Robustness of AI-generated Text Detection Under Paraphrasing Attacks

Dustin Hayes

***Abstract*—Recent advances in large language models have enabled the generation of highly convincing human-like text, creating new opportunities but also significant risks—such as phishing, fraud, and misinformation. Despite active research, no reliable method for detecting AI-generated text currently exists. As models improve, distinguishing human and synthetic content becomes increasingly difficult.**

This challenge is exacerbated by the possibility of deliberate evasion. One such method is a paraphrasing attack, in which the output of an LLM is reworded to obscure telltale signals. In this study, we evaluate the robustness of AI-text detectors to such attacks.

We begin with a RoBERTa-based model developed by OpenAI to detect GPT-2 outputs. We establish baseline performance, apply a paraphrasing attack using a T5 model, and measure the degradation in detection accuracy. We then fine-tune the detector on a new dataset of GPT-3-generated scientific abstracts and repeat the attack. Performance is recorded at each step.

I. Introduction

IT IS a point of contention among scholars whether or not reliable detection of AI-generated text is feasible. A group at the University of Maryland released a paper in June 2023 purporting that reliable detection of AI-generated text is impossible, especially when fed through a paraphraser. They offered both theoretical and empirical evidence to the effect (1). More optimistically, a group at IBM claims in a July 2023 paper to have developed a model which is robust to paraphraser attacks by training both a detector and paraphraser in an adversarial setting (2). Other authors argue that a multi-modal approach is required; although it may be impossible to detect AI-generated text based on the content of the text alone, perhaps such a system could serve as one component of a larger detection system which also employs user metadata to make a prediction.

This work will be exploratory in nature. In order to gather information regarding the susceptibility of AI-generated text detectors, we will both utilize an existing GPT-2 detector model and an updated model trained via transfer learning to detect GPT-3 generated text. The experimental method for each model will be to first observe baseline performance, then apply paraphrasing using a T5 model and re-evaluate performance.

Our results were mixed. The GPT-2 detector model ultimately behaved in a manner which aligned more closely with our expectations. Paraphrasing resulted in a drop in recall of approximately 8%. On the other hand, our updated model, which was trained by ourselves on a set of 28,662 abstracts, half human-generated and half-machine-generated, displayed no such behavior. The effect of paraphrasing on this model was negligible. We suspect that a hidden artifact or some spurious feature of the data that we employed resulted in superficially optimistic results. We will discuss our theories on this point at a later time.

II. Problem Statement

There currently exists a large gap between the sophisticated and believable outputs that GPT-4 and similar models are able to produce and the systems which we currently use to detect content generated by these LLMs. This gap is problematic; as LLMs continue to develop, their potential for misuse only grows.

Although these concerns might seem forward-thinking and speculative by nature, such instances of misuse occur in the present day. There exist unrestricted LLMs, such as FraudGPT and WormGPT, which are marketed on the dark web specifically for illegal and morally bankrupt behavior, such as phishing and social engineering. Bad actors may also use standard LLMs, such as LlaMA2 or GPT-4, to accomplish their goals by carefully evading the preventative measures set in place by these models (3). Although it is by nature difficult to determine the means by which bad actors are designing their attacks, experts suspect that, as of April 2023, increases in novel phishing attacks can be attributed to the utility provided by the advent of LLMs (4). It has been speculated that this technology could also be used for highly efficient, cheap and convincing propaganda and misinformation dissemination (5). Non-withstanding reports of verifiable, current misuse, the potential for effective misuse is well clarified (3); It remains to be seen how this potential will be leveraged in the future.

The capacity to determine whether or not text was generated by an AI system could prove crucial. Such a capacity might allow communications platforms to identify when these highly effective phishing techniques are being used.. The feasibility of a reliable AI-generated text detector is questionable. Specifically, paraphrasing attacks have proven to be difficult to detect (1), (2). In this work we seek to better understand how robust deep learning based detection methods using models like RoBERTa are to these paraphrasing attacks.

III. Technical approach

There are two models which are key to our experimental setup: a RoBERTa model and a T5 model.

