Does Human Collaboration Enhance the Accuracy of Identifying LLM-Generated Deepfake Texts?

Adaku Uchendu [*1,2], Jooyoung Lee[*2], Hua Shen[*2,4], Thai Le[3], Ting-Hao 'Kenneth' Huang[2], Dongwon Lee[2]
[1] MIT Lincoln Laboratory, USA
[2] The Pennsylvania State University, USA
[3] The University of Mississippi, USA
[4] University of Michigan, USA
adaku.uchendu@ll.mit.edu,
{jfl5838, huashen218, txh710, dongwon}@psu.edu,
thaile@olemiss.edu

Abstract

Advances in Large Language Models (e.g., GPT-4, LLaMA) have improved the generation of coherent sentences resembling human writing on a large scale, resulting in the creation of so-called *deepfake texts*. However, this progress poses security and privacy concerns, necessitating effective solutions for distinguishing deepfake texts from human-written ones. Although prior works studied humans' ability to detect deepfake texts, none has examined whether "collaboration" among humans improves the detection of deepfake texts. In this study, to address this gap of understanding on deepfake texts, we conducted experiments with two groups: (1) non-expert individuals from the AMT platform and (2) writing experts from the Upwork platform. The results demonstrate that *collaboration among humans can potentially improve the detection of deepfake texts for both groups*, increasing detection accuracies by 6.36% for non-experts and 12.76% for experts, respectively, compared to individuals' detection accuracies. We further analyze the explanations that humans used for detecting a piece of text as deepfake text, and find that the strongest indicator of deepfake texts is their lack of coherence and consistency. Our study provides useful insights for future tools and framework designs to facilitate the collaborative human detection of deepfake texts. The experiment datasets and AMT implementations are available at: https://github.com/huashen218/llm-deepfake-human-study.git

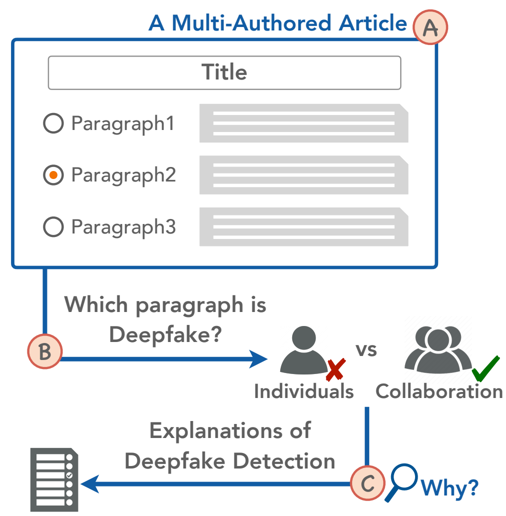

Figure 1: An overview of human studies on detecting deepfake texts. (A) A multi-authored article with 3 paragraphs, including both human-written & LLM-generated paragraphs; (B) We conduct human studies to ask either individuals or collaborative humans to detect deepfake texts; (C) In-depth analysis of the categorical explanations for deepfake text detection from both groups.

Introduction

In recent years, significant advancements in AI technologies have revolutionized the generation of high-quality artifacts across various modalities, including texts, images, and videos (Fagni et al. 2021; Zhang 2022; Pu et al. 2023; Shen and Wu 2023). These AI-generated artifacts, commonly referred to as Deepfakes, have garnered considerable attention. Specifically, the progress made in Natural Language Generation (NLG) techniques, leveraging Large Language Models (LLMs) like GPT-4 (OpenAI 2023) or T5 (Raffel et al. 2020), has facilitated the production of long and coherent machine-generated texts without human intervention (Wu et al. 2023). For the purpose of this study, we designate such neural or LLM-generated texts as **deepfake texts**, while the generative language models themselves are referred to as Neural Text Generators (NTG) (Zhong et al. 2020). While NTGs offer numerous benefits, it is essential to acknowledge the potential misuse associated with this technological advancement (Shevlane et al. 2023). For instance, NTGs can be employed by students to complete their essay assignments, leading to potential plagiarism due to NTGs' memorization of training samples (Lee et al. 2023). Moreover, scammers may exploit NTGs to craft sophisticated phishing messages, or stereotyping, misrepresenting, and demeaning content (Weidinger et al. 2021), while malicious code generation (Chen et al. 2021) and disinformation attacks by state-backed operators are also plausible scenarios (Bagdasaryan and Shmatikov 2022). Given these con-

[*]Equal contribution.

cerns, it becomes imperative to prioritize research efforts towards developing effective methodologies for distinguishing deepfake texts from those authored by humans.

Both computational and non-computational approaches for detecting deepfake texts have received significant attention in recent years (Uchendu et al. 2021; Clark et al. 2021; Dou et al. 2022; Brown et al. 2020), and have been comprehensively surveyed by Uchendu, Le, and Lee (2023). However, emerging literature (Uchendu et al. 2021; Dou et al. 2022) suggests that humans, on average, struggle to detect deepfake texts, performing only slightly better than random guessing. Even with training, the performance of humans in deepfake text detection has shown limited improvement (Clark et al. 2021; Dou et al. 2022; Tan, Plummer, and Saenko 2020). These findings highlight the need to explore alternative strategies, such as collaborative detection or leveraging advanced technological solutions, to address the challenges posed by deepfake texts effectively.

Online fact-checking efforts, as highlighted by Juneja and Mitra (2022), can be achieved collaboratively to detect online misinformation. Previous research has demonstrated that collective intelligence, often referred to as the "wisdom of the crowd", can surpass individual sensemaking capabilities (Surowiecki 2005). Similarly, aggregating multiple human labels has also been shown to yield higher-quality results (Zheng et al. 2017). However, limited attention has been given to understanding how collaboration affects the performance of deepfake detection. Consequently, the primary objective of this study is to **investigate the impact of human collaboration on the detection of deepfake texts**. See an overview of the task presented in Figure 1, wherein we generate a three-paragraph article authored by both humans and LLM. Individuals or collaborative human groups are then tasked with identifying the paragraph that has been generated by LLMs. Furthermore, we delve into the detailed explanations provided by humans to detect the deepfakes. It is worth noting that this deepfake detection design bears resemblance to the *Turing Test*.[1] As a result, our study focuses on addressing the following research questions:

- **RQ1:** Do collaborative teams or groups outperform individuals in deepfake text detection task?
- **RQ2:** What types of reasoning explanations are useful indicators for deepfake text detection?

To conduct comprehensive human studies on evaluating the effectiveness of human collaboration in deepfake text detection (*i.e.,* RQ1), we focus on two distinct stakeholder groups of online workers: Amazon Mechanical Turk (AMT) workers as English non-experts and Upwork workers as English experts. The term "English experts" refers to individuals who possess at least a Bachelor's degree in English or a related field (Please see the Methodology section for detailed filtering criteria for identifying experts). These two groups also represent the conventional micro-task crowdsourcing setting and the freelance marketplace setting, respectively. The next challenge is to facilitate human collaboration on these two platforms. For AMT workers, we have devised an asynchronous collaboration approach, while for Upwork workers, a synchronous collaboration method has been implemented (please refer to the Methodology section for more information on the implementation details). Furthermore, during the study, we request both groups to provide their explanations for detecting deepfake texts (*i.e.,* RQ2). They are given a predefined set of seven explanation types to choose from or the option to supplement their own explanations. By collecting these explanations, we aim to delve deeper into the reasoning process behind human collaborative deepfake text detection.

