Understanding Individual and Team-based Human Factors in Detecting Deepfake Texts

ADAKU UCHENDU*, Pennsylvania State University, USA
JOOYOUNG LEE*, Pennsylvania State University, USA
HUA SHEN, Pennsylvania State University, USA
THAI LE, University of Mississippi, USA
TING-HAO 'KENNETH' HUANG, Pennsylvania State University, USA
DONGWON LEE, Pennsylvania State University, USA

In recent years, Natural Language Generation (NLG) techniques in AI (e.g., T5, GPT-3, ChatGPT) have shown a massive improvement and are now capable of generating human-like long coherent texts at scale, yielding so-called *deepfake texts*. This advancement, despite their benefits, can also cause security and privacy issues (e.g., plagiarism, identity obfuscation, disinformation attack). As such, it has become critically important to develop effective, practical, and scalable solutions to differentiate deepfake texts from human-written texts. Toward this challenge, in this work, we investigate how factors such as skill levels and collaborations impact how humans identify deepfake texts, studying three research questions: (1) do collaborative teams detect deepfake texts better than individuals? (2) do expert humans detect deepfake texts better than non-expert humans? (3) what are the factors that maximize the detection performance of humans? We implement these questions on two platforms: (1) non-expert humans or asynchronous teams on Amazon Mechanical Turk (AMT) and (2) expert humans or synchronous teams on the Upwork. By analyzing the detection performance and the factors that affected performance, some of our key findings are: (1) expert humans detect deepfake texts significantly better than non-expert humans, (2) synchronous teams on the Upwork detect deepfake texts significantly better than individuals, while asynchronous teams on the AMT detect deepfake texts weakly better than individuals, and (3) among various error categories, examining coherence and consistency in texts is useful in detecting deepfake texts. In conclusion, our work could inform the design of future tools/framework to improve collaborative human detection of deepfake texts.

Additional Key Words and Phrases: deepfake, individual, collaboration, expert, non-expert

ACM Reference Format:
Adaku Uchendu*, Jooyoung Lee*, Hua Shen, Thai Le, Ting-Hao 'Kenneth' Huang, and Dongwon Lee. 2018. Understanding Individual and Team-based Human Factors in Detecting Deepfake Texts. In . ACM, New York, NY, USA, 26 pages. https://doi.org/XXXXXXX.XXXXXXX

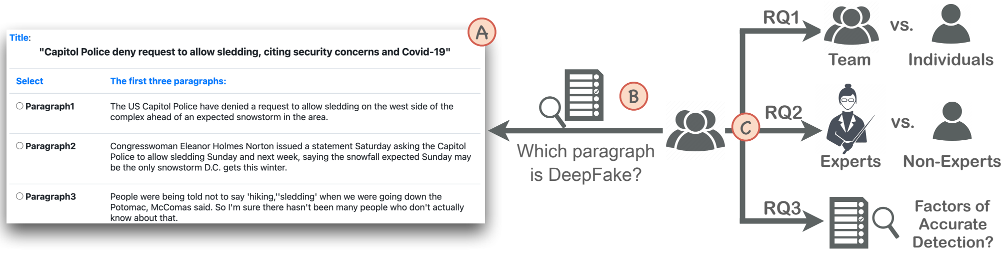
1 INTRODUCTION

In recent years, AI technologies have drastically advanced, enabling the generation of human-quality artifacts in various modalities, including texts, images, and videos [12, 31, 40]. Collectively, these AI-generated artifacts are known as **Deepfakes**. In particular, the advanced *Natural Language Generation* (NLG) techniques, especially large language models (e.g., GPT-3, T5, ChatGPT), are now able to generate long coherent texts without human intervention. In this work, we refer to

*Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX



Title:
"Capitol Police deny request to allow sledding, citing security concerns and Covid-19"

Select	The first three paragraphs:
○ Paragraph1	The US Capitol Police have denied a request to allow sledding on the west side of the complex ahead of an expected snowstorm in the area.
○ Paragraph2	Congresswoman Eleanor Holmes Norton issued a statement Saturday asking the Capitol Police to allow sledding Sunday and next week, saying the snowfall expected Sunday may be the only snowstorm D.C. gets this winter.
○ Paragraph3	People were being told not to say 'hiking,''sledding' when we were going down the Potomac, McComas said. So I'm sure there hasn't been many people who don't actually know about that.

Fig. 1. (A) Example of a multi-authored (Human & Deepfake) 3-paragraph article; (B) Task: Detecting DeepFake texts; (C) Description of three research questions.

such machine-generated texts as **Deepfake Texts**[1] and generative language models as *Neural Text Generator* (NTG). Such deepfake texts have many obvious benefits in diverse applications. For instance, individuals can use deepfake texts in writing draft blog profiles and program codes for learning while business can use them for writing bulk emails or product descriptions at scale. However, as it is true for any technology, deepfake texts also can be misused in many applications. For instance, students may abuse deepfake texts in writing essay homework, scammers may generate sophisticated phishing messages from deepfake texts, and state-backed operators may use deepfake texts as part of disinformation attack. Therefore, a great need to differentiate deepfake texts from human-written texts has naturally risen. In essence, this task (i.e., given a text T, determine if T is a deepfake text or human-written text) resembles so-called *Turing Test*[2] or binary classification in Machine Learning.

In this work, we investigate this task of determining if a given text is a deepfake text or not. While the field of open-ended text generation is still relatively new, both computational and non-computational detection of deepfake texts have been extensively studied in recent years and well surveyed in Uchendu et al. [35]. In particular, what we are interested in answering is how "humans" are able to detect deepfake texts better. Clearly understanding human capacity and their limitations in detecting deepfake texts would help the development of both computational and non-computational (and even hybrid) tools for detecting deepfake text better. Recent literature (e.g., [5, 6, 10, 37]) has shown that, by and large, humans are *not* good at detecting deepfake texts, performing only slightly better than the level of random guessing. Even if humans are trained to detect deepfake texts, the performance has not improved significantly (e.g., [6, 9, 34]).

Therefore, in this work, we aim to find ways to improve humans in detecting deepfake texts better and understand human factors at play. Especially, we wonder if individual humans, trained or not, are not good at detecting deepfake texts, does their collaboration or expertise matter? As such, we pose the following three research questions (RQs):

RQ1 Do collaborative teams/groups perform better than individuals in deepfake text detection?
RQ2 Do experts perform better than non-experts in deepfake text detection?
RQ3 What are the factors that maximize the performance gain?

RQ1 aims to investigate what improves human performance from the baseline - team collaboration or individuals, and if collaboration improves human performance significantly, which collaboration technique matter - synchronous or asynchronous collaboration? The hypothesis here is that synchronous collaboration will improve human performance in deepfake text detection because humans

[1] This is also known as *neural texts*.
[2] 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.

perform better when there is informal discussion and sharing of ideas,as shown in prior literature (e.g., [26], [18], [28]). Given the benefits of collaboration, we hypothesize that collaboration will improve human performance. **RQ2** aims to investigate how English experts vs. English non-experts detect deepfake texts differently. English experts are defined as individuals with at least a Bachelor's degree in English (and related programs). We aim to investigate the characteristics that make one or more settings significantly outperform others. An example here is that we hypothesize that experts will focus more on high-level errors such as logical fallacies and non-experts will focus more on low-level errors such as grammar issues. **RQ3** aims to investigate the human factors that may help improve human performance in deepfake text detection. See Figure 1(C) for a visual representation of the research questions.

Finally, our main contribution is investigating human participants' ability to detect deepfake texts with different settings – non-expert vs. expert and individual vs. collaborative. Our key findings are summarized as follows: (1) both expert and non-expert in the individual settings outperform the baseline significantly; (2) experts improve significantly from individual to collaborative settings; and (3) experts more frequently use strong indicators of deepfake texts as justification (which explains their superior performance).

2 RELATED WORK

2.1 Automatic Evaluation of Deepfake Texts

As Neural Text Generators (NTGs) such as GPT-2 can be maliciously used to generate misinformation at scale, several techniques have been employed to detect deepfake texts. Using *stylometric*[3] classifiers, researchers adopted stylometry from traditional authorship attribution solutions to achieve automatic deepfake text detection [14, 36]. However, due to the flaws of *stylometric* classifiers, *deep-learning* techniques have been proposed [1, 3, 19–21, 39]. 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 [15, 16, 29, 30]. 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 et al. [35] defines these classifiers as *hybrid* classifiers and they achieve superior performance [23, 24, 41]. 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 NTGs 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.

2.2 Human Evaluation of Deepfake Texts

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. This has been studied in clever and nuanced ways which include training and not training.

2.2.1 Human Evaluation Without Training. GROVER [39], a 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

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

are more believable and GROVER-generated fake news was found to be more trustworthy [39]. Donahue et al. [8] recruits human participants from Amazon Mechanical Turk (AMT) to detect machine-generated words in a sentence. Uchendu et al. [37] also recruits human participants from AMT and asks them to detect which one of two articles is machine-generated and given one article, decide if it is machine-generated or not. Ippolito et al. [20] evaluates the human ability to perform comparably given 2 different generation strategies. Brown et al. [5] 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.