RoBERTa (Robustly Optimized BERT-Pretraining Approach), is a transformer model which represents an improvement on the BERT model developed by Google in 2018. It can be used for a range of tasks, including classification, question answering, and named entity recognition. RoBERTa was developed at Facebook AI. Like BERT, the RoBERTa architecture consists of the encoder part of the transformer architecture. We used RoBERTa base, which consists of 110M parameters organized across 12 encoders and 12 attention heads. It is the pre-training process which differentiates RoBERTa and BERT. The dataset that RoBERTa was trained on was larger than that which was used to train BERT by a factor of ten. The authors of RoBERTa neglected to employ the next sentence prediction task which was used to train BERT, stating that this task was not very fruitful. Furthermore, great care was taken by the authors to search for ideal training hyper-parameters. Relative to BERT, RoBERTa utilizes longer training sequences, a higher learning rate, and larger batch sizes. RoBERTa achieved state of the art performance on SQuAD, GLUE and RACE (6)

The T5 model was developed by Google in 2019. It is a model noted for its versatility; T5 treats each task as a text-to-text task, allowing for a ”unified framework” which can be applied to a wide range of NLP tasks. Indeed, the T5 model achieved state-of-the art results on a number of data sets, including SQuAD, GLUE and RACE (7), as was the case with RoBERTa. We will use T5 base, which consists of 220M parameters, organized according to the transformer based architecture, with both the encoder and decoder halves.

The choice to use RoBERTa and T5 respectively was informed by a number of factors. Perhaps most significantly, we were able to find evidence of each model being used for our chosen purpose, and were able to find checkpoints of each model that were fine-tuned on a task similar to our own. OpenAI chose to use RoBERTa for detecting the outputs produced by GPT-2, and we will begin our training process with the result of their efforts.

T5 is a model which is specifically designed for text-to-text tasks, which is well suited to paraphrasing. We were able to find an implementation of the T5 model which has been trained for this purpose on huggingface: https://huggingface.co/Vamsi/T5_Paraphrase_Paws. Our experimentation indicated that this pretrained T5 model accomplishes our paraphrasing task mostly well, although we will find that it is perhaps not aggressive enough for our particular goals. Allow us to present a pair of examples to demonstrate how the T5 paraphraser behaves:

A. T5 usage examples

The T5 model was configured to generate five alternative wording for each input sequence. Note that it usually only introduces light alternation.

1) Input, human generated: ”For each input sentence, five alternative wordings are output”

output: [”For each input sentence five alternative wordings are sent.”, ”For each input sentence, five alternative wordings are output.”, ”Five alternative words are output for each input sentence.”, ”For each input sentence, five alternate wordings are provided.”, ”For each input sentence, five alternative wordings are output.”]

2) Input, GPT-4 generated: ”In a world filled with wonder and mystery, every moment is an opportunity for discovery and growth.”

output: [”In a world filled with wonder and mystery, every moment is an opportunity for discovery and growth.”, ”In a world filled with wonder and mystery, every moment is a moment for discovery and growth.”, ”In a world filled with wonder and mystery, every moment is an opportunity for discovery and growth.”, ”In a world filled with wonder and mystery, every moment offers an opportunity for discovery and growth.”, ”In a world filled with wonder and mystery, every moment is an opportunity for discovery and growth.”]

IV. Experimental setup

We have divided our efforts into three experiments:

A. Paraphrasing attack on GPT-2 detector.
B. Fine-tuning on GPT-3 generated text.
C. Paraphrasing attack on fine-tuned model.

A. Paraphrasing attack on GPT-2 detector

OpenAI released, concurrently with GPT-2, a RoBERTa based detector which was able to distinguish between GPT-2 output and human-written content pulled from WebText. They likewise released a repository connected to this detector, which contained instructions for downloading both the model checkpoint, and the data used to train the model: https://github.com/openai/gpt-2-output-dataset/tree/master/detector. It also contained many functions and utilities which were useful for our own research. This repository formed the basis for our experimental setup, and, excluding the T5 detector model, contained much of the tools necessary for carrying out our first experiment (8).