[1]Turing Test measures how human-like a model is. If a model shows intelligent behavior usually attributed to a human and is thus, labeled a human, the model is said to have passed the Turing Test.

Through the execution of two human studies and a comparative analysis of human collaborative and individual evaluations within both the expert and non-expert groups, our research reveals that **human collaboration has the potential to enhance the performance of deepfake text detection for both stakeholder groups**. The key findings of our study can be summarized as follows:

- Human collaboration leads to a 6.36% improvement in deepfake text detection among non-experts and a 12.76% improvement among experts;
- The detection of deepfake texts is influenced by indicators such as "consistency", "coherency", "common sense", and "self-contradiction" issues;
- Experts outperform non-experts in both individual and collaborative scenarios when it comes to detecting deepfake texts.

Overall, this work focuses on investigating the impact of human collaboration on the detection of deepfake texts and demonstrates that collaborative efforts within representative groups yield superior results compared to individuals. The study sheds light on the underlying reasoning explanations, highlights limitations, and emphasizes the need for the development of computational and non-computational (including hybrid) tools to promote more robust and accurate detection methods.

Related Work

Evaluating Deepfake Texts with Laypeople

The quality of deepfake texts has always been compared to human-written texts. Thus, since humans still remain the gold standard when evaluating machine-generated texts, several works have investigated human performance in distinguishing between human-written and machine-generated texts. GROVER (Zellers et al. 2019), an NTG trained to generate news articles can easily be used maliciously. To evaluate the quality of GROVER-generated news (fake) articles, they are compared to human-written news articles. Humans are asked to pick which articles are more believable and GROVER-generated fake news was found to be more trustworthy (Zellers et al. 2019). Donahue, Lee, and Liang (2020) recruits human participants from Amazon Mechanical Turk (AMT) to detect machine-generated words in a sentence. Uchendu et al. (2021) also recruits human participants from AMT and asks them to detect which one of

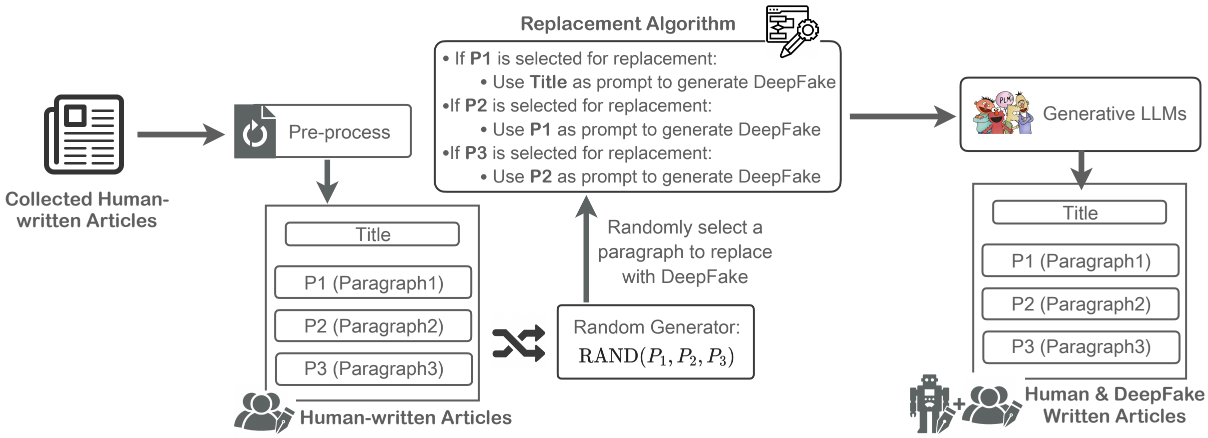

Figure 2: Illustration of the data generation process.

two articles is machine-generated and given one article, decide if it is machine-generated or not. Ippolito et al. (2020) evaluates the human ability to perform comparably given 2 different generation strategies. Brown et al. (2020) evaluates human performance in distinguishing human-written texts from GPT-3-generated texts. Finally, in all these works, the themes remain the same - humans perform poorly at detecting machine-generated texts, achieving about or below chance-level during evaluation.

Training Humans to Evaluate Deepfake Texts

Since human performance in deepfake text detection is very poor, a line of studies have attempted to train the humans first and then ask them to detect the deepfake texts. For example, (Gehrmann, Strobelt, and Rush 2019) proposed a color-coded tool named GLTR (Giant Language Model Test Room). GLTR color codes words based on the distribution level which improves human performance from 54% to 72% (Gehrmann, Strobelt, and Rush 2019). Dugan et al. (2020) gamifies machine-generated text detection by training humans to detect the boundary at which a document becomes deepfake to earn points. Humans are given the option to select one of many reasons or include their own reasons for which a sentence could be machine-generated (Dou et al. 2022). Our framework is modeled more closely after Dugan et al. (2020)'s work. Next, Clark et al. (2021) proposes 3 training techniques - *Instruction-based, Example-based,* and *Comparison-based*. *Example-based* training improved the accuracy from 50% to 55% (Clark et al. 2021).

Despite persistent efforts in human training, all methods except for GLTR did not yield significant improvements in human performance. However, GLTR achieved an average of 56% F1 score on 19 pairs of human vs. state-of-the-art (SOTA) NTGs (Uchendu et al. 2021), suggesting that older deepfake text detectors are inferior/obsolete to modern models. This further necessitates more thorough investigation into advanced human train methods, instead of relying on detectors. We hypothesize that previous training techniques failed because they did not consider that collaboration and skill levels could affect performance. Hence, while we implement the *example-based* training technique, we also take into account expertise and collaboration elements.

Automatic Evaluation of Deepfake Texts

As LLMs such as GPT-2, ChatGPT, LLaMA, etc. are able to be used maliciously to generate misinformation at scale, several techniques have been employed to detect deepfake texts. Using *stylometric*[2] classifiers, researchers adopted stylometry from traditional authorship attribution solutions to achieve automatic deepfake text detection (Uchendu et al. 2020; Fröhling and Zubiaga 2021). However, due to the flaws of *stylometric* classifiers, *deep-learning* techniques have been proposed (Bakhtin et al. 2019; Huggingface 2023; Zellers et al. 2019; Ippolito et al. 2020; Ai et al. 2022; Jawahar, Abdul-Mageed, and Lakshmanan 2022). While these *deep-learning* techniques achieved high performance and significantly improved from *stylometric* classifiers, they are not interpretable. To mitigate this issue, *statistical-based* classifiers are proposed (Gehrmann, Strobelt, and Rush 2019; Pillutla et al. 2021; Gallé et al. 2021; Pillutla et al. 2022; Mitchell et al. 2023). Lastly, to combine the benefits of each of the 3 types of classifiers for deepfake text detection, 2 or more of these classifier types are combined to build a more robust classifier. Uchendu, Le, and Lee (2023) defines these classifiers as *hybrid* classifiers and they achieve superior performance (Liu et al. 2022; Kushnareva et al. 2021; Zhong et al. 2020). Lastly, using automatic deepfake text detectors, deepfake detection has been achieved with reasonable performance. However, in the real world, as humans cannot solely depend on these models to detect deepfakes, they need to be equipped at performing the task themselves. A common theme in most of the detectors are that newer LLMs are harder to detect, which can sometimes make the older detectors obsolete. Thus, it is imperative that humans are also able to perform the task of deepfake text detection. For this reason, a few researchers have evaluated human performance in this task under several settings. See below.

