by 8%. Moreover, groups that interacted with the deliberation-enhancing bot achieved the highest detection performance. However, the observed performance difference was not found to be statistically significant compared to the groups without the bot.

RQ2: DeepFakeDeLiBot and Collaboration Dynamics

In our experiments, we assessed collaboration dynamics by utilizing engagement levels, even participation, consensus formation, discussion constructiveness, and probing utterance frequencies as key proxies. These features were computed based on exchanged utterances between group members and submitted responses. Specifically, we tracked the following items at the article level: participant engagement, even participation, consensus formation, the frequency of solution-driven probing utterances, the frequency

Features	T-Test			Linear Regression	
	Group w/o DeepFakeDeLiBot	Group w. DeepFakeDeLiBot	<i>p</i>	Coef	<i>p</i>
Participant engagement	4.92	6.15	0.006	-0.01	0.17
Even participation	0.04	0.03	0.26	-0.01	0.97
Consensus Formation	0.85	0.95	0.005	0.19	0.001
Solution Probing Frequency	1.44	1.04	0.34	1.05	0.19
Reasoning Probing Frequency	1.68	3.48	0.001	-0.48	0.44
Moderation Probing Frequency	6.51	7.26	0.47	0.08	0.93
Diversity of Discussed Solution	0.82	1.05	0.02	0.10	0.04
Diversity of Submitted Reasoning	2.76	2.75	0.92	0.06	0.008

Table 3: T-test results for collaboration dynamic comparison w.r.t. DeepFakeDeLiBot usage and linear regression results of collaboration dynamics and performance gain.

of reasoning-driven probing utterances, the frequency of moderation-driven probing utterances, the diversity of discussed solutions and reasoning. Their definitions and computational measurement approaches can be described as follows:

- **Participant engagement:** we measure the participants’ engagement levels in the discussion by computing the average number of utterances spoken per participant.
- **Even participation:** we measure whether group members contributed evenly to the discussion (not dominated by particular individuals) by computing the variance in the distribution of participants’ engagement rate.
- **Consensus formation:** we measure whether participants successfully formed a consensus by examining their submitted responses. If all group members submitted an identical response, we assign 1. Else, we assign 0.
- **Frequency of probing utterances:** we measure how often participants leveraged probing utterances to drive discussion by computing the percentage of three types of probing utterances (moderation (e.g., *Let’s discuss our initial solutions*), solution (e.g., *Why did you think it wasn’t paragraph 3?*), reasoning (e.g., *Are we going for paragraph 2?*) of all utterances. Here we use the Karadzhev, Stafford, and Vlachos (2023)’s finetuned classification model from DeliToolkit⁴.
- **Diversity of discussed solutions:** we measure how often changes of mind occur by tracking the participants’ solution mentions within the dialogue. To achieve this, we utilized a regular expression to extract their mentions of paragraphs from their utterances.
- **Diversity of submitted reasoning:** we aim to measure whether participants exchanged various justification categories throughout the discussion by counting the unique number of submitted reasoning types.

For this analysis, we excluded utterances generated by DeepFakeDeLiBot. We performed the unpaired t-test to validate whether the mean differences of these features between the two conditions (groups without DeepFakeDeLiBot vs. groups with DeepFakeDeLiBot) were statistically meaningful.

⁴<https://github.com/gkaradzhev/delitookit>

As shown in Table 3, the differences in participants’ engagement levels, consensus formation, the frequency of reasoning probing utterances, and the diversity of solutions considered were found to be statistically significant. In particular, groups that engaged with DeepFakeDeLiBot tended to have 1.25% higher interaction rates with their team members than groups without DeepFakeDeLiBot. The probability of reaching a consensus was also higher for groups with DeepFakeDeLiBot. Among the three types of probing utterances, participants from the Delbiot group exchanged reasoning-probing utterances more often than participants in groups without DeepFakeDeLiBot. The frequency of moderation-probing utterances was slightly higher for the groups with DeepFakeDeLiBot compared to the groups without DeepFakeDeLiBot. Yet, the differences were statistically insignificant. Decision-making processes were measured through discussed solutions and submitted reasoning. Although there was no meaningful difference in the number of submitted reasoning between the two groups, groups with DeepFakeDeLiBot were more likely to discuss more diverse solutions than groups without DeepFakeDeLiBot.

These findings naturally led us to the next question: *what is the relationship between collaboration dynamics and performance as a whole?*. To answer this question, we ran a linear regression model by setting 8 features as independent variables and group performance as a dependent variable. The dependent variable was represented by the percentage of group members answering the question correctly during the group session. For example, if two out of three people within the group submitted an accurate response, their performance as a whole is equivalent to 66.66%. Table 3 demonstrates the linear regression results. The model suggests that consensus formation, the diversity of discussed solutions, and submitted reasoning are strong predictors of performance gain. The coefficient of 0.1981 indicates that for every one-unit increase in consensus formation, the performance gain is expected to increase by 19.81% units, holding other factors constant. Similarly, the diversity of discussed solution and submitted reasoning have positive relationships with performance gain; for every one-unit increase in the diversity of discussed solution and reasoning, the performance gain is expected to increase by 10.5% and 6% units, respectively.