Our first experiment is conducted as follows: First, we sample a 500 machine written and 500 human written subset from the data originally used to test the GPT-2 output detector. We will then perform inference on this subset and establish a baseline level of performance. We will subsequently apply our T5 paraphraser to the same test data according to the following scheme. For each machine written example:

1) Collect sentences one at a time, in order and without skipping, until the token limit (512) of the T5 paraphraser would be exceeded by the addition of the next sentence.
2) Generate five paraphrased version of the collected excerpt with the T5 model with the prompt: “paraphrase: $< text >$”
3) Select one of the five paraphrased versions according to a discrete, uniform distribution.
4) Insert the paraphrased text back to its original place in the test set.
5) Continue until each phrase in the machine-generated test set has been passed through the paraphraser once.

This scheme was the result of trial and error. Alternate schemes were attempted by changing the frequency with which paraphrasing occurred within our test set, modifying the prompt given to the T5 model with more specific instructions, and changing hyper-parameters such as top_k, which sets the number of next words considered during the generation process. The scheme presented here represents the most aggressive paraphrasing we accomplished under our time and computational constraints that consistently maintained the meaning of the original text. Likewise, this scheme, unlike some others which we attempted, did not, as far as we are aware, result in the creation of unwanted artifacts which might affect our results.

After our paraphrased test set is generated, we perform inference using the GPT-2 output detector model once more. Any drop in performance metrics relative to our baseline run may be attributed to the paraphrasing attack, the magnitude of the drop will, to some extent, allow us to quantify this vulnerability.

The following is a sample of the human generated text and machine-generated text used for this experiment. Note that the GPT-2 text, although syntactically solid, is somewhat odd in terms of content and overall meaning. This oddness was a consistent feature of our GPT-2 output data. Some special characters have been removed, and quotation marks have been modified to suit LaTeX formatting, otherwise these excerpts are identical to that found in the test data.

1) Human generated; GPT-2 detector set: *In these days of anxiety and alienation, Thanksgiving offers the warm embrace of inclusiveness. Particularly for many people with families and faiths rooted in other lands, no other holiday, not even the Fourth of July, has so great a capacity to make them feel American. A child of Orthodox Jewish immigrants could feel his apartness on other festivals celebrated by the larger society. Christmas, Easter, Halloween all are distinctly Christian observances, no matter how temporal and commercialized they have become. They are inevitable reminders for some Americans that they are different. Thanksgiving's origins are also Christian. But it has evolved into something both secular and spiritual, a day devoted to family and amity. Perhaps that explains its unwavering appeal for believers and nonbelievers alike (even if many Native Americans understandably choose not to partake). Thanksgiving is at heart more than parades, or football or even country; there's no flag-waving or chest-thumping. It is about shared bounty and shared humanity. That's why the writer Saadia Faruqi, a Pakistan-born Muslim, welcomes the day. "For a Pakistani-American, Thanksgiving is as wholesome and normal a holiday as one can get," she said in a 2015 essay. "It is a time to be grateful, to spend time with family, and to have a little bit of fun." Though she never developed a taste for turkey, Ms. Faruqi wrote, the Thanksgiving table would most definitely be set with tandoori chicken, daal and naan.*

2) machine-generated; GPT-2: *The first time I ever saw "Jurassic World" was more than 12 years ago on the night the movie's director, Colin Trevorrow, took a break from the filmmaking grind of his new "Star Wars: The Force Awakens" trilogy to visit a theater in San Francisco called the Argyle. A couple of months into that trip, when a filmgoer sitting a few rows in front of me politely asked whether he would be interested in a preview of the two-hour runtime movie, I thought it a ridiculous question. I was there to see an arthouse hit, and I didn't take any time to look up which genres the movie fit within. Over the next 12 years, I have seen "Jurassic World" almost a dozen times. I have seen it more times than I can count. For the uninitiated, "Jurassic World" is an off-beat science-fiction action film about a boy (played by Bryce Dallas Howard) who stumbles upon a massive prehistoric dinosaur egg in San Francisco and is given the unique opportunity to bring the dinosaur back to life. It's a film about dinosaurs, boy genius, and humanity's quest to understand the evolution of life on Earth. It is also a science-fiction adaptation of "Jurassic Park," written by Steven Spielberg with a screenplay by Frank Darabont, that takes place during the late 1980s...*

B. Fine-tuning on GPT-3 generated text

Since GPT-2 is considerably less convincing than modern LLMs, we thought it prudent to train a model to distinguish between human written text and text generated by a more recently developed, more capable model. In doing so, we hoped to elucidate the difficulties or lack-thereof in training a model to perform such a task, as well use the resultant model to further study the robustness of such models against paraphrasing attacks, as we will do in our third experiment.