2.2.2 Human Evaluation with Training. Since human performance in deepfake text detection is very poor, researchers proposed improving there experimental framework by training human participants before the task. Humans without training achieve a 54% accuracy in distinguishing GPT-2-generated texts from human-written texts [16]. To improve performance, GLTR (Giant Language Model Test Room), a color-coded tool is proposed. GLTR color codes words based on the distribution level which improves human performance from 54% to 72% [16]. Dugan et al. [11] 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 [10]. Our framework is modeled more closely after Dugan et al. [11]'s work. Next, Clark et al. [6] proposes 3 training techniques - *Instruction-based, Example-based,* and *Comparison-based*. *Example-based* training improved the accuracy from 50% to 55% [6]. Lastly, Dou et al. [9] recruits human participants to annotate the error types of machine-generated texts. Participants were evaluated on an extensive qualification task which trains them [9]. A score ≥ 90 out of 100 is considered a pass so the participant can move to the next round.

Finally, 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. NTGs [37]. This means that while GLTR outperformed all other methods, it was built in 2019, and the results from Uchendu et al. [37] suggest that newer NTG render older deepfake text detectors inferior/obsolete. We hypothesize that previous techniques to improve human performance failed because they did not consider that collaboration and skill levels could affect performance. Based on the skill levels of humans, they will understand hints and potentially use them differently. For instance, if given a task to highlight the grammar issues in a piece of text, a person with college or post-college level of reading & writing will find higher-level grammar errors (e.g., re-worded repetition, run-on sentences) than a person with 11th-grade level. Furthermore, the point of collaboration is to encourage the exchange of ideas which could de-mystify the task for humans and improve performance. Therefore, while we implement the *example-based* training technique, we also improve human performance by incorporating collaboration.

3 METHODOLOGY

To improve human performance in deepfake text detection, we first, define a realistic problem - detecting deepfake texts in an article authored by both human and an AI (i.e., GPT-2). This is done by randomly replacing 1 out 3 human-written paragraphs with a GPT-2-generated paragraph. Next, we define 2 variables for our study - *individual vs. collaboration* and *non-expert vs. expert*. After using these variables to facilitate crowdsoursing recruitment, we ask the human participants to select 1/3 paragraphs that is deepfake and provide justification for selection.

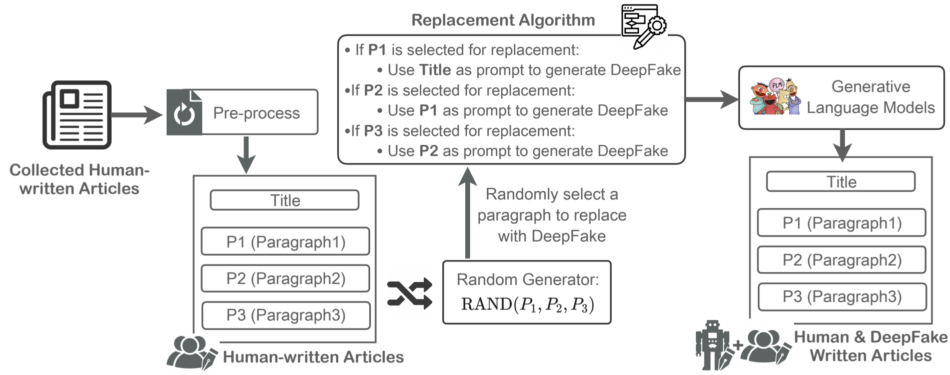

Replacement Algorithm
• If P1 is selected for replacement:
• Use Title as prompt to generate DeepFake
•If P2 is selected for replacement:
• Use P1 as prompt to generate DeepFake
•If P3 is selected for replacement:
• Use P2 as prompt to generate DeepFake
Collected Human-written Articles
Pre-process
Title
P1 (Paragraph1)
P2 (Paragraph2)
P3 (Paragraph3)
Human-written Articles
Randomly select a paragraph to replace with DeepFake
Random Generator: RAND(P_1, P_2, P_3)
Generative Language Models
PLM
Title
P1 (Paragraph1)
P2 (Paragraph2)
P3 (Paragraph3)
Human & DeepFake Written Articles

Fig. 2. Illustration of the data generation process

3.1 Data Generation

Our task is modeled similarly to the *Turing Test* problem defined by Alan Turing in the 1950s. The *Turing Test* is a test administered by a human as the judge who has a conversation with an unknown entity and decides if they are speaking with a human or machine/AI model. If the machine is labeled as human, then the machine has passed the test. However, for our task, due to the security risks deepfake texts pose, it is imperative that the NTG does not pass the test. To that end, we propose a framework that increases the probability of GPT-2 failing the *Turing Test*.

Therefore, we first define the goal of distinguishing deepfake texts from human-written texts. As this problem has been studied by several researchers [6, 11, 16, 20, 36] and shown to be very non-trivial and difficult to solve, we develop a novel way of administering the *Turing Test*. Thus, we implement a realistic setting of the problem - *detecting deepfake texts in an article authored by both humans and an NTG*. This setting is motivated by the fact that while NTGs currently have very impressive generations, humans still produce more natural speech than NTGs. This means that it will be natural for humans to edit machine-generated articles to make them sound more authentic. We study this non-trivial problem by asking human evaluators to: (1) Of 3 paragraphs, two written by a human, and one machine-generated, select the machine-generated paragraph. (2) Please check all explanations that satisfy the reason(s) for your choice. See Figure 1(B) for a visual representation of our task. We provide seven pre-defined rationales that correspond to flaws typically observed in deepfake text [10] - 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) or choose to write on their own.

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 AI), we randomly select 1/3 paragraphs to be replaced by GPT-2's [32] generated texts. We used only GPT-2 to generate the deepfake texts because: (1) GPT-2 and GPT-3 are similar. Based on [6, 36], 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 since GPT-2 is open-source and GPT-3 is not. For generation, we used GPT-2 XL

Label	Paragraph1	Paragraph2	Paragraph3
Count	23	16	11

Table 1. Data labels of deepfake texts

which has 1.5 billion parameters and *aitextgen*[4], a robust implementation of GPT-2 to generate texts with the default parameters. We followed the following replacement process:

(1) If paragraph 1 is to be replaced: Use Title as a prompt to generate GPT-2 replacement
(2) If paragraph 2 is to be replaced: Use Paragraph 1 as a prompt to generate GPT-2 replacement
(3) If paragraph 3 is 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 to choose the best GPT-2 replacement that fits well with the article. We use a BERT-base [7] as the Masked Language model to get the probability 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

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

We use a random number generator to select which paragraphs are to be replaced and got the following deepfake text replacement for the paragraphs in Table 1. 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 Figures 2 for the data generation process and 1(A) for an example of the final multi-authored article.

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.

Finally, we 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.

3.2 Participant Recruitment

3.2.1 AMT.

Inspired by Clark et al. [6], Dugan et al. [11], and Van Der Lee et al. [38], we used Amazon Mechanical Turk (AMT) to collect responses from non-expert evaluators. We deployed a two-stage

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

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 2. Upwork participant demographics

process to conduct the non-expert human studies. First, we posted a *Qualification* Human Intelligence Task (HIT) that pays $0.50 per assignment on MTurk to recruit 240 qualified workers In terms of the qualification requirements, in addition to our custom qualification used for worker grouping, three built-in worker qualifications are used in all the HITS, including *i)* HIT Approval Rate (≤98%), Number of Approved HITs (≥3000), and Locale (US Only) Qualification.

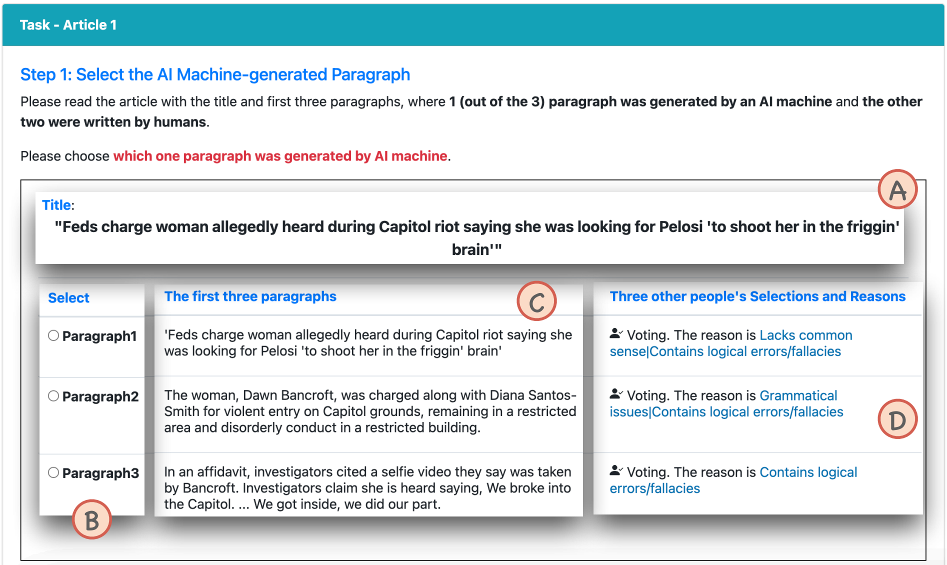
Next, we only enable the qualified workers to enter the large-scale labeling tasks. The approximate time to finish each labeling task is around 5 minutes (*i.e.,* the average time of two authors on finishing a random HIT). Therefore, we aim for $7.25 per hour and set the final payment as $0.6 for each assignment. Further, we provide "double-payment" to workers who made correct submissions as the extra bonus.