[2]stylometry is the statistical analysis of an author's writing style/signature.

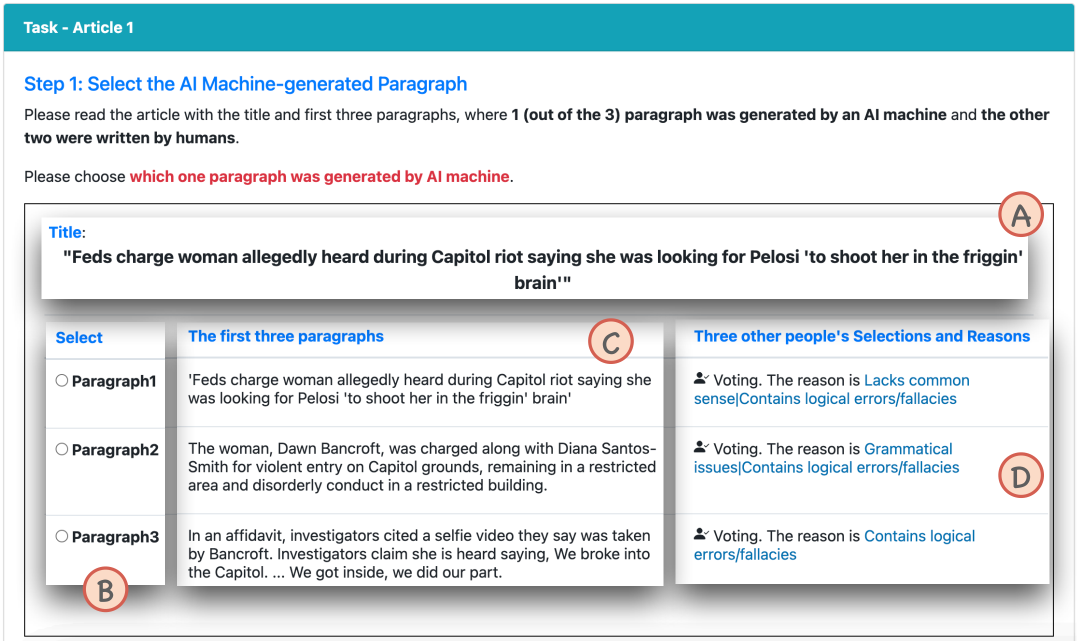

Figure 3: User interface for the AMT collaborative group workers to choose the LLM-generated one paragraph, whereas the individual group workers can only see A, B, and C panels.

Methodology

The collective body of prior research has consistently highlighted the inherent difficulty involved in solving the deepfake detection problem (Uchendu et al. 2020; Clark et al. 2021; Ippolito et al. 2020; Dugan et al. 2020; Gehrmann, Strobelt, and Rush 2019). Building upon the concept of "collective intelligence" that has exhibited superior performance in online misinformation detection tasks (Horowitz et al. 2022; Mercier and Sperber 2011; Liu 2018; Seo, Xiong, and Lee 2019) this study aims to investigate whether human collaboration can enhance the detection of deepfake texts. Specifically, the research methodology involves the creation of articles comprising two paragraphs authored by humans and one paragraph generated by an LLM (*i.e.,* GPT-2). Non-expert participants from Amazon Mechanical Turk (AMT) are then engaged in an asynchronous collaboration setting to discern the LLM-generated paragraph from the human-written paragraphs within the mixed-up articles. Additionally, English experts sourced from Upwork are enlisted to perform the same task but in a synchronous collaboration manner. To gain deeper insights into the reasoning process of humans, we analyzed the explanations provided by participants in the deepfake detection tasks. This study design is rooted in the practical reality that, with the increasingly impressive capabilities of LLMs, humans are increasingly inclined to employ LLMs to amend or replace portions of their own written content. The subsequent sections provide a detailed account of the data generation procedure, the design of the human study, and the analysis of explanations.

Data Generation

As an overview of the data generation process shown in Figures 2, to build this dataset, we collected 200 human-written news articles (mostly politics since this work is motivated by mitigating the risk of mis/disinformation or fake news dissemination) from reputable news sources such as CNN and Washington Post. Next, of the 200 articles, we took the first suitable 50 articles with at least 3 paragraphs. Then, we removed all paragraphs after the 3rd paragraph. Since the goal is to have a multi-authored article (human and LLM), we **randomly select one out of the three paragraphs** to be replaced by LLM-generated texts. We use a random number generator to select which paragraphs are to be replaced. As a result, we replaced the *Paragraph 1* in 23 articles, *Paragraph 2* in 16 articles, and *Paragraph 3* in 11 articles.

For Deepfake text generation, we used GPT-2 (Radford et al. 2019) XL which has 1.5 billion parameters, and the *aitextgen*[3], a robust implementation of GPT-2 to generate texts with the default parameters[4]. We then followed the following mechanism to replace the article with the LLM-generated paragraph:

- If paragraph 1 is selected to be replaced: Use Title as a prompt to generate GPT-2 replacement;

[3]https://github.com/minimaxir/aitextgen

[4]We used only GPT-2 instead of GPT-3 or above to generate the deepfake texts because: (1) GPT-2 and GPT-3 or above are using the similar algorithms. Based on (Uchendu et al. 2020; Clark et al. 2021), human performance on detecting GPT-2 and GPT-3 texts have similar accuracies; and (2) GPT-2 is cheaper to generate texts with than GPT-3 or above since GPT-2 is open-source and GPT-3 or above is not.

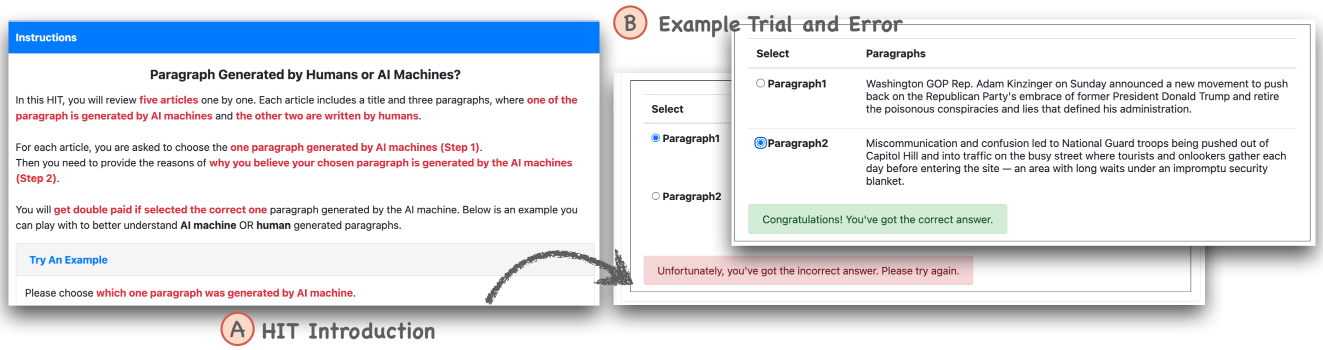

Figure 4: Instructions to train users by providing prompt feedback.