We utilize a data set consisting of 28,662 abstracts regarding the 2019 Coronavirus pandemic, half of which were generated by a human authors, the other half being generated by GPT-3, DaVinci. The machine-generated abstracts were created by offering the title of one of the human-written papers to GPT-3 and prompting it to create an abstract according to the following instructions: "Create an abstract for a scientific journal with a formal tone, academic language, and a background story of the topic in a unique paragraph with the title: $< giventitle >$". This data was obtained from a GitHub repository, created by researchers at the University of Thessaly, Greece in connection to their paper "Detection of Fake Generated Scientific Abstracts".

Initial experimentation revealed an artifact that we were concerned would make training superficially easy: almost all of the machine-generated abstracts began with the phrase "This study", or something very similar. We subsequently processed the data in an attempt to remove this artifact. Two strategies were employed: 1) "This study" was removed if it occurred at the beginning of the abstract, 2) LLaMA 2 was used to paraphrase such that this artifact was avoided while maintaining coherency. Unfortunately, each of these strategies produced artifacts of their own. The simple removal of the words "This study" also leaves a clue for our detector: that machine-generated abstracts seem to be missing a few words at the beginning. The LLaMA2 approach resulted in LLaMA2 occasionally inserting thoughts of its own, such as statements like "Sure, I can paraphrase that for you". Due to time constraints, we chose to combine each approach, using the simple pre-processing step for our training data and the LLaMA2 approach for our validation and test sets, in the hope that doing so would prevent this artifact from affecting our test and validation performance, regardless of whether or not the model learned anything from it.

A sample of the data is included. For this researcher – not a biologist – it is quite hard to tell which abstract was produced by a human and which by GPT-3:

1) machine-generated; Abstract Set: The paper examines the prevalence of influenza C in young children with respiratory infections using retroactive data from the national influenza surveillance system in Germany from 2012 to 2014. Recent evidence suggests that the influenza C virus infection manifesting as a primary onset is becoming increasingly common among very young children, posing a public health risk and necessitating continued research. This study seeks to provide fresh insights into this critical issue by analyzing past information on the occurrence of Infectious Bronchitis Virus (IBV) caused by an Influenza C virus in Germany's collective healthcare system among children with respiratory infections for three years. By utilizing the existing data collected during this period for anthropological analysis, the study aims to shed further light on this important problem and inform future initiatives to address it more effectively.

2) Human generated; Abstract Set: Respiratory RNA viruses are constantly evolving, thus requiring development of additional prophylactic and therapeutic strategies. Harnessing the innate immune system to non-specifically respond to viral infection has the advantage of being able to circumvent viral mutations that render the virus resistant to a particular therapeutic agent. Viruses are recognized by various cellular receptors, including Toll-like receptor (TLR) 3 which recognizes double-stranded (ds)RNA produced during the viral replication cycle. TLR3 agonists include synthetic dsRNA such as poly (IC), poly (ICLC) and poly (AU). These agents have been evaluated and found to be effective against a number of viral agents. One major limitation has been the toxicity associated with administration of these drugs. Significant time and effort have been spent to develop alternatives/modifications that will minimize these adverse effects. This review will focus on the TLR3 agonist, poly (IC)/(ICLC) with respect to its use in treatment/prevention of respiratory viral infections.

Preliminary experimentation suggested that transfer-learning via our previously utilized GPT-2 output detector is quite effective. As such, our experiment consists of further fine-tuning our GPT-2 output detector, with all layers unfrozen, and subsequently observing the performance of our trained model on the test data. Hyper-parameters and training settings such as number of epochs, optimizer and learning rate will be determined on a trial-and-error basis, although, as we will find, minimal experimentation was required in order to achieve high accuracy. Our hope is that carrying out this training may provide some information regarding the relative difficulty of this task, as well as set the stage for our final experiment.