3.2.2 Upwork.

We utilized Upwork to recruit expert evaluators, especially those with expertise in writing domains.Upwork is one of the leading freelance websites with a substantial network size: Upwork generates 40 million monthly visits on average, and its gross services volume reached 3.5 billion dollars in 2021.[5] It has facilitated the freelance industry by introducing skilled freelancers in diverse categories like writing, graphic design, and web development. With its automated recommendation system, Upwork is capable of effectively matching clients and workers based on their needs.

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 questions we anticipated that they will solve. 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] https://sellcoursesonline.com/Upwork-statistics



Task - Article 1

Step 1: Select the AI Machine-generated Paragraph

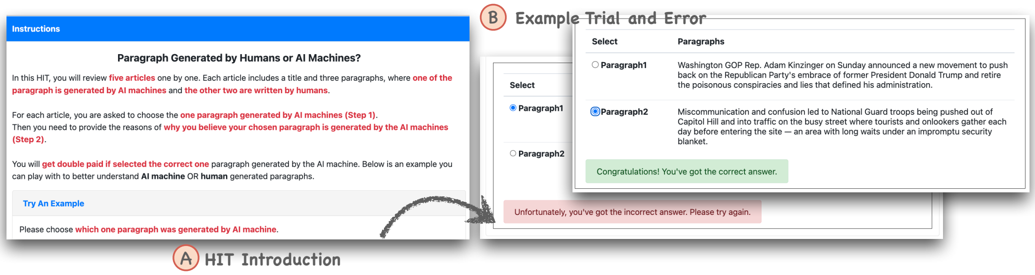
Please read the article with the title and first three paragraphs, where **1 (out of the 3) paragraph was generated by an AI machine** and **the other two were written by humans**.

Please choose **which one paragraph was generated by AI machine**.

Title:
"Feds charge woman allegedly heard during Capitol riot saying she was looking for Pelosi 'to shoot her in the friggin' brain'"

Select	The first three paragraphs	Three other people's Selections and Reasons
○ Paragraph1	'Feds charge woman allegedly heard during Capitol riot saying she was looking for Pelosi 'to shoot her in the friggin' brain'	Voting. The reason is Lacks common sense|Contains logical errors/fallacies
○ Paragraph2	The woman, Dawn Bancroft, was charged along with Diana Santos-Smith for violent entry on Capitol grounds, remaining in a restricted area and disorderly conduct in a restricted building.	Voting. The reason is Grammatical issues|Contains logical errors/fallacies
○ Paragraph3	In an affidavit, investigators cited a selfie video they say was taken by Bancroft. Investigators claim she is heard saying, We broke into the Capitol. ... We got inside, we did our part.	Voting. The reason is Contains logical errors/fallacies

Fig. 3. User Interface for the AMT Collaborative Group workers to choose the machine-generated paragraph.



Instructions

Paragraph Generated by Humans or AI Machines?

In this HIT, you will review **five articles** one by one. Each article includes a title and three paragraphs, where **one of the paragraph is generated by AI machines** and **the other two are written by humans**.

For each article, you are asked to choose the **one paragraph generated by AI machines (Step 1)**.
Then you need to provide the reasons of **why you believe your chosen paragraph is generated by the AI machines (Step 2)**.

You will **get double paid if selected the correct one** paragraph generated by the AI machine. Below is an example you can play with to better understand **AI machine** OR **human** generated paragraphs.

Try An Example

Please choose **which one paragraph was generated by AI machine**.

Select	Paragraphs
○ Paragraph1	Washington GOP Rep. Adam Kinzinger on Sunday announced a new movement to push back on the Republican Party's embrace of former President Donald Trump and retire the poisonous conspiracies and lies that defined his administration.
● Paragraph2	Miscommunication and confusion led to National Guard troops being pushed out of Capitol Hill and into traffic on the busy street where tourists and onlookers gather each day before entering the site — an area with long waits under an impromptu security blanket.

Congratulations! You've got the correct answer.

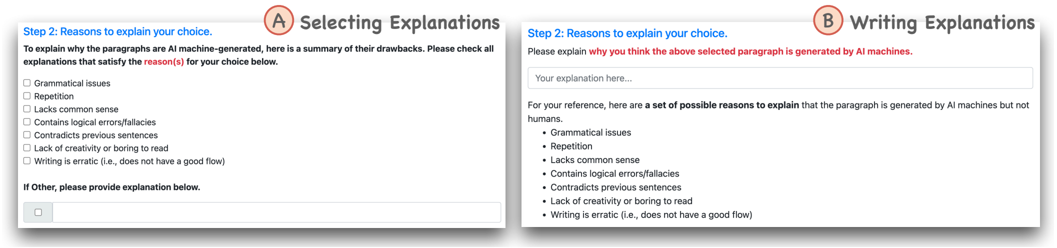
Select
● Paragraph1
○ Paragraph2

Unfortunately, you've got the incorrect answer. Please try again.

Fig. 4. The instructions to train users by providing prompt feedback.

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, 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 2 gives the self-reported demographic breakdown of recruited Upworkers.



Step 2: Reasons to explain your choice.

To explain why the paragraphs are AI machine-generated, here is a summary of their drawbacks. Please check all explanations that satisfy the reason(s) for your choice below.

□ Grammatical issues
□ Repetition
□ Lacks common sense
□ Contains logical errors/fallacies
□ Contradicts previous sentences
□ Lack of creativity or boring to read
□ Writing is erratic (i.e., does not have a good flow)

If Other, please provide explanation below.

□

Step 2: Reasons to explain your choice.

Please explain why you think the above selected paragraph is generated by AI machines.

Your explanation here...

For your reference, here are a set of possible reasons to explain that the paragraph is generated by AI machines but not humans.

- Grammatical issues
- Repetition
- Lacks common sense
- Contains logical errors/fallacies
- Contradicts previous sentences
- Lack of creativity or boring to read
- Writing is erratic (i.e., does not have a good flow)

Fig. 5. *Select* (A) vs. *Write* (B) justification question type

3.3 Experiment Design

3.3.1 AMT.

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 the group1 as *Individual Group*, in which each worker was asked to select the machine-generated paragraph without any references. See Figure 3, for example, humans in *Individual Group* can only see the introduction with panel (A) (B) and (C). On the other hand, we design the 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 the selection results from the group1 in an asynchronously manner, as the example shown in Figure 3(D), to support their own selection.

In addition, we investigate the capability of Individual vs. Collaboration of non-expect human participants to improve human performance in deepfake text detection. To do this, we compare 2 ways the human participants can provide justification for their answer - *Select* & *Write*. For the *select* setting, the question type was inspired by RoFT [11], 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. However, as this may be limiting because we pre-define the justifications, we also investigated another question type - *write*. In this setting, participants were asked to provide their reasoning. To help, we also share the list of justifications in the *select* setting to give the participants an idea of what justifications look like. See Figure 5 for *select* and *write* AMT interface.

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. [6]'s human evaluation study and was found to be the most effective training technique. *ii)* We pay double compensations to the workers who provide correct answers as mentioned in Section 3.2.1. This aims to encourage workers to get high accuracy on 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.

3.3.2 Upwork.

Given that we aim to compare experts' deepfake text detection accuracy with respect to individual vs. collaborative settings, our Upwork study consists of two sub-experiments. 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. The user interface was equivalent to the *select* scenario in Figure 5. Upwork participants were given one week to complete the survey. Upon completion, we randomly grouped 3 participants per team, resulting 6 teams in total for synchronous collaboration (Table 2). All discussions were conducted on the video communications software - Zoom[7] and we leveraged Zoom's built-in audio transcription feature, which is powered by Otter.ai[8] 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.

3.4 Analysis Methods

To investigate RQ1 and RQ2, we first quantitatively compare the performance of human participants (Section 4.2). This is done by measuring the mean accuracy of participants at the article level. Next, we categorize their provided rationales and compare their distributions based on correct and incorrect responses to support RQ3 (Section 4.3 & 4.4). Consistent with RQ1 and RQ2 analyses, for RQ3, we compare the frequency of justifications across *individual vs. collaboration* and *non-experts (AMT) vs. experts (Upwork)* settings. In addition to quantitative examination, we conduct qualitative studies on discussion transcripts from Upwork participants to gain insights into how Upwork groups benefited more from collaboration (Section 4.5). Finally, the implications of preceding investigation are outlined in Section 5.