RQ3: What Conditions Make Group Deliberation With DeepFakeDeLiBot Effective?

This RQ is to identify the conditions or contexts where DeepFakeDeLiBot has the greatest positive impact. Factors such as participants' profiles, the number of group members, user interaction patterns, or the type of utterances provided by DeepFakeDeLiBot can impact the degree of performance boost resulting from the bot.

Participants' backgrounds and experiences. We first analyzed how participants' self-reported characteristics submitted through pre- and post-study surveys influence the performance gap. Specifically, we examine their self-perceived proficiency in writing (Q7), AI-powered tool usage levels (Q9), and their trust levels in AI-powered tools (Q10) from the pre-study survey. In addition, their self-perceived performance after group collaboration (Q2) and self-perceived effectiveness of group collaboration (Q3) from the post-study survey were studied. The performance gap was computed by subtracting the solo performance from the group performance. If the particular individual answered more questions correctly after the group session, the performance gap would be a positive value. If not, it would remain 0 or negative. To investigate the effect of these variables and chatbot usage on the performance gap, we run a linear regression. We also modeled interaction effects between independent variables in order to reveal whether the impact of DeepFakeDeLiBot is stronger, weaker, or even reversed under certain conditions. Our results (Table 8) indicate that neither the chatbot nor the individual predictors (Q7, Q9, etc.) have significant main effects. Yet, we find a positive and significant coefficient regarding interaction terms between Q3 and the involvement of DeepFakeDeLiBot (coefficient= 9.6432, $p = 0.034$). Overall, This suggests that for individuals who perceive group collaboration as effective, the chatbot adds substantial value to their performance.

Group Dynamics. Next, we examine the effect of features related to group dynamics (from RQ2) on the performance gain before and after the group discussion. As these features such as participant engagement, even participation, and consensus formation are calculated at the article level, we calculate the performance gap at the article level as well. This can be done by measuring the gap in the percentage of people who answered the question correctly. For instance, in article 1, if one out of three (i.e., 33.33%) participants submitted the correct response in the solo session and then everyone (i.e., 100%) submitted the correct response after the group session, the performance gap is 66.67%. We also modeled interaction effects between independent variables to reveal whether the impact of DeepFakeDeLiBot is stronger, weaker, or even reversed under certain conditions. Our results (Table 9) show that neither DeepFakeDeLiBot nor the individual group dynamic predictors have significant main effects on the performance. Yet, we find a negative and significant coefficient regarding interaction terms between the moderation probing utterance frequency and the involvement of DeepFakeDeLiBot (coefficient = -0.9091, $p = 0.05$). For each unit increase in moderation probing utterance percentage, the effect of chatbot involvement on performance

gain decreases by -0.9091 units. Alternatively, it means the relationship between chatbot involvement and performance gain becomes more negative as the user-generated moderation probing utterance percentage increases.

Participants' interaction with DeepFakeDeLiBot. We further delve into why certain groups achieve a noticeable performance boost among 10 groups that engaged with DeepFakeDeLiBot. Among 10 groups, 6 groups outperformed their initial solo performance. The remaining group demonstrated no gain or even regress from their original detection performance. We speculate that their interaction patterns with DeepFakeDeLiBot will differ between these two subsets. To validate this assumption, we investigate 5 various categories:

- **DeepFakeDeLiBot's engagement rates** : we measure DeepFakeDeLiBot's engagement levels in discussion by computing the percentage of DeepFakeDeLiBot's utterances out of all participants' utterances.
- **Frequency of DeepFakeDeLiBot-generated probing utterances** : we quantify how often DeepFakeDeLiBot generated three (moderation, solution, reasoning) types of probing utterances by computing the percentage of each type across DeepFakeDeLiBot's responses.
- **Participants' unresponsiveness to DeepFakeDeLiBot's responses** : we measure how often participants ignored DeepFakeDeLiBot's responses by modeling the conversation flow. Motivated by Uchendu et al. (2023), we leverage the pre-trained LLMs⁵' perplexity⁶ scores. Since our objective is to identify if participants respond to DeepFakeDeLiBot's utterances or not, we can take DeepFakeDeLiBot's utterance as the first sentence and consider the following sentence as the participants' response. For measurement, we first take the DeepFakeDeLiBot's utterance as baseline perplexity and check if the combined sequence's perplexity is lower than the baseline.
- **Lexical diversity of DeepFakeDeLiBot's utterances** : we measure the lexical diversity of DeepFakeDeLiBot's responses by counting the number of unique n-grams.
- **Semantic coherence of DeepFakeDeLiBot's utterances** : we measure how semantically coherent DeepFakeDeLiBot's responses are to the ongoing conversation by calculating the cosine similarity between the embedding vectors of previous turns and DeepFakeDeLiBot's response. For text encoding, we use the T5 model.