- If paragraph 2 is selected to be replaced: Use Paragraph 1 as a prompt to generate GPT-2 replacement;
- If paragraph 3 is selected to be replaced: Use Paragraph 2 as a prompt to generate GPT-2 replacement.

Since we are unable to control the number of paragraphs GPT-2 generates given a prompt, we use a Masked Language Model (MLM) to choose the best GPT-2 replacement that fits well with the article. We use a BERT-base MLM (Devlin et al. 2018) to get the probability and calculate the perplexity score of the next sentence. Let us call this model G(.), it takes 2 inputs - the first and probable second sentence/paragraph ($G(Text\_1, Text\_2)$) and outputs a score. The lower the score, the more probable $Text\_2$ is the next sentence. For instance, say GPT-2 texts is to replace Paragraph 2 (P2) of an article:

1. We use P1 as prompt to generate P2 with GPT-2;
2. GPT-2 generates another 3-paragraph article with P1 as the prompt;
3. To find the suitable P2 replacement, we do G(P1, each GPT-2 generated paragraph);
4. Since low scores with G(.) is considered most probable, the P2 replacement is the GPT-2 paragraph that yielded the lowest score with G(.).

After we created these multi-authored articles, we manually did a quality check of a few of these articles by checking for consistency and coherence. See Figure 3(C) for an example of the final multi-authored article. We also observe that based on the replacement algorithm, some bias in detection may be introduced. Replacing paragraph 3 may be seen as easier because there is no other paragraph after it to judge the coherency. However, we keep the generation process fair by only using the text right before the paragraph as a prompt to generate the next paragraph. Thus, to replace paragraph 3, we only use paragraph 2 as a prompt, not the previous paragraphs and title.

Human Study Design

Next, as we have defined this realistic scenario, we hypothesize that collaboration will improve human detection of deepfake texts. Thus, we define 2 variables for this experiment - Individual vs. Collaboration and English expert vs. English non-expert. We investigate how collaboration (both synchronous and asynchronous) improves from individual-based detection of deepfake texts. The hypothesis here is that when humans come together to solve a task, collaborative effort will be a significant improvement from average individual efforts. Additionally, as human detection of deepfake texts is non-trivial, we want to investigate if the task is non-trivial because English non-experts focus on misleading cues as opposed to English experts.

Study1: Collaboration between AMT Participants

Participant Recruitment. Inspired by Clark et al. (2021), Dugan et al. (2020), and Van Der Lee et al. (2019), we used Amazon Mechanical Turk (AMT) to collect responses from non-expert evaluators. We deployed a two-stage process to conduct non-expert human studies. First, we posted a *qualification-required* Human Intelligence Task (HIT) that pays \$0.50 per assignment on AMT to recruit 240 qualified workers. In addition to our custom qualification used for worker grouping, three built-in worker qualifications are used in all HITS, including *i)* HIT Approval Rate ($\leq$98%), Number of Approved HITs ($\geq$3000), and Locale (US Only) Qualification. Next, we only enable the qualified workers to enter the large-scale labeling tasks. The approximate time to finish each labeling task is around 5 minutes (*i.e.,* the average time of two authors on finishing a random HIT). Therefore, we aim for \$7.25 per hour and set the final payment as \$0.6 for each assignment. Further, we provide "double-payment" to workers who made correct submissions as the extra bonus.

Experiment Design. During the large-scale labeling task, we divide the recruited qualified workers into two groups to represent the individual vs. collaborative settings, respectively. We define group1 as *Individual Group*, in which each worker was asked to select the LLM-generated paragraph without any references. See Figure 3, for example, humans in *Individual Group* can only see the introduction with panels (A) (B) and (C). On the other hand, we design group 2 to be *Collaborative Group*, where the workers were asked to conduct the same task after the *Individual Group* finishes all HITs (*i.e.,* see panel (A), (B), (C) in Figure 3). In addition, workers from the *Collaborative Group* could also see

Participant	Gender	Education	Group
P1	Female	Bachelor's degree	G1
P2	Female	Bachelor's degree	
P3	Female	Bachelor's degree	
P4	Female	Bachelor's degree	G2
P5	Male	Bachelor's degree	
P6	Male	Graduate degree	
P7	Female	Graduate degree	G3
P8	Female	Graduate degree	
P9	Female	Bachelor's degree	
P10	Female	Bachelor's degree	G4
P11	Female	Bachelor's degree	
P12	Male	Bachelor's degree	
P13	Female	Graduate degree	G5
P14	Female	Bachelor's degree	
P15	Male	Graduate degree	
P16	Female	Bachelor's degree	G6
P17	Male	Bachelor's degree	
P18	Male	Bachelor's degree	

Table 1: Expert (Upwork) participant demographics.

the selection results from group 1 in an asynchronous manner, as the example shown in Figure 3(D), to support their own selection.

Furthermore, we take actions to incentivize workers to provide qualified results: *i)* in our instruction, we provide immediate feedback on the worker's selection to calibrate their accuracy. In specific, after reading the HIT instruction (*i.e.,* Figure 4 (A)), workers can get a deeper understanding of "which paragraph is generated by AI machine" by trial and error on selecting one example (*i.e.,* Figure 4 (B)). Participants were given unlimited chances to change their answers. This example-based training process was inspired by Clark et al. (2021)'s human evaluation study and was found to be the most effective training technique. *ii)* We pay double compensation to the workers who provide correct answers. This aims to encourage workers to get high accuracy in selecting the correct machine-generated paragraphs. *iii)* We set the minimum time constraint (*i.e.,* one minute) for workers to submit their HITs so that the workers will concentrate on the task for at least one minute instead of randomly selecting one answer and submitting the HIT. Note that we also disabled the copy and paste functions in the user interface to prevent workers from searching for the paragraphs from online resources.

Study2: Collaboration Between Upwork Participants

Participant Recruitment. We utilized Upwork[5] to recruit expert evaluators, especially those with expertise in writing domains.Through Upwork, we first posted a task description as a client to gather participants. We mentioned in the description that this is for research and provided all necessary information such as research objectives and example questions. Our recruitment advertisement also highlighted the mandatory requirements: (1) a participant should be at least 18 years old; and (2) a participant should be a native English speaker. Lastly, if they were willing to proceed, they were asked to submit a proposal answering the following questions: (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.

[5]Upwork is one of the leading freelance websites with a substantial network size. Upwork facilitates the freelance industry by introducing skilled freelancers in diverse categories like writing, design, and web development. With its automated recommendation system, we can effectively match our expert workers to our needs. See link: https://sellcoursesonline.com/Upwork-statistics.