4 RESULTS

4.1 Performance Measurement

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: $acc_n = l_n/m_n * 100$. 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 an independent sample T-test. Since the T-test is grounded on the assumption of normality [17], we ran the Kolmogorov-Smirnov test on our data and confirmed that the requirement was satisfied. Following, we summarize the results of statistical testing.

4.2 Deepfake Text Detection Performance (RQ1 & RQ2)

4.2.1 RQ1: Individual vs. Collaboration. The baseline (i.e., randomly guessing) accuracy is 32% (i.e., 1 out of 3 paragraphs) for AMT participants. AMT-Individual improved from randomly guessing

[6] https://www.qualtrics.com
[7] https://zoom.us
[8] https://otter.ai

SETTING	Select		Write	
	Mean Accuracy	p-value	Mean Accuracy	p-value
Baseline vs. Individual	32% vs. 44.99%	0.0004	32% vs. 45.92%	0.002
Baseline vs. Collaboration	32% vs. 51.35%	0.00007	32% vs. 51.97%	0.00003
Individual vs. Collaboration	44.99% vs. 51.35%	0.187	45.92% vs. 51.97%	0.26

Table 3. T-test Results for AMT Experiments (RQ1)

SETTING	Select	
	Mean Accuracy	p-value
Baseline vs. Individual	31% vs. 56.11%	9.1e-12
Baseline vs. Collaboration	31% vs. 68.87%	7.3e-15
Individual vs. Collaboration	56.11% vs. 68.87%	0.008

Table 4. T-test Results for Upwork Experiments (RQ1)

SETTING	Select		Write	
	Mean Accuracy	p-value	MeanAccuracy	p-value
AMT-Individual vs. Upwork-Individual	44.99% vs. 56.11%	0.005	45.92% vs. 56.11%	0.028
AMT-Collaboration vs. Upwork-Collaboration	51.35% vs. 68.87%	0.002	51.97% vs. 68.87%	0.003

Table 5. T-test Results for AMT vs. Upwork (RQ2)

by achieving 45% and 46% in the *select* and *write* settings, respectively. AMT-Collaboration achieved 51% and 52% accuracy in the *select* and *write* settings, respectively. In terms of AMT experiments, both individual and team-based problem-solving significantly outperformed random guessing with p-value < 0.05 (Table 3). Still, a mean difference between individual and collaborative environments was found to be insignificant. Additionally, Upwork participants had a baseline of 31%. They achieved an accuracy of 56% and 69% accuracy for Individual and Collaboration, respectively. All Upwork results were found to be significant. See Tables 3 and 4 for the average accuracy of AMT and Upwork participants, respectively.

4.2.2 RQ2: Non-Experts (AMT) vs. Experts (Upwork). AMT participants represent the laypeople/non-experts while Upwork participants represent experts. For the Individual setting non-experts vs. experts achieved an accuracy of 45% vs. 56%. Also, for the Collaboration setting non-experts vs. experts achieved an accuracy of 51% vs. 69%. And since all p-value < 0.05 for baseline vs. individual, baseline vs. collaboration, and individual vs. collaboration, the results are statistically significant between non-experts and experts. See Table 5 for a detailed result of the T-test of human performance on deepfake text detection of non-experts vs. experts.

4.3 Justification Patterns - *select* (RQ3)

Using the findings of prior literature [6, 10], we used 7 justifications for participants to select a paragraph as deepfake: *grammar issues, repetition, lacks common sense, contains logical errors/fallacies,*

Justification Type	Correct				Incorrect			
	AMT		Upwork		AMT		Upwork	
	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value
Grammar	13.97 vs. 23.08	0.016	15.33 vs. 24.6	0.013	15.65 vs. 16.89	0.675	14.22 v.s 12.07	0.469
Repetition	6.73 vs. 6.69	0.986	4 vs. 6.4	0.266	8.53 vs. 5.62	0.113	1.67 vs. 2	0.725
Common Sense	9.25 vs. 15.48	0.036	13 vs. 28	9.67e-05	13.02 vs. 9.94	0.206	3.33 vs. 5.56	0.171
Logical Errors	11.64 vs. 10.24	0.589	7.78 vs. 14.4	0.006	18.54 vs. 7.7	4.66e-05	3.89 vs. 4	0.942
Self-Contradiction	9.35 vs. 5.57	0.092	7.67 vs. 14.8	0.011	18.01 vs. 6.7	1.53e-06	6.56 vs. 3.6	0.054
Lack of Creativity	12.87 vs. 13.49	0.843	8.33 vs. 7.6	0.734	16.9 vs. 14.13	0.322	8.11 v.s 3.6	0.004
Coherence	14.64 vs. 19.29	0.174	20.56 vs. 32	0.019	11.65 vs. 10.06	0.513	13.78 vs. 9.2	0.05
Other	0 vs. 0	N/A	12.22 vs.18.4	0.053	0 vs. 0	N/A	6.78 vs. 8.4	0.519

Table 6. T-test Results of *select* Justification Frequency w.r.t. Correctness (Individual vs. Collaboration)

contradicts previous sentences, lack of creativity or boring to read, writing is erratic/incoherent. If none of 7 error types are found suitable, participants choose an additional justification, *other*, and write their own justification. To compare the frequency of justifications regarding two variable pairs (individual vs. collaboration & non-experts vs. experts), we first compute the frequency at which each of 4 groups cited a justification category. We calculate the overall frequency of justification and the frequency of justification used for incorrect and correct responses. Lastly, we calculate the statistical significance test with the independent sample T-test for these error types.

4.3.1 Individual vs. Collaboration. Figure 6 and 7 display the justification frequency based on correct and incorrect responses, respectively. Table 6 details the significance scores of individual vs. collaboration comparison. For the correct responses, the top-3 dominant categories that AMT mentioned were 'grammar', 'common sense', and 'coherence', and 'lack of creativity'. The top-3 least frequent justifications for correct responses that AMT participants used are 'other', 'repetition', and 'self-contradiction'. During the experiment, none of them selected 'other' to provide explanations for their choices that did not fall into the main 7 existing types.

Three categories ('grammar', 'common sense' and 'coherence') had the most increase in use from Individual to Collaboration, but only a 9.11% increase for grammar error type and a 6.23% increase for common sense was found to be significant (p = 0.016, p = 0.036).

Next, for incorrect responses, AMT participants frequently used 'grammar', 'logical errors', 'self-contradiction', and 'lack of creativity'. The frequency of all justifications reduced from Individual to Collaboration, except for 'grammar'. Among those categories, a drop in 'logical errors', and 'self-contradiction' are statistically significant.

The most common reasons Upwork participants gave for their correct responses, regardless of the presence of discussion, are 'coherence', 'grammar', and 'common sense'. They also chose 'other' considerably often. Our manual inspection of written justifications under the 'other' category revealed that off-topic and off-prompt were the most frequent ones. Similar to AMT workers, Upwork participants least frequently cited 'repetition' to justify their decisions. As opposed to AMT, we find that when they collectively solved the task the overall frequency of providing particular reasons surged. Specifically, the following justifications exhibited a significant spike in frequency: 'grammar' (+9.27%, p = 0.013), 'common sense' (+15%, p < 0.0001), 'logical errors' (+6.62%, p = 0.006), 'self-contradiction' (+7.13%, p = 0.011), 'coherence' (+11.44%, p = 0.019) types. 'Lack of Creativity' is the only category that was used less often in collaborative problem-solving than individual problem-solving, but it was statistically insignificant.

Next, the top-3 most frequent justifications are 'grammar', 'coherence', and 'lack of creativity' for individual-based responses from Upwork participants regarding their incorrect answers. Also, all

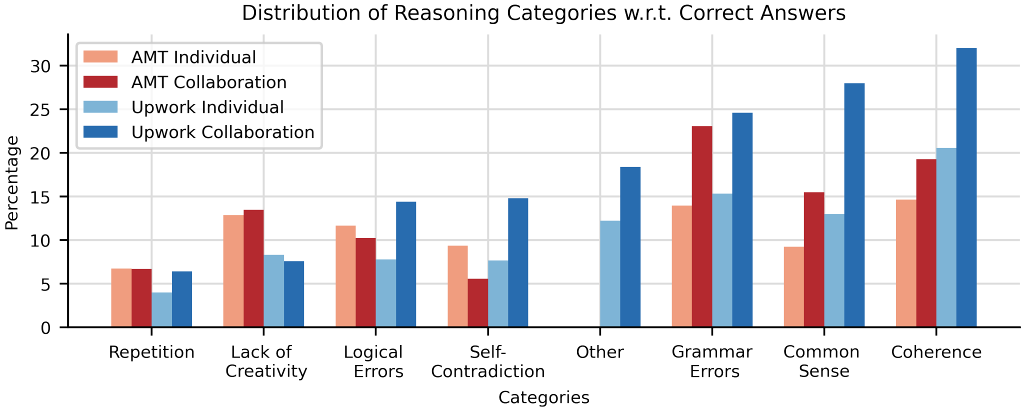


Fig. 6. Justification Category Distribution w.r.t. Correct Responses

justification types, excluding 'repetition', 'common sense', 'logical errors', and 'other', reduced in frequency from Individual to Collaboration. The surge in the frequency of four justification types was not statistically supported. Finally, 'lack of creativity' and 'coherence' categories were less prevalent during the team-based approach in comparison to the individual-based approach.