For statistical analysis, we first performed the t-test to test whether the means of these factors were significantly different between two subgroups (a group without performance gain vs. a group with performance gain).

As illustrated in Table 4, the means of DeepFakeDeLiBot's engagement rates, its generation frequency of

⁵Since the authors had to run the model in the local machine, we used the GPT-2 small model.

⁶Perplexity quantifies how well a language model predicts the next word in a sequence. Lower perplexity indicates better predictive confidence, which can signal whether the transition between two sentences flows naturally or not.

moderation-probing utterances, and lexical diversity are found to be statistically different. For instance, while DeepFakeDeLiBot's participation rate was 8% on average for groups with a performance gain, DeepFakeDeLiBot engaged less often (5%) within groups that did not have the performance boost. Additionally, DeepFakeDeLiBot more frequently generated moderation-driven probing utterances in groups that experienced performance gain compared to those without any gain. Reasoning-probing utterances were less frequent in groups with performance gain, but the difference was not statistically significant. In comparing dialogue quality through the lens of lexical diversity and semantic coherence, only the lexical diversity of DeepFakeDeLiBot tended to be higher within groups with performance gain than groups without performance gain. Unresponsiveness rates were moderately high (41%-44%) in both groups, but the mean difference was not statistically meaningful.

Subsequently, we conducted a linear regression analysis to determine whether these factors could predict performance gain before and after collaborative problem-solving with DeepFakeDeLiBot. Interestingly, despite observed differences in means, none of these features were found to be significant predictors of the measured performance gain.

Qualitative Discourse Analysis

To better understand why the chatbot's interventions were ineffective for certain groups, we employ an inductive qualitative approach, allowing codes and themes to emerge organically from the dialogue transcripts. The lead author iteratively reviewed the data and identified recurring patterns, which were then grouped into higher-order categories reflecting group responsiveness to DeepFakeDeLiBot. See Table 10 for the codebook and dialogue examples. Our qualitative analysis suggests that this outcome stems from four core limitations.

DeepFakeDeLiBot's temporal misalignment with conversation flow: In many low-performing groups, DeepFakeDeLiBot's prompts were poorly timed relative to the group's conversational flow or decision points. Rather than entering at moments of uncertainty or open debate, the bot often interjected after participants had already reached consensus or shifted to a different topic. In such cases, the bot's input was perceived as irrelevant or even disruptive to the group's flow. This misalignment diminished the influence of DeepFakeDeLiBot on group reasoning.

Participants' lack of receptiveness: Low-performing groups frequently exhibited a high rate of dismissal toward DeepFakeDeLiBot's prompts. Even when the bot issued reasonable, well-formed questions, participants often showed low contextual receptiveness and seldom treated the bot as a legitimate contributor to the dialogue. Without active uptake and collaborative elaboration, even well-crafted prompts had minimal downstream impact on collective reasoning. This pattern suggests that Delibot's effectiveness is shaped not only by the quality of its interventions, but also by the group's disposition toward external input and its overall orientation toward deliberative engagement.

Participants' accelerated task completion: In some

groups, participants moved through the task so quickly that Delibot had little opportunity to intervene. These groups prioritized rapid consensus over deliberative reasoning, often finalizing decisions before the bot's prompts could appear or be meaningfully engaged with. As a result, even well-timed or contextually relevant interventions were rendered ineffective—not because of their content, but because the conversational window had already closed. This dynamic reveals a structural mismatch between the pace of group interaction and the bot's ability to scaffold reasoning in real time.

Discussion and Conclusion

RQ1. While group collaboration improved deepfake detection performance, the involvement of the deliberation-enhancing bot did not further enhance performance.

Our findings that group deliberation significantly enhances deepfake detection performance compared to individual efforts. This improvement likely arises from the exchange of diverse perspectives and the ability to critically evaluate information (Fleenor 2006). According to the exit survey, participants noted that group collaboration provided a “second pair of eyes” from individuals with different professional backgrounds, boosting their confidence in responses. Moreover, working together helped them identify unusual patterns in deepfake content that might be missed by individuals working alone.

While groups with DeepFakeDeLiBot slightly outperformed those without it in terms of performance gain, the effect was not statistically significant. This finding challenges our expectations, as deliberation is typically linked to improved group outcomes in decision-making (Karadzhov, Stafford, and Vlachos 2023; Iaryczower, Shi, and Shum 2018). There are two possible explanations for our findings. First, some participants in the exit survey noted that DeepFakeDeLiBot felt redundant, as the group was already effectively using its collective knowledge and reasoning. This minimized reliance on the bot, reducing its impact on the discussion. Second, deepfake detection may require specialized skills beyond what group deliberation alone can offer. Uchendu et al. (2023) found that experts outperformed laypeople in detecting deepfake texts, suggesting that advanced linguistic skills matter. Future research should explore models that provide task-specific cues while supporting deliberation based on the group's progress.