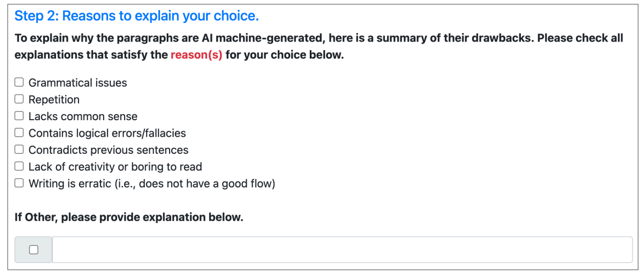

Figure 5: User interface for participants to select explanations for identified deepfake paragraphs.

One useful feature for accelerating the recruitment process in Upwork is that not only workers can apply to the postings but also clients like us can invite prospective candidates that seem suitable for the task to submit proposals. We manually reviewed workers' profile descriptions who specified their skill sets as copywriting, editing/proofreading, and content writing and then sent them invites.

While making recruitment decisions, we verified participants' eligibility by checking their self-reported age, language, and education in the profile, in addition to evaluating their proposal responses. It resulted in a total of 18 finalists to officially begin the study. Next, we sent them the consent form via the platform's messaging function and activated Upwork contracts only after they returned the signed form. A primary purpose of the contracts was for clients to compensate workers based on submitted hours through the Upwork system. Participants' requested hourly wages ranged from $25-$35 per hour depending on their prior experiences and education levels. All 18 individuals successfully signed both documents and were compensated accordingly. Table 1 gives the self-reported demographic of recruited Upworkers.

Experiment Design. To compare experts' deepfake text detection accuracy with respect to individual vs. collaborative settings, our Upwork study consists of two subexperiments. The first experiment asks Upwork participants to perform a given task on their own. The second experiment requires three individuals to solve the questions as one group in a synchronous manner. We used Qualtrics[6] service to generate and disseminate the study form. Upwork participants

[6]https://www.qualtrics.com

Setting	Mean Accuracy	P-Value
Baseline vs. Individual	33.33% vs. 44.99%***	3.8e-05
Baseline vs. Collaboration	33.33% vs. **51.35%*****	2.8e-05
Individual vs. Collaboration	44.99% vs. **51.35%**	0.054

Table 2: Paired t-test results for AMT experiments. (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$)

Setting	Mean Accuracy	P-Value
Baseline vs. Individual	33.33% vs. 56.11%***	8.2e-11
Baseline vs. Collaboration	33.33% vs. **68.87%*****	1.2e-12
Individual vs. Collaboration	56.11% vs. **68.87%*****	1.3e-05

Table 3: Paired t-test results for Upwork experiments. (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$)

were given one week to complete the survey. Upon completion, we randomly grouped 3 participants per team, resulting in 6 teams in total for synchronous collaboration (Table 1). All discussions were conducted on the video communications software - Zoom and we leveraged Zoom's built-in audio transcription feature, which is powered by Otter.ai[7] for discourse analyses. In addition to the written consent obtained during the recruitment procedure, verbal consent for participation in the discussion and for audio recording was obtained prior to the start of each session. One member of the study team served as a moderator for the meetings. Depending on the participant's schedule and level of commitment in their group, each meeting lasted 1.5 - 3 hours.

In-depth Analysis on Detection Explanations

We build the explanation section similar to RoFT (Dugan et al. 2020), a gamification technique for improving human performance in deepfake text detection. In the RoFT framework, participants were asked to select from a pre-defined list one or more reasons such as repetition, grammar errors, etc. Participants were also given another option, where they can enter their own justification if they do not find any suitable selection from the provided list.

To determine the list of pre-defined reasoning explanations in deepfake text detection, we first refer to Dou et al. (2022), which provides a detailed list of 10 errors in which annotators have been indicated to be good indicators of deepfake texts. However, these errors are general errors and thus some are not applicable to the task of detecting deepfake paragraphs. Therefore, due to this novel application, we only select the most relevant errors. Additionally, we also include relevant errors from Dugan et al. (2020) including the selection of other, a gamification of deepfake text detection. As the result shown in Figure 5, we consequently provide seven pre-defined rationales that correspond to flaws typically observed in deepfake texts (Dou et al. 2022; Dugan et al. 2020), including "Grammatical issues", "Repetition", "Lacks common sense", "Contains logical errors", "Contradicts previous sentences", "Lack of creativity or boring to read", "Writing is erratic" (i.e. does not have a good flow), and an additional open-ended selection - other for participants to write more of their own.

Given the pre-defined reasoning explanation list, we ask both individuals and collaborative groups to provide their explanations for each corresponding detection instance. We apply this implementation for both non-expert and expert groups, resulting in the in-depth explanation analysis with respect to four scenarios (*i.e.,* individual-expert, collaboration-expert, individual-non-expert, collaboration-non-expert). To provide more fine-grained insights, we further separate the deepfake detection results into correct detection and incorrect detection subgroups.

[7]https://otter.ai

Experimental Results

Evaluation Metrics and Baselines

Objective Metrics. We measure how well participants perform the tasks and compared them across different experiment settings. To quantify the detection performance of each setting, we computed the proportion of people who got the answer correct given a set of 50 questions Q={q_1, q_2,..., q_{50}}. Suppose l_n is the number of participants with correct answers, and m_n is the total number of participants for the question q_n, we calculated the accuracy using this formula: $acc_n = l_n/m_n \times 100$. This resulted in a list of accuracy scores ACC={acc_1, acc_2, ..., acc_{50}}, representing the participants' performance of 50 articles. To further evaluate whether the means of two groups (individual vs. collaborative & non-experts vs. experts settings) are statistically different, we conducted a paired independent sample T-test. Since the T-test is grounded on the assumption of normality (Gerald 2018), we ran the Kolmogorov-Smirnov test on our data and confirmed that the requirement was satisfied. Following, we summarize the results of statistical testing.

Baseline. Each of the 50 3-paragraphed articles has 2 paragraphs authored by human and 1 paragraph deepfake-authored. Therefore, participants have a 1/3 chance of selecting the deepfake paragraph, and we developed a random generator to randomly identify one of the paragraph as deepfake. As such, the baseline performance of the random-guessing accuracy is approximately to be 33.33%.