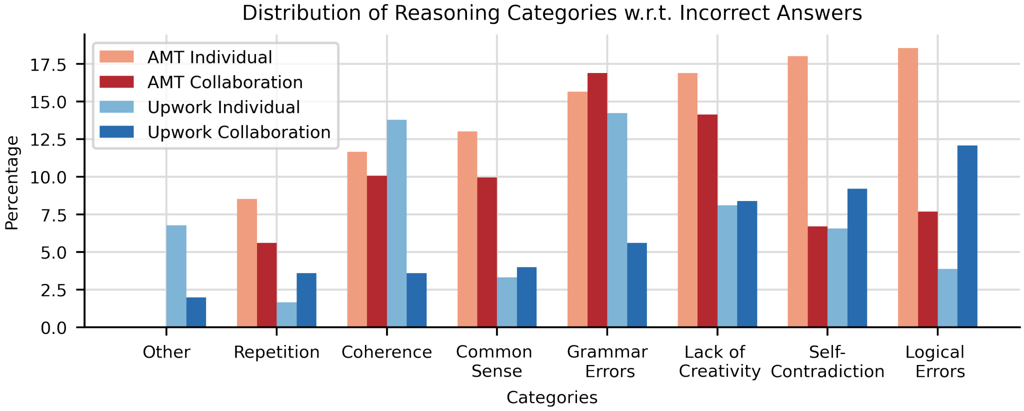


Fig. 7. Justification Category Distribution w.r.t. Incorrect Responses

4.3.2 Non-Experts vs. Experts. We now report the results of comparative analyses in terms of AMT-Individual vs. Upwork-Individual and AMT-Collaboration vs. Upwork-Collaboration, respectively (Table 7). AMT participants are non-experts while Upwork participants are experts.

Our analyses of correct responses reveal that 'grammar' and 'coherence' were the most commonly submitted rationales, regardless of individuals' expertise. The least frequent justification for both AMT- and Upwork-Individuals was 'repetition'. While Upwork participants cited 'grammar', 'common sense', 'coherence' and 'other' more often than AMT participants, those differences were found to be insignificant, except for 'other' (0% vs. 12.22%, $p < 0.0001$). This is because AMT participants

Justification Type	Correct				Incorrect			
	Individual		Collaboration		Individual		Collaboration	
	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value
Grammar	13.98 vs. 15.33	0.57	23.08 vs. 24.6	0.747	15.65 vs. 14.22	0.53	16.89 vs. 12.07	0.175
Repetition	6.73 vs. 4	0.107	6.69 vs. 6.4	0.919	8.54 vs. 1.67	3.46e-07	5.62 vs. 2	0.029
Common Sense	9.25 vs. 13	0.086	15.48 vs. 28	0.004	13.02 vs. 3.33	6.49e-07	9.95 vs. 5.6	0.06
Logical Errors	11.64 vs. 7.77	0.056	10.24 vs. 14.4	0.151	18.54 vs. 3.89	1.14e-10	7.7 vs. 4	0.089
Self-Contradiction	9.35 vs. 7.67	0.398	5.57 vs. 14.8	0.002	18.01 vs. 6.56	4.7e-08	6.7 vs. 3.6	0.097
Lack of Creativity	12.87 vs. 8.33	0.026	13.49 vs. 7.6	0.071	16.9 vs. 8.11	1.17e-05	14.13 vs. 3.6	7.26e-05
Coherence	14.64 v.s 20.56	0.066	19.29 vs. 32	0.011	11.65 vs. 13.78	0.283	10.06 vs. 9.2	0.75
Other	0 vs. 12.22	1.1e-11	0 vs. 18.4	1.11e-09	0 vs. 6.78	6.5e-08	0 vs. 8.4	0.0003

Table 7. T-test Results of *select* Justification Frequency w.r.t. Correctness (Non-experts vs. Experts)

never described their rationale in writing form. Meanwhile, the frequency of 'lack of creativity' (12.87% vs. 8.33%) was significantly lower for Upwork-Individuals than for AMT-Individual s (p = 0.0264).

While 'grammar' and 'coherence' remain as popular categories both AMT and Upwork participants cited for their correct answers even in the collaborative environment, the 'common sense' justification was cited more often than 'grammar' in Upwork-Collaboration. Moreover, reasons such as 'contradicts previous sentences' (+9.23%, p = 0.002), 'lacks common sense' (+12.7%, p = 0.011), and 'coherence' (+12.51%, p = 0.004) were more strongly associated with correct responses in Upwork-Collaboration compared to AMT-Collaboration.

For incorrect responses, 'logical errors', 'self-contradiction', and 'lack of creativity' were top-3 dominant justifications that AMT-Individual s provided, whereas Upwork-Individuals frequently chose 'grammar', 'coherence' and 'lack of creativity'. All justification categories, except for 'coherence' and 'other', decreased in frequency from AMT-Individual s to Upwork-Individuals. Yet, the observed drop in 'grammar' was found to be insignificant with $p > 0.05$. Also, despite an increase in the mentions of coherence, the gap was statistically insignificant. Grammatical errors were mentioned the most within erroneous answers in the collaborative context by both AMT and Upwork participants. Yet, 'repetition' (-3.63%, p = 0.029) and 'lack of creativity' (-10.53%, $p < 0.0001$) were the only categories that demonstrated a substantial drop in frequency from AMT-Collaboration to Upwork-Collaboration.

4.4 Justification Patterns - *write* (RQ3)

Unlike the *select* setting, the *write* setting does not offer any selection alternatives, allowing us to conduct more comprehensive studies that are not confined to the 7 error types (excluding 'other'). Motivated by Clark et al. [6]'s findings that failing participants concentrated on the form of the text rather than content, we manually inspected and categorized the justifications into three subsets: low-level, high-level, and hybrid. A low-level justification is related to the format, style, and tone of the text; high-level justification is an error type that can be determined based on the text's meaning; and hybrid justification represents a case where evaluators cited both low-level and high-level justifications. For annotation, one researcher initially created a codebook (Table 8) using 7 pre-defined justification categories and coded written responses submitted by AMT workers. We integrated new justifications into the code book as we noticed new information in the data. Next, two additional researchers independently labeled them. Lastly, Fleiss' Kappa coefficient [13] was used to calculate the agreement amongst three annotators (0.924 for individuals & 0.94 for collaboration), supporting the reliability of the generated labels.

Justifications	Code
Writing is erratic (i.e., does not have a good flow)	Low
Grammatical issues	
Repetition	
Sentence structure issues	
Lacks common sense	High
Lack of creativity or boring to read	
Contains logical errors/fallacies	
Contradicts previous sentences	
Too much/too little information	
Off-prompt	
Off-topic	
Incorrect Information	
low-level+ high-level	Hybrid

Table 8. Code book for error level annotation

Justification Level	Correct				Incorrect			
	AMT		Upwork		AMT		Upwork	
	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value
Low	19.54 vs. 21.08	0.675	21.89 vs. 20.4	0.668	24.95 vs. 19.78	0.121	23.11 vs. 12.8	0.0009
High	20.94 vs. 21.88	0.812	20.33 vs. 19.13	0.744	25.75 vs. 24	0.68	15.11 vs. 8.4	0.006
Hybrid	5.56 vs. 9.67	0.103	13.67 vs. 28.67	0.0003	3.27 vs. 3.6	0.825	5.67 vs. 9.6	0.067

Table 9. T-test Results of Justification Level Frequency w.r.t. Correctness (Individual v.s. Collaboration)

As Upwork experiments do not have the writing setting, we relied on the justifications provided in the *select* style. This time we also annotated writings submitted in the 'Other' type. Following, we describe the T-test results which were used to test the distribution differences across three labels.

4.4.1 Individual vs. Collaboration. As shown in Table 9, AMT participants' correct responses were more strongly associated with either low or high error levels rather than both when they solved the task individually. In the collaborative setting, their mentions of hybrid reasoning surged as well as low and high-level rationales. Yet, none of the increases was statistically significant. For incorrect answers, 25% of AMT individuals on average used low-level or high-level justifications to explain their judgments. Meanwhile, a relatively smaller percentage of participants (3%) provided both low and high-level justifications. When they collectively performed the task, the frequency of low-level rationale decreased to 20%. Still, the drop was found to be insignificant. Also, a change in high and hybrid-level errors was minimal.