RQ2. The deliberation-enhancing bot positively influences collaboration dynamics that are important factors in achieving high performance.

Although performance gain between groups with and without DeepFakeDeLiBot did not differ significantly, we hypothesize that DeepFakeDeLiBot may bring secondary benefits, particularly in enhancing collaboration dynamics within the group. Through comparative analyses, we observed that groups utilizing DeepFakeDeLiBot exhibited higher rates of participant engagement, increased consensus formation, more frequent usage of reasoning-probing utterances, and greater diversity in the solutions discussed during collaborative tasks, compared to groups that did not have DeepFakeDeLiBot. Similar findings have been re-

Features	T-Test			Linear Regression	
	Group w/o Performance Gain	Group w. Performance Gain	<i>p</i>	Coef	<i>p</i>
Engagement Rates	0.05	0.08	0.01	0.69	0.13
Solution Probing Frequency	0.0	0.01	0.27	0.48	0.14
Reasoning Probing Frequency	0.23	0.16	0.27	-0.08	0.55
Moderation Probing Frequency	0.26	0.55	0.0001	-0.08	0.48
Lexical Diversity	0.44	0.68	0.0001	0.008	0.95
Semantic Coherence	0.58	0.57	0.69	0.69	0.13
Unresponsiveness Rates	0.41	0.44	0.12	0.22	0.17

Table 4: T-test results for interaction patterns of DeepFakeDeLiBot w.r.t. group performance gain and linear regression results.

ported in studies in AI-assisted group decision-making processes (Kim et al. 2021; Shin et al. 2022; Dortheimer et al. 2024), where chatbots can foster more active and constructive dialogue.

Among these factors, consensus formation, the diversity of discussed solutions, and submitted responses emerged as critical predictors of group performance. Specifically, an increase in the strength of consensus was positively correlated with improved group performance. The importance of consensus building has been suggested in the strategic decision-making procedure (Dess and Origer 1987; Whyte 1989). Furthermore, it could promote a more deliberative process (Hare 1980) and can reduce the probability of errors, producing higher quality decisions (Davis et al. 1993; Feddersen and Pesendorfer 1998). Our finding is well-aligned with existing literature. We also found that a higher diversity of solutions and justification types submitted by participants was associated with increased group performance. This result aligns with the previous studies (e.g., Post et al. (2009), Hundschell et al. (2022)) such that diversity of thought stimulates innovation and problem-solving by encouraging the exploration of a broader range of possibilities.

Contrary to our initial expectations, neither the overall levels of participant engagement nor the distribution of engagement among group members demonstrated significant relationships with group performance. While engagement is often viewed as a key driver of group effectiveness (Yoerger, Crowe, and Allen 2015), our findings challenge this assumption in the context of the deepfake detection task. They suggest that simply increasing engagement levels or ensuring equitable participation may not be sufficient to improve performance. Similarly, the generation of probing utterances, often expected to enhance group outcomes by stimulating deeper analysis, did not demonstrate a significant impact on group performance.

RQ3. Participants’ backgrounds, experiences, and interaction patterns differ significantly based on performance gain.

Several studies (Liu, Joy, and Griffiths 2010; Huang 2018) argued that group composition (e.g., the members’ backgrounds and beliefs) and group interaction patterns shape the success of the group decision-making. For example, in the context of deepfake text detection, prior literature such as Clark et al. (2021); Uchendu et al. (2023) reported that ex-

perts in writing domains could identify deepfake texts more accurately than laypeople, as they are more knowledgeable to capture unnatural flow or subtle topical changes often found in deepfake texts. According to our analyses, participants’ background information, including their proficiency in writing, experiences in AI tools, and their trust in these tools, were not significant predictors of performance gain. We also examined relationships between their self-perceived detection performance and the effectiveness of group collaboration and performance gain. Our result reveals that, for individuals who perceived group collaboration as effective, the involvement of DeepFakeDeLiBot can positively impact their performance. This indicates that, in groups where collaboration is already perceived as smooth and effective, DeepFakeDeLiBot can further amplify these dynamics to achieve even greater performance outcomes. However, in groups where collaboration was deemed unhelpful, DeepFakeDeLiBot’s impact may be shadowed by group-specific challenges.

We empirically validated that, depending on the degree of performance gain, DeepFakeDeLiBot’s engagement rates, its generation frequency of moderation-probing utterances, and lexical diversity differed significantly. Still, these factors did not strongly correlate with predicting performance improvements. Our findings in the discourse analysis suggests that the effectiveness of the bot is not uniform, but contingent on participants’ contextual receptiveness and the bot’s interjection timing. In groups that were already deliberating effectively, Delibot’s prompts were more likely to be integrated and acted upon.