Study 1: Collaboration between AMT Workers

Detection Performance. From Table 2 we observe that English non-experts achieve an average accuracy of 44.99% individually, which is a 11.66% increase from the baseline (random-guessing) of 33.33%. Using a paired T-test to measure statistical significance, the baseline vs. individual performance comparison achieve a p-value of $3.8e{-}05$ which indicates strong statistical significance. Next, for the collaborative setting, the non-experts collaborate asynchronously, achieving an average accuracy of 51.35%. The p-value of Individual vs. Collaboration comparison is 0.054, indicating weak statistical significance. However, the comparison of Baseline vs. Collaboration yields a p-value of $2.8e{-}05$ which indicates strong significance. Thus, all comparison

Explanation Type	Correct Detection				Incorrect Detection			
	Non-Expert (I vs. C)		Expert (I vs. C)		Non-Expert (I vs. C)		Expert (I vs. C)	
	Percentage (%)	P-Value	Percentage (%)	P-Value	Percentage (%)	P-Value	Percentage (%)	P-Value
Grammar	13.97 vs. 23.08**	0.004	15.33 vs. 24.6***	0.001	15.65 vs. 16.89	0.587	14.22 v.s 12.07	0.329
Repetition	6.73 vs. 6.69	0.98	4 vs. 6.4	0.125	8.53* vs. 5.62	0.044	1.67 vs. 2	0.703
Common Sense	9.25 vs. 15.48*	0.011	13 vs. 28***	4.8e-07	13.02 vs. 9.94	0.166	3.33 vs. 5.56	0.163
Logical Errors	11.64 vs. 10.24	0.496	7.78 vs. 14.4**	0.002	18.54*** vs. 7.7	3.3e-05	3.89 vs. 4	0.938
Self-Contradiction	9.35 vs. 5.57	0.077	7.67 vs. 14.8***	0.001	18.01*** vs. 6.7	9.6e-07	6.56 vs. 3.6	0.054
Lack of Creativity	12.87 vs. 13.49	0.776	8.33 vs. 7.6	0.714	16.9 vs. 14.13	0.322	8.11** v.s 3.6	0.002
Coherence	14.64 vs. 19.29*	0.045	20.56 vs. 32***	0.0008	11.65 vs. 10.06	0.318	13.78** vs. 9.2	0.003
Other	0 vs. 0	N/A	12.22 vs 18.4*	0.014	0 vs. 0	N/A	6.78 vs. 8.4	0.264

Table 4: The percentage of frequency for each reasoning explanation category w.r.t. correct & incorrect detection (I vs. C = Individual vs. Collaboration) and corresponding t-test results (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

groups for non-experts indicate strong significant improvement, except for Individual vs. Collaboration in which the improvement observed during collaboration is weak.

Analysis of Reasoning Explanations. In Table 4, we measure the statistical significance of explanations used by participants individually and collaboratively for each of the seven reasoning explanations, where we divide based on both correct and incorrect detection responses. For AMT (*i.e.,* non-experts), we observe only a few statistically significant explanations. Correct responses show significant scores for grammar, common sense, and coherence. While incorrect responses have significant scores for repetition, logical errors, and self-contradiction. Furthermore, we visualize these explanations for both correct and incorrect responses in Figures 6 and 7, respectively. In these figures, we observe that non-experts, both Individually and collaboratively, do not show any patterns in response. Thus, in summary, these factors yielded a minimal improvement in performance when non-experts collaborated. Another reason for the minimal improvement is the style of collaboration utilized by non-experts - asynchronous collaboration. We further elaborate on the potential hypothesis in the Discussion section below.

Study 2: Collaboration Between Upwork Participants

Detection Performance. The English experts achieve an average accuracy of 56.11% and a p-value of 8.2e-11 for Baseline vs. Individual, indicating strong significance. In the collaborative (synchronous) setting, the participants achieve an average accuracy of 68.87% with a p-value of 1.3e-05 for Individual vs. Collaboration, suggesting a strong statistical significance. Also the p-value for the comparison of Baseline vs. Collaboration (1.2e-12) indicates an even stronger significance.

Analysis of Reasoning Explanations. In Table 4, we measure the statistical significance of explanations used individually and collaboratively. We measure two categories when responses are correct and incorrect. For Upwork (experts), we observe more statistically significant explanations for correct responses than for incorrect responses. Correct responses had 6 statistically significant types from collaborations out of 8 - grammar, common sense, logical errors, self-contradiction, coherence, and other. Next, incorrect responses recorded only 2 statistically significant responses - lack of creativity and coherence. Furthermore, we observe in Table 4 that experts show a much stronger significance (p-value < 0.01) in the frequency of explanations used than non-experts in the correct detection cases.

Furthermore, Figures 6 and 7 visualize the frequency of explanations used by participants for correct and incorrect responses, respectively. We observe that experts used coherence, common sense, grammar errors, other,[8] and self-contradiction more frequently collaboratively for correct responses. However, individually, they used grammar errors, coherence and other more frequently for incorrect responses. This suggests that coherence, common sense, and self-contradiction are strong indicators for distinguishing deepfake texts from human-written texts, since they are the only explanations do not overlap in frequency between correct and incorrect responses.

Discussion

Deepfake Text Detection is Non-Trivial for Humans

In order to further confirm the difficulty of the task of detecting 1/3 of paragraphs as deepfake, we asked ChatGPT to perform the task. Recently, ChatGPT[9] has been found to have emergent abilities (OpenAI 2023), one of which is being able to accurately perform many text classification tasks accurately. Given the recent observation that using personas with ChatGPT improves accuracy, we use a persona-themed prompt - *You are an expert. Which of the 3 paragraphs is AI-generated? Answer choices: paragraph_1, paragraph_2, or paragraph_3*. Using this prompt, ChatGPT achieves a 38% accuracy, only 5% above the baseline. In fact, ChatGPT got confused with certain paragraphs that it deviated from the

[8]off-topic and off-prompt were the most frequent justifications.
[9]https://openai.com/blog/chatgpt

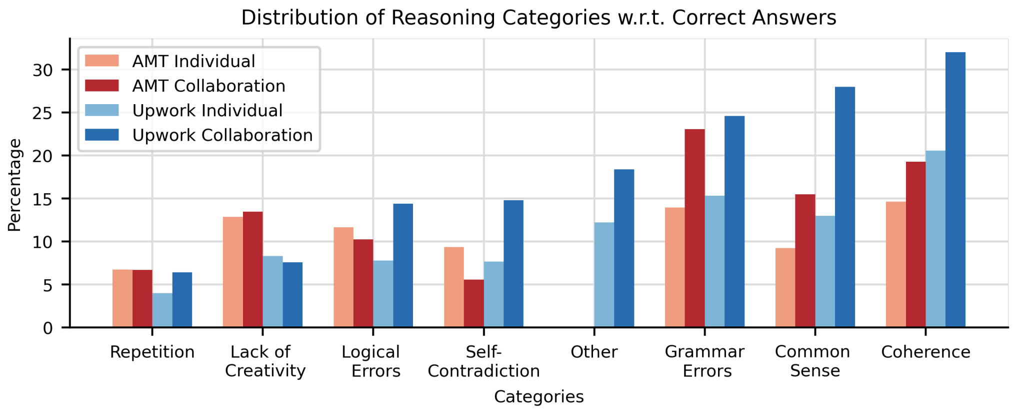

Figure 6: The percentage of frequency for selected reasoning explanation w.r.t. correct human detection.