Next, Upwork participants, similar to AMT participants cited one error type more often than both for correct responses in individual problem-solving. The frequency of hybrid reasoning almost doubled and became the most frequent justification type after group discussions, with $p < 0.05$. A decrease in low or high levels, on the other hand, was not significant. For incorrect responses, low-level errors were most frequently mentioned when Upwork participants performed the task individually. Per question, 23.11% of participants on average utilized low-level errors to justify their decisions. In a collaborative environment, its percentage dropped to 12.8%, and this drop was statistically significant. Similarly, we observe that the shift in frequency of high-level justifications—from

Justification Level	Correct				Incorrect			
	Individual		Collaboration		Individual		Collaboration	
	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value	Mean Accuracy	p-value
Low	19.54 vs. 21.89	0.482	21.08 vs. 20.4	0.858	24.95 vs. 23.11	0.558	19.78 vs. 12.8	0.032
High	20.94 vs. 20.33	0.838	21.88 vs. 19.13	0.545	25.75 vs. 15.11	0.003	24 vs. 8.4	8.9e-06
Hybrid	5.56 vs. 13.67	0.0007	9.67 vs. 28.67	1.3e-05	3.27 vs. 5.67	0.083	3.59 vs. 9.6	0.008

Table 10. T-test Results of Justification Level Frequency w.r.t. Correctness (Non-Experts v.s. Experts)

Code	Description
Grammar	mentions of the spelling and grammar of the text
Repetition	mentions of words/phrases/content being repetitive
Factuality	mentions of whether the text describes things that are true
Consistency	mentions of how the text relates to the context and other pieces of the text
Common sense	mentions of whether the text makes sense within the world that it is written
Coherence	mentions of the structure, wording, or coherence of the text
Self-Contradiction	mentions of whether the text contradicts itself
Creativity	mentions of whether the text seems boring to read
Writers' capabilities	mentions of writer's intent or capabilities

Table 11. Annotation categories for Upwork transcript

15.11% to 8.4%—was significant. Hybrid-level errors, on the contrary, became more frequent after the collaboration but were not substantial.

4.4.2 Non-Experts vs. Experts. Table 10 includes T-test results of AMT vs. Upwork experiments. In both individual and collaborative problem-solving environments, the proportion of people who stated low- or high-level reasoning for their correct responses did not differ that much across AMT and Upwork studies. Specifically, all percentages associated with low- or high-level ranged from 20%-23%, regardless of participants' linguistic abilities. Meanwhile, the frequency of hybrid reasoning was 8.11% greater for Upwork than for AMT when the task was performed independently, and the measured gap was statistically significant. When participants complete the task together, the difference between AMT and Upwork in the frequency of both low- and high-level reasons increased from 8.11% to 19%, yielding greater statistical significance.

Now, for incorrect responses, AMT-Individual participants provided the low-level error as their justification slightly more often than Upwork-Individual participants, but the gap is negligible with $p > 0.05$. Likewise, the observed difference in hybrid-level errors was not substantial. Upwork had a much lower frequency of high-level explanations than AMT, and this result was statistically significant. We also discover that in the collaborative setting, AMT participants cited more low- and high-level experiments than on Upwork participants. Especially, the rate of high-level justifications was approximately 3 times greater in the AMT setting compared to the Upwork setting. Hybrid-level errors, on the contrary, were more common in Upwork, and the difference was significant.

4.5 Qualitative Analysis of Upwork Transcripts

To better understand why the synchronous collaboration of Upwork evaluators yielded better performance than other scenarios, we run a transcript analysis. Transcript analysis is a common methodology that has been applied in the HCI field to qualitatively examine discourse content or

Upwork Groups	G1	G2	G3	G4	G5	G6
Fleiss' Kappa	0.6078	0.7890	0.8883	0.8745	0.8602	0.7764

Table 12. The Fleiss' Kappa score of human annotation on the categories

patterns. To begin, we first construct a coding scheme covering the mentions of syntactic or semantic elements of texts (Table 11). We primarily follow categories from [6, 10]. The annotation process was conducted by one researcher (responsible for assigning the codes) and supported by another researcher (responsible for double-checking whether the code assignment made sense, i.e., if a code would fit to a given statement). Computed Fleiss's Kappa values are shown in Table 12.

Following, we introduce three key strategies that were commonly adopted by Upwork-Collaboration participants: *deductive reasoning*, *advanced linguistic skills*, and *leveraging their prior experiences*. We further support them with selected quotes.

4.5.1 Deductive reasoning. Deduction, by definition, produces valid conclusions, which must be true if the premises are true [22]. Based on the transcript analyses, Upwork participants preferred to use deductive reasoning to come to a conclusion; they first attempted to come up with reasons why 2 out of 3 paragraphs could not be deepfake, and then picked the remaining one as deepfake. Specifically, there was an instance where P10 from G4 chose the paragraph that lacked common sense the most:

> *"Number two and three makes sense. Actually they all make sense, but it's just this first one that makes the least sense." - [P10]*

In this example, she hypothesized that human-written texts do not lack common sense. G2 members also sometimes used elimination strategies when the answer was not clear:

> *"All three have grammatical issues. I think it's I feel the best about number three like actually being part of the article." - [P4]*
> *"yeah. let's say one or two. The way the last sentence and paragraph three, is written, I feel like that's that was just written by a human getting short shrift." - [P5]*
> *"I agree. Let's go with two then." - [P6]*

They relied on their intuitions to make a final judgment after confirming all 3 paragraphs contain grammar errors. Although they did not provide detailed explanations, they could come to a consensus depending on their beliefs about human-written articles. Interestingly, P16 from G6 cited the following sentences when he applied the deductive logic:

> *"I have to admit, all three sound perfectly fine to me. I'd say one actually makes the most sense to me so I'm going to buy again process of elimination, I think one is. That could be the AI just because the other two people can't write very well." - [P16]*

He chose the paragraph that makes the most sense as deepfake due to a similarity in writing styles of the other two paragraphs. This could have been influenced by his focus on the consistency of texts.

4.5.2 Advanced linguistic skills. The quantitative findings showed that Upwork participants cited grammar errors very often as part of their justifications. Congruent with the fact that Upwork participants are proficient English speakers, their discussion around linguistic properties was complex and comprehensive. For instance, several participants (e.g., P4, P13) talked about the length of sentences and how they are not joined with the proper conjunction or punctuation:

> *"I mean, I feel like number two is just hard to read and long and confusing and if I were writing it or editing it I would want to trim it maybe turn it into more sentences." - [P4]*

> *"One of those calculations, paragraph one, has three of them I would assume that and they would be programmed to pump out the longer sentences." - [P13]*

Not limited to the identification of run-on sentences, they could capture more sophisticated grammatical errors resulting from tense shifts, dashes or capitalization. For example, P2 from G1 mentioned a missing dash, and P4 from G2 discussed inconsistent tense usage:

> *"I was wondering, to the Washington at the beginning, there should be a dash after." - [P2]*

> *"Studies on historical facts and policy talking about the Capitol Hill security issue is in the present tense and this is taking us to 13 and 2012." - [P4]*

Our quantitative analysis on justification frequencies (Section 4.3) revealed that Upwork participants frequently selected the 'other' option to cite 'off-topic' and 'off-prompt' as their rationales. Transcript annotation results further confirm their recurring mentions of consistency issues within given paragraphs. See these examples below:

> *"Why, I didn't get it up here well but yeah it is that is bad and nothing I didn't get about paragraph two is recently bringing the Russians in when the first paragraph, had nothing to do with Russia. " - [P2]*

> *"For this one, I thought it was paragraph two, just because that question seems a little random and irrelevant to the other paragraphs." - [P8]*

> *"Paragraph one is talking about North Korea, while paragraph two and three are talking about something else. So it seems like the AI was just gotten into that place without having information of what the rest of the article was talking about." - [P15]*

Another topic that Upwork participants actively discussed during the process of synchronous collaboration was text structure or awkward word selections. Specifically, P14 from G5 and P10 from G4 cited the following:

> *"I think this one I choose, I chose number two just because 'jumping too much for joy' seems like an odd statement. The word orders are unusual." - [P14]*

> *"Number three says, this was some way to pay back and so love for the country like this was some way, I thought that sounded a little odd." - [P10]*

These cases overall indicate Upwork participants' abilities to analyze a piece of text in greater depth and contextualize its relationship to other paragraphs.