Limitations

Our study has two limitations to consider. First, our sample size is relatively small, primarily due to difficulties in recruiting qualified individuals on the Upwork platform and the associated costs of compensation. Moreover, the experiment required synchronous communication, further constraining the sample size due to practical considerations such as participant availability and scheduling. These factors may have limited the statistical power of our analyses. However, given the exploratory nature of this research, our primary aim was to identify trends and relationships rather than to generalize findings to a larger population. Future studies with larger samples can help validate and extend these find-

ings. Second, the generalizability of our findings regarding detection performance may be limited, as we exclusively used two of OpenAI’s LLMs to generate deepfake paragraphs. Future research should examine how a broader range of model families, such as Meta’s Llama or Google’s Gemini, influence human detection performance.

Ethical Statement

Our research protocol was approved by the Institutional Review Board (IRB) at our institution. We exclusively recruited participants aged 18 years or older and ensured they were fully informed about the nature of the study. We also explicitly notified them that the dialogue histories and survey responses collected during the study would be shared publicly upon manuscript acceptance. To safeguard privacy, any personally identifiable information, such as names and email addresses, will be removed, and participants will be assigned numerical identifiers to ensure anonymity. In line with the FAIR principles (Findable, Accessible, Interoperable, and Reusable), the data will be managed by Github.

Before the experiment, we explicitly informed participants that the presented articles, including one of the three paragraphs, contained deepfake content. Consequently, we believe that exposure to these news articles with deepfake paragraphs is unlikely to have negatively influenced the participants. Participants were compensated regardless of whether they completed the entire task and were compensated at rates exceeding the minimum wage.

Acknowledgments

This work was in part supported by NSF awards #1820609, #2114824, and #2131144, and PSU CSRE seed grant 2023. We appreciate our study participants for delivering the results. We also thank the anonymous reviewers for their valuable feedback.

References

Agarwal, D.; Shahid, F.; and Vashistha, A. 2024. Conversational Agents to Facilitate Deliberation on Harmful Content in WhatsApp Groups. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–32.

Barari, S.; Lucas, C.; and Munger, K. 2025. Political deepfakes are as credible as other fake media and (sometimes) real media. *The Journal of Politics*, 87(2): 510–526.

Barman, D.; Guo, Z.; and Conlan, O. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 100545.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Clark, E.; August, T.; Serrano, S.; Haduong, N.; Gururangan, S.; and Smith, N. A. 2021. All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings*

of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7282–7296. Online: Association for Computational Linguistics.

Davis, J. H.; Stasson, M. F.; Parks, C. D.; Hulbert, L.; Kameda, T.; Zimmerman, S. K.; and Ono, K. 1993. Quantitative decisions by groups and individuals: Voting procedures and monetary awards by mock civil juries. *Journal of Experimental Social Psychology*, 29(4): 326–346.

Dess, G. G.; and Origer, N. K. 1987. Environment, structure, and consensus in strategy formulation: A conceptual integration. *Academy of Management Review*, 12(2): 313–330.

Dortheimer, J.; Martelaro, N.; Sprecher, A.; and Schubert, G. 2024. Evaluating large-language-model chatbots to engage communities in large-scale design projects. *AI EDAM*, 38: e4.

Dou, Y.; Forbes, M.; Koncel-Kedziorski, R.; Smith, N. A.; and Choi, Y. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7250–7274. Dublin, Ireland: Association for Computational Linguistics.

Drapeau, R.; Chilton, L.; Bragg, J.; and Weld, D. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 32–41.

Dugan, L.; Ippolito, D.; Kirubarajan, A.; and Callison-Burch, C. 2020. RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 189–196. Online: Association for Computational Linguistics.

Dugan, L.; Ippolito, D.; Kirubarajan, A.; Shi, S.; and Callison-Burch, C. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12763–12771.

Feddersen, T.; and Pesendorfer, W. 1998. Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review*, 92(1): 23–35.

Fleener, J. W. 2006. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies and nations. *Personnel Psychology*, 59(4): 982.

Garbacea, C.; Carton, S.; Yan, S.; and Mei, Q. 2019. Judge the Judges: A Large-Scale Evaluation Study of Neural Language Models for Online Review Generation. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Confer-*