C. Paraphrasing attack on fine-tuned model

Having obtained a model which is capable of distinguishing between human written and GPT-3 written data, or at least human written and machine written abstracts, our final experiment consists of repeating our paraphrasing attack experiment on our new model. Paraphrasing will be accomplished according to the same scheme. We have elected to only paraphrase a subset of our test data once more; in this case we will utilize a sample of 1983 data points. Paraphrasing is a rather expensive operation and, as we will find, this sample is sufficient for determining the effect that our paraphrasing attack has.

V. Results and Discussion

We will likewise treat each experiment separately during this section in order to maintain clarity. Will will also include a discussion subsection at the end to discuss our results in general as they relate to our overarching purpose.

A. Paraphrasing attack on GPT-2 detector: Results

We observed a moderate drop in performance as a result of the paraphrasing attack. Before paraphrasing, our GPT-2 output detector produced the following metrics on our sample of 1000 examples, taking ”machine-generated” to be our positive class:

- Accuracy: 0.958
- Precision: 0.956
- Recall: 0.960

	Predicted	
	Positive	Negative
Actual Positive	480	20
Actual Negative	22	478

TABLE I
Confusion Matrix, GPT-2 Output Detector, Pre-paraphrasing

After perturbing our data according to the scheme described in our section regarding experimental setup, we observed a degradation in the capacity of the GPT-2 output detector to detect our paraphrased AI-generated text. The recall of our detector dropped by approximately 8%; 42 additional true positive examples were misclassified.

- Accuracy: 0.916
- Precision: 0.952
- Recall: 0.876

	Predicted	
	Positive	Negative
Actual Positive	438	62
Actual Negative	22	478

TABLE II
Confusion Matrix, GPT-2 Output Detector, Post-paraphrasing

B. Fine-tuning on GPT-3 generated text: Results

Our training was largely successful, although concerns persist regarding the accidental inclusion of some artifact besides that mentioned prior made this task superficially easy.

Although not included here due to the apparently inferiority of this approach, we did initially attempt to train beginning with RoBERTa base. Transfer learning, beginning with the GPT-2 output detector model, proved very effective by comparison. All 28,662 abstracts were used in training. We employed a 60/20/20 training/validation/test split. We were prepared to perform a rigorous hyper-parameter search, but early runs provided highly accurate results. Our most successful run employed the following settings:

- Epochs: 2
- Learning Rate: 2e-5
- Batch Size: 8
- Loss: Cross Entropy
- Optimizer: Adam
- Tokenizer: RoBERTa Base
- Max Sequence Length = 512

We subsequently observed the following performance on the test set

- Accuracy: 0.989
- Precision: 0.998
- Recall: 0.980

	Predicted	
	Positive	Negative
Actual Positive	2809	57
Actual Negative	4	2862

TABLE III
CONFUSION MATRIX, UPDATED ROBERTA, PRE-PARAPHRASING

These metrics are very high, although the paper from which our abstract data was sourced reported similarly high performance using a BERT based approach (9). Nonetheless, we remain skeptical of our results and hope to ensure that this performance is genuine and suitable.

C. Paraphrasing attack on fine-tuned model: Results

We found that our paraphrasing attack was ineffective in reducing the accuracy of our updated detector on our paraphrased data. Our observed performance on our paraphrased data set is very similar to that of our un-paraphrased data set. We will offer some theories as to why we think this might be at a later point.

Our performance metrics on the paraphrased subset of our abstract data was as follows:

- Accuracy: 0.988
- Precision: 0.978
- Recall: 0.998

	Predicted	
	Positive	Negative
Actual Positive	931	21
Actual Negative	2	1029

TABLE IV
CONFUSION MATRIX, GPT-2 OUTPUT DETECTOR, POST-PARAPHRASING

D. Discussion

Here we present a discussion regarding each of our experiments and our findings as a whole. Note that we do offer theories as to why we observed the results that we did, but many of these theories are speculative and further research would be required to validate them.