4.5.3 Leveraging one's prior experiences. As could be expected given the professional backgrounds of Upwork participants, some of them shared their experiences with AI-powered writing and advised a course of action. For example, P17 from G6 said that:

> *"The thing is, I use AI writing to help me with my own writing and least the software that I use it's it does not make those mistakes when it comes to just like spelling (...). It seems like more human error is for it to be grammatically related because that's something we'd be making more mistakes compared to logic and fallacy and flow." - [P17]*

> *"If you give the AI prompt it'll start writing on that and I've had it happen, sometimes when it doesn't know what to write it just repeats back the crop." - [P17]*

> *"Since all of them have been clear grammatical errors we're supposed to be looking for something else that is Missing logically." - [P17]*

Another G2 member (P6) voiced his opinion on deepfake content:

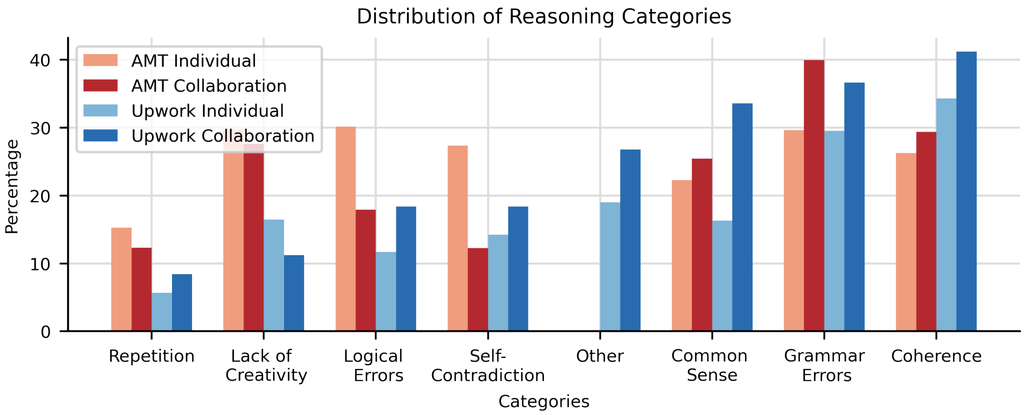


Fig. 8. The distribution of eight justification categories.

> *"I tried to generate machine-written content for the purpose of this research. And there's a grammatical errors that we find here, sometimes a I think a more human related than the machines, because (...)." - [P6]*

Following, P4 formulated a hypothesis and further reshaped her viewpoint accordingly:

> *"The thing so interesting what P5 is saying is like we are assuming that the most clearly written material will be written by humans and you're saying well, maybe we should assume that the most clearly written material is written by a computer. I would say that I rely a lot less on grammatical errors and more on logical errors like if something has a grammatical error." - [P4]*

The Upwork participants also anticipated certain writing styles from human-written texts. Using, this knowledge, they concluded that the paragraphs are from news articles that incorporate formal writing. For instance, P1, P3, and P13 mentioned as follows:

> *"yeah this was just filtering to me. Where would a machine get the details, but also this is so odd I feel like a human journalist when right this weird tangent about these people." - [P1]*

> *"I feel like there could be more like journalistic voice here, or it could have been presented in a more interesting way." - [P3]*

> *"Another reason I chose paragraph three is because of the 'so i'm sure' which usually isn't a way that any writer would begin." - [P13]*

In conclusion, the examples above suggest that Upwork participants took an advantage of their previous experiences and skills to complete the task. Further, the overall analysis supports the role of synchronous discussion in allowing participants to exchange stories and thoughts to appropriately reshape their perspectives.

5 DISCUSSION

5.1 Summary of Results

This paper entails a holistic examination of the performance of AMT *(laypeople)* and Upwork *(English professionals)* participants in detecting deepfake texts, as well as providing insights into the influence of asynchronous and synchronous collaboration on their performance. It also attempts to elucidate the relationship between various textual components and detection performance. Our major findings from Section 4 can be summarized as below:

Individual vs. Collaboration

(1) **Experts benefited from collaboration while non-experts did not**: While there is no difference in the deepfake text detection performance between AMT-Individual and AMT-Collaboration, Upwork-Collaboration's (69%) performance is significantly higher than that of Upwork-Individual (56%).

(2) **Experts' mentions of coherence, logical fallacies, and self-contradiction errors as justifications for deepfake text detection were significantly higher in the collaborative setting than the individual setting**: There are frequency differences between coherence, logical fallacies, and self-contradiction for Upwork-Individual and Upwork-Collaboration, with Collaboration more frequently citing these errors, whereas AMT experiments show no difference.

(3) **In contrast to experts, non-experts' usage of three error levels did not differ between individual and collaborative scenario**: While the frequency of low, high, and hybrid error levels did not differ between AMT-Individual and AMT-Collaboration, Upwork-Collaboration's mentions of the hybrid level was significantly more common than Upwork-Individual - (Individual vs. Collaboration usage percentage - 13.67 vs. 28.67 for correct responses).

Experts vs. Non-experts

(1) **Experts were better at deepfake text detection than non-experts in terms of detection accuracy**: Both AMT (45%) and Upwork (56%) Individual groups significantly outperformed the baseline (32% & 33%, respectively) performance. Still, Upwork participants outperformed AMT participants in both Independent and Collaborative environments.

(2) **Experts paid less attention to the creativity of articles than non-experts**: Upwork participants mentioned a lack of creativity less frequently than AMT participants for correct responses. For analysis of non-experts vs. experts (12.87 vs. 8.33 for correct and 13.49 vs. 7.6 for incorrect responses) frequent usage of creative errors as justification.

(3) **Experts put more emphasis on consistency issues in the articles than non-experts**: Upwork participants were capable of capturing consistency-related errors such as off-topic and off-prompt while AMT participants could not.

(4) **Experts relied on hybrid-level justifications more often than non-experts**: Upwork participants in general cited the hybrid-level justification more frequently than AMT participants for correct responses. For non-experts vs. experts analysis frequency comparison of hybrid-level errors in the collaboration setting, we have 9.67 vs. 28.67 for correct responses.

5.2 Implications

We discuss the implications of these summarized results below to extrapolate the reasoning or phenomena of the findings.

5.2.1 *Individual vs. Collaboration.*

(1) **Experts benefited from collaboration while non-experts did not**: It has long been established that the performance of a group may surpass that of even the most knowledgeable person [27]. However, we discovered that there was no performance difference between independent and asynchronous collaboration regarding AMT experiments, whereas Upwork-Individual participants not only detected significantly better than the baseline but also benefited from the synchronous collaboration. A reason for this could be because synchronous collaboration context, for example, could have encouraged workers to be more involved, creative, and social. A body of CSCW literature (e.g., [4, 25, 33]) has also argued that the gains of synchronous

collaboration outweigh the benefits of asynchronous collaboration. There is another possible reason to explain these mixed results. Since Upwork collaborators share the same background in English expertise, their familiarity with each other's fields and the advanced degree of individual intelligence may have positively impacted the group's intelligence.

(2) **Experts' mentions of coherence, logical fallacies, and self-contradiction errors as justifications for deepfake text detection were significantly higher in the collaborative setting than the individual setting**: Non-experts showed no pattern differences in coherence, logical errors, and self-contradiction justifications 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. [10]'s finding that machines are prone to fall short of those categories. Moreover, coherence errors and self-contradiction are considered high-level errors (Table 8). This implies that it is more challenging to find them and thus, advanced linguistic skills might have shown to be helpful. Therefore, it is expected that English professionals citing these errors more frequently than non-experts. 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. Regarding logical fallacy 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 texts.

(3) **In contrast to experts, non-experts' usage of three error levels did not differ between individual and collaborative scenarios**: We have three error levels - low (i.e., form of text) vs. high (i.e., content of text) vs. hybrid (i.e., form & content). We observe no patterns in non-experts' use of error levels between the independent and collaborative settings for both incorrect and correct answers. Given their similarities in error type usage, it is reasonable that non-experts' performance gain from the collaboration was marginal. The prevalence of hybrid-level errors, on the other hand, is more than doubled for correct responses between individual and collaborative settings of experts. This suggests that when experts collaborate, they are possibly more able to discuss more and find more errors than when they work independently. Furthermore, they could detect deepfake texts more effectively using these hybrid-level errors than they could individually. This hints that considering both form and content-wise errors is a useful strategy for deepfake text detection.

5.2.2 Experts vs. Non-experts.

(1) **Experts were better at deepfake text detection than non-experts**: As opposed to previous works where humans performed either at chance or below chance level [6, 9, 11, 37], both experts and non-experts from our studies exceeded the baseline performance. For the non-experts, we tested various interface designs to improve performance as well as the incorporation of *example-based training*. These experimental designs may have impacted the performance of non-experts. Given Upwork participants' profound experiences in writing domains, it is expected that they performed significantly better than laypeople like AMT workers. What is more interesting is that, despite additional training processes of non-experts, they were unable to outperform experts who had not received any training. Thus, these results indicate that existing training procedures are insufficient for non-experts, while experts may not need as much training as non-experts. Finally, more nuanced approaches that contain patterns that English professionals are likely to adopt should be explored further.