- ence on Natural Language Processing (EMNLP-IJCNLP), 3968–3981. Hong Kong, China: Association for Computational Linguistics.
- Gosse, C.; and Burkell, J. 2020. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, 37(5): 497–511.
- Hare, A. P. 1980. Consensus versus majority vote: A laboratory experiment. *Small Group Behavior*, 11(2): 131–143.
- Hu, X.; Chen, P.-Y.; and Ho, T.-Y. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36: 15077–15095.
- Huang, C. Y. 2018. How background, motivation, and the cooperation tie of faculty members affect their university–industry collaboration outputs: an empirical study based on Taiwan higher education environment. *Asia Pacific Education Review*, 19(3): 413–431.
- Hundscheil, A.; Razinkas, S.; Backmann, J.; and Hoegl, M. 2022. The effects of diversity on creativity: A literature review and synthesis. *Applied Psychology*, 71(4): 1598–1634.
- Hutson, J. 2024. Rethinking Plagiarism in the Era of Generative AI. *Journal of Intelligent Communication*, 4(1): 20–31.
- Iaryczower, M.; Shi, X.; and Shum, M. 2018. Can words get in the way? The effect of deliberation in collective decision making. *Journal of Political Economy*, 126(2): 688–734.
- Ippolito, D.; Duckworth, D.; Callison-Burch, C.; and Eck, D. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1808–1822. Online: Association for Computational Linguistics.
- Jakesch, M.; Hancock, J. T.; and Naaman, M. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11): e2208839120.
- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Karadzov, G. 2024. *DEliBots: Deliberation Enhancing Bots*. Ph.D. thesis.
- Karadzov, G.; Stafford, T.; and Vlachos, A. 2023. DeliData: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–25.
- Kim, S.; Eun, J.; Oh, C.; Suh, B.; and Lee, J. 2020. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.
- Kim, S.; Eun, J.; Seering, J.; and Lee, J. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–26.
- Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, 12–24.
- Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; and Iyyer, M. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.
- Lee, J.; Le, T.; Chen, J.; and Lee, D. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, 3637–3647.
- Lee, S.-C.; Song, J.; Ko, E.-Y.; Park, S.; Kim, J.; and Kim, J. 2020. Solutionchat: Real-time moderator support for chat-based structured discussion. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–12.
- Li, Y.; Li, Q.; Cui, L.; Bi, W.; Wang, Z.; Wang, L.; Yang, L.; Shi, S.; and Zhang, Y. 2024. Mage: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 36–53.
- Lingard, L. 2023. Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on medical education*, 12(1): 261.
- Liu, S.; Joy, M.; and Griffiths, N. 2010. Students’ perceptions of the factors leading to unsuccessful group collaboration. In *2010 10th IEEE International Conference on Advanced Learning Technologies*, 565–569. IEEE.
- Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 22631–22648. PMLR.
- Lucas, J.; Uchendu, A.; Yamashita, M.; Lee, J.; Rohatgi, S.; and Lee, D. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14279–14305. Singapore: Association for Computational Linguistics.
- Mireshghallah, N.; Mattern, J.; Gao, S.; Shokri, R.; and Berg-Kirkpatrick, T. 2024. Smaller Language Models are Better Zero-shot Machine-Generated Text Detectors. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 278–293. St. Julian’s, Malta: Association for Computational Linguistics.
- Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 24950–24962. PMLR.
- Ni, J.; Abrego, G. H.; Constant, N.; Ma, J.; Hall, K. B.; Cer, D.; and Yang, Y. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.
- Post, C.; De Lia, E.; DiTomaso, N.; Tirpak, T. M.; and Borwankar, R. 2009. Capitalizing on thought diversity for innovation. *Research-Technology Management*, 52(6): 14–25.

- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.
- Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2023. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.
- Sahab, S.; Haqbeen, J.; and Ito, T. 2024. Conversational AI as a Facilitator Improves Participant Engagement and Problem-Solving in Online Discussion: Sharing Evidence from Five Cities in Afghanistan. *IEICE TRANSACTIONS on Information and Systems*, 107(4): 434–442.
- Shin, J.; Hedderich, M. A.; Lucero, A.; and Oulasvirta, A. 2022. Chatbots facilitating consensus-building in asynchronous co-design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–13.
- Stromer-Galley, J.; and Muhlberger, P. 2009. Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy. *Political communication*, 26(2): 173–192.
- Taber, C. S.; and Lodge, M. 2006. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3): 755–769.
- Tulchinskii, E.; Kuznetsov, K.; Kushnareva, L.; Cherniavskii, D.; Nikolenko, S.; Burnaev, E.; Barannikov, S.; and Piontkovskaya, I. 2024. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36.
- Uchendu, A.; Lee, J.; Shen, H.; Le, T.; Lee, D.; et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, 163–174.
- Uchendu, A.; Ma, Z.; Le, T.; Zhang, R.; and Lee, D. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2001–2016. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Venkatraman, S.; Uchendu, A.; and Lee, D. 2024. GPT-who: An Information Density-based Machine-Generated Text Detector. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 103–115. Mexico City, Mexico: Association for Computational Linguistics.
- Wang, P.; Li, L.; Ren, K.; Jiang, B.; Zhang, D.; and Qiu, X. 2023. SeqXGPT: Sentence-Level AI-Generated Text Detection. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1144–1156. Singapore: Association for Computational Linguistics.
- Wang, Y.; Mansurov, J.; Ivanov, P.; Su, J.; Shelmanov, A.; Tsvigun, A.; Whitehouse, C.; Mohammed Afzal, O.; Mahmoud, T.; Sasaki, T.; Arnold, T.; Aji, A.; Habash, N.; Gurevych, I.; and Nakov, P. 2024. M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1369–1407. St. Julian’s, Malta: Association for Computational Linguistics.
- Wason, P. C. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3): 273–281.
- Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šigut, P.; and Waddington, L. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1): 26.
- Whyte, G. 1989. Groupthink reconsidered. *Academy of Management Review*, 14(1): 40–56.
- Yoerger, M.; Crowe, J.; and Allen, J. A. 2015. Participate or else!: The effect of participation in decision-making in meetings on employee engagement. *Consulting Psychology Journal: Practice and Research*, 67(1): 65.