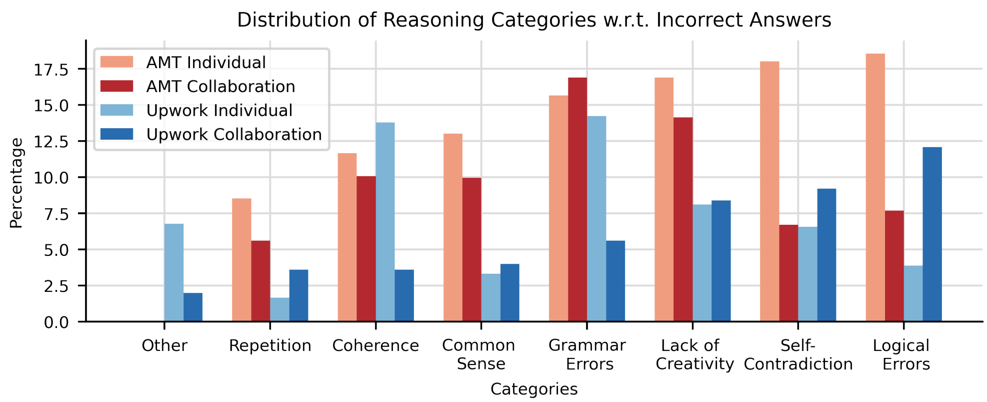

Figure 7: The percentage of frequency for selected reasoning explanations w.r.t. incorrect human detection.

answer choices and generated other responses, such as none of the paragraphs are AI-generated, all are AI-generated, or picking two instead of one paragraph, e.g., paragraph_1 & paragraph_3. While using our framework, humans achieve an average accuracy of 44.99% (non-experts) and 51.35% (experts), individually. Through collaboration, their performances increase to an average accuracy of 51.35% and 68.87% for non-experts.

Detection Performance Comparison Between Experts and Non-Experts

In the aforementioned sections, we observe that collaboration is an effective approach to improve human performance in deepfake text detection. Further, the results in Tables 2 and 3 also suggests that **experts achieve a more significant improvement with collaboration than non-experts**. There are two potential reasons for this: (1) expert participants are able to utilize their expert knowledge more efficiently when collaborating. This is further confirmed in Figures 6, where Upwork (experts) Collaboration show more frequent use of coherence, common sense, grammar errors, other, self-contradiction, and logical errors than Individually. The intuition here is individually, expert participants did not use these explanations as frequently which yielded an average accuracy of 56.11%, however, during collaboration, they used these explanations more frequently and accurately, improving the average accuracy to 68.87%. (2) The second reason is argued by the body of CSCW literature (e.g., (Birnholtz and Ibara 2012; Shirani, Tafti, and Affisco 1999; Mabrito 2006)) which suggests that the gains of synchronous collaboration outweighs the benefits of asynchronous collaboration.

However, for non-experts due to the erratic usage of explanations as observed in Figures 6 and 7, it is difficult to ascertain a pattern. This is potentially the reason why, although collaboration is statistically significant for non-experts, it is a weak significance (p-value=0.054). Furthermore, this suggests that while experts are able to collaborate well, non-experts may require a guided synchronous collaboration strategy to further improve performance. It is worth noting that when comparing the Non-expert with AMT and Expert with Upwork performance, the difference may potentially also be resulted from different collaboration modes (*i.e.,* "asynchronous" vs. "synchronous" settings) and different compensation levels. However, with the respective rational settings with two groups, the Experts can outperform non-experts in detecting deepfake texts.

Which Explanation Categories Can Potentially Be Helpful for Deepfake Text Detection?

Experts' mentions of **coherence, logical errors, and self-contradiction errors** as explanations for deepfake text detection were significantly higher in the collaborative setting than in the individual setting, specifically for correct responses (Figure 6). Non-experts showed no pattern differences in coherence, logical errors, and self-contradiction explanations between individuals and collaboration. However, expert participants used them, especially coherence and self-contradiction, more in collaboration when they detected the deepfake texts successfully and less in collaboration when they detected deepfake texts inaccurately. This result corroborates Dou et al. (2022)'s finding that machines are prone to fall short of those categories. Taking into account experts' superior performance in deepfake text detection, we conclude that both coherence errors and self-contradiction errors are strong indicators of deepfake text. Table 4 confirms this finding as well since both explanations have a p-value < 0.001 for correct responses, suggesting very strong significance. Regarding logical errors, expert participants used them more frequently in the collaborative setting for both correct and incorrect responses. That said, our findings imply that logical flaws may be a weak predictor of deepfake.

Which Explanation Categories Should Be Cautious Indicators for Deepfake Text Detection?

In line with previous works (Dou et al. 2022; Clark et al. 2021), we observe that participants' mentions of **grammar errors, repetition, creativity** were associated with incorrect detection of deepfake texts. These markers are identified as deceptive indicators of deepfake texts.

1. **Grammar Errors**: Experts use grammar errors frequently for both correct (collaboration) and incorrect (individual) responses. This could be attributed to the fact that humans are equally prone to making grammatical mistakes. As a result, employing this explanation can result in both accurate and inaccurate detection. Still, our results indicate that experts can use grammar errors for detection signals more correctly as opposed to non-experts. Non-experts use grammar errors frequently for both incorrect and correct responses, although a bit more frequently for incorrect responses. Furthermore, this phenomenon is confirmed by the findings in (Clark et al. 2021; Dou et al. 2022) that grammar errors are weak indicators of deepfake texts. Therefore, we conclude that grammar errors are weak indicators of deepfake texts.
2. **Repetition:** We observe in Figure 6 and 7 that repetition is the last and second-last frequently used explanations for correct and incorrect responses, respectively. This was a good indicator of deepfake texts when NTGs were still in their infancy. However, NTGs have improved significantly such that the quality of generations can be misconstrued as human-written. In addition, we took measures to ensure high-quality generations, which is discussed in detail in the Method Section. Thus, repetition was not prevalent in the deepfake texts, making repetition a weak indicator of deepfake texts.
3. **Creativity:** News is supposed to be the unbiased reporting of factual events. Therefore, as these events remain non-fiction, news articles are not creative and should not be judged by their level of creativity. This is the reason why experts used creativity very sparingly because English experts they are aware of which style of writing should creative or not-creative. Therefore, experts use repetition as explanation second-to-last for correct responses (Figure 6). Unsurprisingly, experts used creativity a bit more frequently for incorrect responses, with the more frequent usage observed in individuals (Figure 7). However, for non-experts, creativity is also used sparingly for correct responses but frequently for incorrect responses. Therefore, due to the frequent usage of creativity for incorrect responses vs. infrequent usage for correct responses, it follows that for the task of detecting deepfake news paragraphs, creativity is a false indicator of deepfake texts.

Limitation

To implement design choices and run manageable experiments, we made a few simplifications that may limit our findings. First, since we only use GPT-2 to generate deepfake texts, our findings may not be directly applicable to other NTGs. However, our choice of using GPT-2 is reasonable because: (1) prior research reported that human detection performance of deepfake texts by the later GPT-3 and GPT-2 is similar (Uchendu et al. 2021; Clark et al. 2021), and (2) using the largest parameter size of GPT-2 enabled us to generate deepfake texts more effectively that closely resembles GPT-3 quality. Furthermore, as we use the default hyperparameters of GPT-2 to generate the texts, the results may be limited to that sampling technique. However, we mitigated this issue by manually checking the quality of a few of the articles and found the deepfake texts to be coherent and consistent with the rest of the paragraphs. This preserved the integrity of the experiments as the task remained non-trivial.