(2) **Experts paid less attention to the creativity of articles than non-experts**: While the statistical significance of the difference was partially supported, experts, in general, mentioned a lack

of creativity less frequently than non-experts for correct responses. Since experts significantly outperformed non-experts in both independent and collaboration scenarios, their lack of focus on the creativity error category suggests that it is not a strong indicator of deepfake texts. We also observe non-experts' frequent mentions of lack of creativity in both incorrect and incorrect responses, reinforcing our argument that lack of creativity is a misleading indicator of deepfake texts. Finally, our transcript analysis (Section 4.5) illustrates that UpWork participants were able to leverage their previous knowledge in writing domains and expected specific writing styles from the political news articles when making a decision. Since political news contain reporting of events/facts and are not fictional, their writing styles are expected to be formal and factual. In summary, creativity errors were misleading error types for our task possibly because the goal is for humans to detect deepfake texts in the politics-related news domain. Thus, future works are needed to draw further conclusions on other news topics.

(3) **Experts put more emphasis on consistency issues in the articles than non-experts**: Unlike non-experts, expert participants were able to detect off-topic and off-prompt errors. According to Badaskar et al. [2], language models find it challenging to stick to a single topic but cover diverse, often unrelated topics in a single text. This explains experts' enhanced performance over non-experts. Moreover, off-prompt/off-topic error detection is more likely to require careful reading and thinking than the detection of other error types. From Section 4.5, we also observe that Upwork participants were able to identify one paragraph that introduce a new topic, causing it to be off-prompt. As a whole, this demonstrates that English experts, as opposed to laypeople like AMT participants, can detect errors that go beyond the provided categories, such as off-topic and off-prompt. This also implies that consistency errors are good markers of deepfake texts.

(4) **Experts tended to rely on hybrid-level justifications more often than non-experts**: Analyzing non-experts vs. experts, the frequency of the hybrid-level error level was found to be statistically different. These results suggest that using a mix of both low-level and high-level errors rather than solely one yields more accurate detection performance. Hybrid-level errors require more careful analysis to cite compared to, for instance, repetition or grammar issues. Since low-level errors are easiest to detect but can lead to erroneous judgments, experts' abilities to look beyond obvious error types appear to be beneficial for deepfake detection tasks. Lastly, the results imply that experts' ability to reason deductively, their use of advanced linguistic skills, and leveraging their own prior experiences informed their use of hybrid-level errors. The discussions of experts in Section 4.5 showcased their ability to detect deepfake texts by relying on their professional skills. It also showed their ability to use linguistic cues that will have been otherwise missed by non-experts.

6 ETHICS, LIMITATIONS, AND FUTURE WORK

6.1 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.

6.2 Limitations

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, we believe that the choice of 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 [6, 37], 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, we believe that 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 human-like. This preserved the integrity of the experiments as the task remained non-trivial.

Next, for the *non-expert vs. expert* analysis, we compared AMT to Upwork, where AMT are the non-experts and Upwork, experts. However, since they are different forms of collaboration, AMT being asynchronous and Upwork being synchronous, the comparison may be different. Although a few prior works explored the space of synchronous collaboration between AMT workers, these systems tend to be engineering-heavy and are rarely used in real-world applications. We thus believe asynchronous collaboration is a more realistic way of using AMT. We understand that this design decision might compromise the comparability of the two settings but believe it is a reasonable trade-off. Furthermore, Upwork's interface supports the recruitment of English experts, as well as provides a framework for synchronous collaboration. To mitigate this limitation, we calculate the accuracy of AMT and Upwork participants in the same way (per-article accuracy) to have a fairer comparison.

6.3 Future Work

In the future, we aim to improve the AMT *write* setting framework by proposing a *Turing Test* framework that supports non-experts' creativity. Furthermore, using our framework, we will investigate and re-run all experimental studies on Upwork. This experiment will have the following variables - non-experts vs. experts, Individual vs. Collaboration, and asynchronous vs. synchronous collaboration. Next, we will improve the performance of collaborating for deepfake text detection using a highlight tool to color code the error types. As indicated above, hybrid-level errors, self-contradiction, and coherence are good indicators of deepfake texts. Therefore, our highlighting tool can be used to color codes these errors to reduce their cognitive load (especially for non-experts). Next, using this highlighting tool, we will train human participants before testing their ability to use such a tool to improve detection.

Finally, most human evaluation research on deepfake text detection setup the articles to be either fully human-written or fully deepfake. However, in the real world, deepfake articles will be edited by humans, creating a multi-authored article. Therefore, there are several ways to generate a more realistic dataset for this study. These include: (1) given a 3-paragraph article, 2 deepfake, and 1 human-written, ask human participants to select the human-written paragraph; (2) given a 3-paragraph article, each paragraph is generated by a unique NTG, ask humans select which paragraph is deepfake. The goal here is to see which of the state-of-the-art NTGs are easier for humans to detect; (3) given a 3-paragraph article, all written by humans, ask humans to select which is deepfake. The goal here is to see if humans have an affinity for selecting a particular paragraph (e.g. mostly the second paragraph).

7 CONCLUSION

In this paper, we studied human performance in deepfake text detection. To be more realistic, we built a 3-paragraph article with 1/3 paragraphs, machine-generated (deepfake) and 2/3 paragraphs, human-written. We ask human participants to select which paragraph is deepfake and to provide justification for their selection out of 7 error types. Specifically, we studied human performance with two variables - *individual vs. collaboration* and *English non-expert vs. English expert*. To achieve this, we recruit non-expert human participants from AMT and experts from Upwork. Furthermore, we run asynchronous collaboration with AMT and compared it to synchronous collaboration with Upwork. Finally, our results suggest that synchronous collaboration of expert human participants significantly improves human performance in deepfake text detection. We further identify several factors (such as coherence and consistency) that deepfake texts excel at or fail by analyzing their justification patterns. Lastly, the enhanced performance of participants (particularly non-experts) from baseline in the individual setting indicates that our *Turing Test* framework facilitates the improvement of humans' deepfake text detection performance.

REFERENCES

[1] Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 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.

[2] Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

[3] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351* (2019).

[4] Jeremy Birnholtz and Steven Ibara. 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.

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

[6] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 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)*. Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

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

[8] Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling Language Models to Fill in the Blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2492–2501.

[9] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294* (2021).

[10] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 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.

[11] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 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.

[12] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *Plos one* 16, 5 (2021), e0251415.

[13] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[14] Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science* 7 (2021), e443.

[15] Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and Distributional Detection of Machine-Generated Text. *arXiv preprint arXiv:2111.02878* (2021).

[16] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 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.

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

[18] Stefan Hrastinski. 2008. The potential of synchronous communication to enhance participation in online discussions: A case study of two e-learning courses. *Information & Management* 45, 7 (2008), 499–506.

[19] Huggingface. 2019. GPT-2 Output Detector Demo. *https://huggingface.co/openai-detector/* (2019).

[20] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 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*. Association for Computational Linguistics, Online, 1808–1822. https://doi.org/10.18653/v1/2020.acl-main.164

[21] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan. 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.

[22] Philip N Johnson-Laird. 1999. Deductive reasoning. *Annual review of psychology* 50, 1 (1999), 109–135.

[23] Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 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.

[24] Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Yu Lan, and Chao Shen. 2022. CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Data Limitation With Contrastive Learning. *arXiv preprint arXiv:2212.10341* (2022).

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

[26] Florence Martin, Ting Sun, Murat Turk, and Albert D Ritzhaupt. 2021. A meta-analysis on the effects of synchronous online learning on cognitive and affective educational outcomes. *International Review of Research in Open and Distributed Learning* 22, 3 (2021), 205–242.

[27] Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences* 34, 2 (2011), 57–74.

[28] Amy T Peterson, Patrick N Beymer, and Ralph T Putnam. 2018. Synchronous and asynchronous discussions: Effects on cooperation, belonging, and affect. *Online Learning* 22, 4 (2018), 7–25.

[29] Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2022. MAUVE Scores for Generative Models: Theory and Practice. *arXiv preprint arXiv:2212.14578* (2022).

[30] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. An information divergence measure between neural text and human text. *arXiv preprint arXiv:2102.01454* (2021).

[31] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, Bimal Viswanath, Virginia Tech, and LUMS Pakistan. 2023. Deepfake Text Detection: Limitations and Opportunities. *44th IEEE Symposium on Security and Privacy* (2023).

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[33] Ashraf I Shirani, Mohammed HA Tafti, and John F Affisco. 1999. Task and technology fit: a comparison of two technologies for synchronous and asynchronous group communication. *Information & management* 36, 3 (1999), 139–150.

[34] Reuben Tan, Bryan Plummer, and Kate Saenko. 2020. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2081–2106.

[35] Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *SIGKDD Explorations* (2023), vol. 25.

[36] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

[37] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2001–2016.

[38] Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural*

Language Generation. 355–368.

[39] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems* 32 (2019).

[40] Tao Zhang. 2022. Deepfake generation and detection, a survey. *Multimedia Tools and Applications* 81, 5 (2022), 6259–6276.

[41] Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural Deepfake Detection with Factual Structure of Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2461–2470.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009