Paper Checklist

1. Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
2. Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes**
3. Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes**
5. Did you describe the limitations of your work? **Yes**
6. Did you discuss any potential negative societal impacts of your work? **Yes**
7. Did you discuss any potential misuse of your work? **Yes**
8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**
9. Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
10. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes**
11. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Not applicable, we examined generation quality manually.**
12. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes**

13. Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes**
14. Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **Yes**
15. If your work uses existing assets, did you cite the creators? **Yes**
16. Did you mention the license of the assets? **No, because the license of the assets are mentioned within the cited paper**
17. Did you include any new assets in the supplemental material or as a URL? **No, because we mentioned that the dataset and source code will be made publicly available upon acceptance of the manuscript.**
18. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Yes**
19. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes**
20. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **Yes**
21. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? **Not yet, as it is a simple and small dataset and will not be publicly shared upon acceptance. However, when we release it publicly, we will create one.**
22. Did you include the full text of instructions given to participants and screenshots? **Yes**
23. Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **Yes**
24. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **Yes**
25. Did you discuss how data is stored, shared, and deidentified? **Yes**

model was trained for 3 epochs, completing within approximately 1 GPU hour, and achieved satisfactory performance.

Appendix

A. DeepFakeDeLiBot Training Details

Data	Dialogues	Avg. Turns	Avg. Probing
DeliData	500	28	3.488
Transcribed	5	1044.0	114.4
Pilot	10	224.3	37.2

Table 5: Summary of datasets with statistics on dialogues, average turns, and average probing.

For fine-tuning the Flan-T5 Base model, we constructed a synthetic dataset using GPT-3.5-Turbo with in-context learning examples. After manually reviewing and filtering 500 generated data points, we retained 371 for training, 46 for validation, and 47 for testing. Learning rate was set to $5e-5$ with minibatch size 2 and 0.01 weight decay using Adam optimizer. The fine-tuning process was conducted on a single Quadro RTX 8000 GPU with 48 GB of memory. The

Survey Type	Questions
Pre-Study	<ul style="list-style-type: none"> - What is your gender? - Which category below includes your age? - Which race/ethnicity best describes you? - What is the highest level of school you have completed? - Please briefly describe your occupation in one or two sentences. - Please rate your self-perceived proficiency in writing on a scale of 1 to 5, with 1 being not proficient at all and 5 being highly proficient. - Have you worked on a similar project before? If not, please insert "N/A". Else, please describe. - Have you ever used AI-powered tools before? If so, how often do you use them? - On a scale of 1 to 5, with 1 being "Not Trusting at All" and 5 being "Highly Trusting", how much would you say you trust AI-powered tools in general? - On a scale of 1 to 5, with 1 being "Very Easy" and 5 being "Very Difficult", please rate the overall difficulty level of this task. - On a scale of 1 to 5, with 1 being "Poorly" and 5 being "Exceptionally Well", how well do you believe you performed this task?
Post-Study	<ul style="list-style-type: none"> - How well do you believe you performed this task? - To what extent did group collaboration benefit your ability to accomplish the task? - In what ways did group collaboration benefit your ability to accomplish the task? - In what ways did group collaboration not benefit your ability to accomplish the task? - During the experiment, did you engage with DEliBot (i.e., a deliberation enhancing dialogue agent)? - How would you rate the quality of Delibot's probing utterances?- How would you describe your overall experience with Delibot's engagement frequency during your interactions? - If you interacted with DEliBot during the experiment, to what extent did the DEliBot benefit the group collaboration? - On a scale of 1 to 5, with 1 being "Not Trusting at All" and 5 being "Highly Trusting", how much would you say you trust DEliBot? - Kindly provide any suggestions you may have for improving DEliBot.