Conclusion

Our study investigated the impact of human collaboration on improving the detection of deepfake texts. To create a realistic experimental setup, we constructed a three-paragraph article comprising one LLM-generated (deepfake) paragraph and two human-written paragraphs. Participants were tasked with identifying the deepfake paragraph and providing explanations based on seven explanation types. For participant recruitment, we recruited non-expert participants from AMT for asynchronous collaboration and experts from Upwork for synchronous collaboration. The results revealed that collaboration is likely to enhance the detection performance of both non-experts and experts. We further identified several strong and weak indicators of deepfake texts through the explanation analysis. Notably, the improved performance of participants compared to the baselines indicated that our *Turing Test* framework effectively facilitated the enhancement of human deepfake text detection performance.

Ethical Statement

Our research protocol was approved by the Institutional Review Board (IRB) at our institution. We only recruited human participants 18 years old or over. Participants did not have to complete the entire task to be paid. Using AMT, participants' identification was already anonymized, but for Upwork we anonymized participants by assigning them numerical values for the analysis. For performing the deepfake text detection task, all our human participants, from both AMT and Upwork, were paid over minimum wage rate. Next, the articles that we used for the experiments are the first 3-paragraphs of news articles. While we did not share the answer to the task, we clearly informed participants that the presented texts (and one of three paragraphs therein) contains deepfake texts. Therefore, we believe that participants are unlikely to be negatively influenced by their exposure to the test news articles with deepfake paragraphs.

Acknowledgments

This work was in part supported by NSF awards #1820609, #2114824, and #2131144, and PSU CSRE seed grant 2023. We thank the crowd workers from AMT and the experts from Upwork for participating in this study. We also thank the anonymous reviewers for their constructive feedback.

References

Ai, B.; Wang, Y.; Tan, Y.; and Tan, S. 2022. Whodunit? Learning to Contrast for Authorship Attribution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 1142–1157.

Bagdasaryan, E.; and Shmatikov, V. 2022. Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*, 769–786. IEEE.

Bakhtin, A.; Gross, S.; Ott, M.; Deng, Y.; Ranzato, M.; and Szlam, A. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Birnholtz, J.; and Ibara, S. 2012. Tracking changes in collaborative writing: edits, visibility and group maintenance. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 809–818.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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 *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.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donahue, C.; Lee, M.; and Liang, P. 2020. Enabling Language Models to Fill in the Blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2492–2501.

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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7250–7274.

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

Fagni, T.; Falchi, F.; Gambini, M.; Martella, A.; and Tesconi, M. 2021. TweepFake: About detecting deepfake tweets. *Plos one*, 16(5): e0251415.

Fröhling, L.; and Zubiaga, A. 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7: e443.

Gallé, M.; Rozen, J.; Kruszewski, G.; and Elsahar, H. 2021. Unsupervised and Distributional Detection of Machine-Generated Text. *arXiv preprint arXiv:2111.02878*.

Gehrmann, S.; Strobelt, H.; and Rush, A. M. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116.

Gerald, B. 2018. A brief review of independent, dependent and one sample t-test. *International Journal of Applied Mathematics and Theoretical Physics*, 4(2): 50–54.

Horowitz, M.; Cushion, S.; Dragomir, M.; Gutiérrez Manjón, S.; and Pantti, M. 2022. A framework for assessing the role of public service media organizations in countering disinformation. *Digital journalism*, 10(5): 843–865.

Huggingface. 2023. GPT-2 Output Detector Demo. https://huggingface.co/openai-detector/. Accessed: 2023-09-18.

Ippolito, D.; Duckworth, D.; Callison-Burch, C.; and Eck, D. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1808–1822. Online: Association for Computational Linguistics.

Jawahar, G.; Abdul-Mageed, M.; and Lakshmanan, L. 2022. Automatic Detection of Entity-Manipulated Text using Factual Knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 86–93.

Juneja, P.; and Mitra, T. 2022. Human and Technological Infrastructures of Fact-Checking. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Kushnareva, L.; Cherniavskii, D.; Mikhailov, V.; Artemova, E.; Barannikov, S.; Bernstein, A.; Piontkovskaya, I.; Piontkovski, D.; and Burnaev, E. 2021. Artificial Text Detection via Examining the Topology of Attention Maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 635–649.

Lee, J.; Le, T.; Chen, J.; and Lee, D. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, 3637–3647.

Liu, I. J. 2018. CekFakta: Collaborative Fact-Checking in Indonesias. https://blog.google/outreach-initiatives/google-news-initiative/cekfakta-collaborative-fact-checking-indonesia/. Accessed: 2023-09-18.

Liu, X.; Zhang, Z.; Wang, Y.; Lan, Y.; and Shen, C. 2022. CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Data Limitation With Contrastive Learning. *arXiv preprint arXiv:2212.10341*.

Mabrito, M. 2006. A study of synchronous versus asynchronous collaboration in an online business writing class. *The American Journal of Distance Education*, 20(2): 93–107.

Mercier, H.; and Sperber, D. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2): 57–74.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Pillutla, K.; Liu, L.; Thickstun, J.; Welleck, S.; Swayamdipta, S.; Zellers, R.; Oh, S.; Choi, Y.; and Harchaoui, Z. 2022. MAUVE Scores for Generative Models: Theory and Practice. *arXiv preprint arXiv:2212.14578*.

Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; and Harchaoui, Z. 2021. An information divergence measure between neural text and human text. *arXiv preprint arXiv:2102.01454*.

Pu, J.; Sarwar, Z.; Abdullah, S. M.; Rehman, A.; Kim, Y.; Bhattacharya, P.; Javed, M.; Viswanath, B.; Tech, V.; and Pakistan, L. 2023. Deepfake Text Detection: Limitations and Opportunities. *44th IEEE Symposium on Security and Privacy*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 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. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Seo, H.; Xiong, A.; and Lee, D. 2019. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science*, 265–274.

Shen, H.; and Wu, T. 2023. Parachute: Evaluating interactive human-lm co-writing systems. *arXiv preprint arXiv:2303.06333*.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Shirani, A. I.; Tafti, M. H.; and Affisco, J. F. 1999. Task and technology fit: a comparison of two technologies for synchronous and asynchronous group communication. *Information & management*, 36(3): 139–150.

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Tan, R.; Plummer, B.; and Saenko, K. 2020. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2081–2106.

Uchendu, A.; Le, T.; and Lee, D. 2023. Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *SIGKDD Explorations*, vol. 25.

Uchendu, A.; Le, T.; Shu, K.; and Lee, D. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 8384–8395.

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 *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2001–2016.

Van Der Lee, C.; Gatt, A.; Van Miltenburg, E.; Wubben, S.; and Krahmer, E. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wu, S.; Shen, H.; Weld, D. S.; Heer, J.; and Ribeiro, M. T. 2023. ScatterShot: Interactive In-context Example Curation for Text Transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 353–367.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Zhang, T. 2022. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5): 6259–6276.

Zheng, Y.; Li, G.; Li, Y.; Shan, C.; and Cheng, R. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5): 541–552.

Zhong, W.; Tang, D.; Xu, Z.; Wang, R.; Duan, N.; Zhou, M.; Wang, J.; and Yin, J. 2020. Neural Deepfake Detection with Factual Structure of Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2461–2470.