Table 6: Questions for pre-study and post-study surveys

Model	Title	Paragraph Number	Content
GPT-2	Andrew Cuomo's Covid-19 performance may have been less stellar than it seemed	P1	If Donald Trump was seen as the public face of the failed government response to the coronavirus pandemic, Andrew Cuomo was seen by some as the opposite – a politician who understood the myriad challenges created by Covid-19 and moved quickly to address them in the most transparent way possible.
		P2	One day, Cuomo took the podium at a state event at the hospital where an Ebola patient was being treated.
		P3	It was, for many, a refreshing palate cleanser from the obfuscation, spin and denialism that defined how Trump and his administration responded to the virus through the spring and summer of 2020.
GPT-3.5	'People are angry': House Republicans who voted to impeach Trump face backlash at home	P1	House Republicans who voted to impeach former President Donald Trump are facing intense backlash from GOP voters in their home districts, putting their 2022 primaries in jeopardy. The backlash highlights the continued influence of Trump in Republican politics and raises questions about the loyalty of GOP voters.
		P2	The backlash has turned their 2022 primaries into tests of how long Trump can hold the stage in Republican politics and whether GOP voters are willing to turn the midterms into tests of loyalty to him.
		P3	The group of 10 Republicans includes moderates in swing districts, as well as some reliable conservatives, including the No. 3-ranking House Republican, Wyoming Rep. Liz Cheney, and South Carolina Rep. Tom Rice

Table 7: Deepfake article example. Text in blue indicates a paragraph written by LLMs.

Main Features	Coef	<i>p</i>	Interaction Features	Coef	<i>p</i>
Q7	0.66	0.89	Q7_DeepFakeDeLiBot	3.88	0.54
Q9	4.37	0.21	Q9_DeepFakeDeLiBot	2.53	0.54
Q10	7.35	0.14	Q10_DeepFakeDeLiBot	0.54	0.96
Q2	0.26	0.96	Q2_DeepFakeDeLiBot	2.02	0.81
Q3	3.34	0.27	Q3_DeepFakeDeLiBot	9.64	0.03

Table 8: Linear regression results of participants' background/experiences and detection performance gain. Investigated features include: self-perceived proficiency in writing (Q7), AI-powered tool usage levels (Q9), and their trust levels in AI-powered tools (Q10), their self-perceived performance after group collaboration (Q2), and self-perceived effectiveness of group collaboration (Q3).

Main Features	Coef	<i>p</i>	Interaction Features	Coef	<i>p</i>
Participant engagement	-0.007	0.47	Participant engagement_DeepFakeDeLiBot	0.005	0.67
Even participation	-0.26	0.56	Even participation_DeepFakeDeLiBot	-0.39	0.59
Consensus formation	0.006	0.89	Consensus formation_DeepFakeDeLiBot	0.05	0.55
Solution probing frequency	0.52	0.42	Solution probing frequency_DeepFakeDeLiBot	0.60	0.62
Reasoning probing frequency	-0.59	0.44	Reasoning probing frequency_DeepFakeDeLiBot	0.63	0.5
Moderation probing frequency	0.48	0.14	Moderation probing frequency_DeepFakeDeLiBot	-0.9	0.05
Diversity of discussed solutions	0.004	0.27	Diversity of discussed solutions_DeepFakeDeLiBot	-0.02	0.72
Diversity of submitted reasoning	0.006	0.76	Diversity of submitted reasoning_DeepFakeDeLiBot	-0.0007	0.98

Table 9: Linear regression results of group dynamics and detection performance gain.

Code Category	Description	Example Bot Utterance
Temporal Misalignment	Bot prompt is mistimed—occurs after group consensus or before relevant context is established.	Zebra: So we're all good with Paragraph 2? Penguin: Yep. Bot: What is your reason for choosing Paragraph 3? Zebra: We already picked 2... Penguin: Yeah, too late now.
Dismissal	Participants ignore or reject the prompt without elaboration.	Bot: What do you think about Paragraph 1? Penguin: Anyway, 3 is the weirdest one. Let's move on. Zebra: Sure.
Task Acceleration	Group moves too quickly for bot to intervene effectively.	Zebra: Let's just pick Paragraph 1. It's probably fake. Penguin: Agreed. Done. Zebra: What about for the next question? I chose 3. Penguin: Same, let's move on.
Contextual Receptiveness	Group engages with the prompt, elaborates or builds on the idea in meaningful ways.	Zebra: Paragraph 2 feels kind of off, but I can't tell why. Bot: What makes Paragraph 2 stand out to you? Penguin: It's more narrative—it reads like a story. Zebra: Yeah. Also, that paragraph talks about Ebola while the other two talk about Covid.
Prompt Relevance	Bot's utterance matches the group's current topic or reasoning trajectory.	Penguin: I chose Paragraph 1 as the answer due to some logical gaps. Zebra: I chose Paragraph 1 because of grammatical errors. Bot: Do we all agree that Paragraph 1 is the answer? Penguin: Yes! Zebra: Yup.
Prompt Quality	Prompt is clear, focused, and designed to elicit deeper reasoning.	Penguin: I think Paragraph 1 is trying too hard to be emotional. Bot: Can you further explain on that? Penguin: Yeah, it's full of dramatic language. That's not how news sounds. Zebra: Good catch. Real articles are usually more neutral.

Table 10: Dialogue annotation scheme and examples.