Our experiment regarding the base GPT-2 output detector was perhaps most in line with our initial expectations. Although the drop in performance was somewhat modest, we did observe a clear reaction to our paraphrasing attack. Nonetheless, it seems that our intervention was not aggressive enough to "break" our detector, as has been observed in existing literature (1), at least in our estimation. We speculate that perhaps it was, in part, the nonsensical nature of GPT-2 output which causes our paraphraser to be less effective than expected; GPT-2 writes in a manner which is notably less logically coherent than most human writers. If this pattern was detectable by our RoBERTa model, there is no amount of paraphrasing which retains the original meaning of the text which will remove it.

We were concerned that an artifact in the training data may have driven the model's high accuracy. The realization that the paper in which this data was first used reported a similar level of accuracy (.987) with a BERT based approach assuaged our fears to a certain degree (9). It was also notable how effective transfer learning seemed to be. We only required two epochs to reach a high level of performance, and much of the learning had, in fact, already been accomplished towards the beginning of our first epoch. However, this task is certainly much easier than achieving general AI text detection capabilities. Firstly, this data is very narrow in scope. Only abstracts written on topics related to the Coronavirus pandemic were available. Secondly, all machine-generated texts were made using the same prompt, with the title of the study that GPT-3 was to create an abstract for being the only variable. Thirdly, this data was produced using only GPT-3. As such, this task must be considerably easier than identifying AI-generated text in general.

Our third experiment was perhaps the most surprising, and requires the most in terms of continued research. Although we were prepared for our paraphrasing approach to have only a modest effect on our performance, as we observed with our first experiment, we did not observe any notable drop at all. It was at this juncture that we once more attempted to modify our approach towards aggressively so as to observe our expected drop, which we thought appropriate due to the nature of the problem. A motivated bad actor could be prepared and capable of designing a more thorough and disruptive scheme, and we are required to do the same to verify how effective such an attack might be. Among our attempted strategies, we tried changing our prompt from "paraphrase: ¡text¿" to "paraphrase fully, retain meaning: ¡text¿". Unfortunately, the model responded by inserting our directive into odd places in its output. Further study, additional processing steps, or some other strategy is required to more accurately simulate a true attack.

1) Aggressive paraphrase attempt: A meta-analysis of in vitro and in vivo studies examines the biological effects of low-level millimeter waves (MMWs) fully paraphrased, but retain meaning : The existing research on this topic is limited and disconnected, making it difficult to determine the actual impacts that MMWs may have on human physiology. To address this knowledge gap we conducted an extensive search of five electronic databases (Pubmed/Medline, Web of Science, Embase etc.) published before 2020 for relevant studies assessing MMW exposure among various animal and cellular models. After a comprehensive review based on predetermined criteria, 25 articles were selected for detailed analysis of tissue responses to different doses and parameters associated with MMW - radiation. Our results suggest possible modest effects on cells or animals after relatively long exposures at high power densities and specific frequencies previously not studied ; however, further research needs to support these findings using more scientific protocols. Given the lack of sufficient knowledge concerning the health risks posed by contemporary exposure to MMWs in everyday scenarios such as those emitted by 4G telecommunication systems - there is an urgent need for more precise experiments to test such long-term consequences that this form of nonionising radiation presents.

VI. Future Work

We believe that expanding our work and producing a better representation of a paraphrasing attack for study involves: gathering a more diverse data set, experimenting with an alternative paraphraser, or ideally, training a paraphraser not just to paraphrase, but to paraphrase with the intention of avoiding detection. Once such a paraphraser is obtained, the task of creating a detector which can manage in the presence of such a disruption can begin. There exists work which manages to accomplish both of these tasks simultaneously: “RADAR: Robust AI-Text Detection via Adversarial Learning” contributed by Xiaomeng Hu, Pin-Yu Chen and Tsung-Yi Ho (2023) (2). They have designed a means by which a detector and paraphraser are trained simultaneously in an adversarial setting. Their results indicate that a detector trained in this manner displays a much greater degree of robustness to these attacks than extant statistics and machine learning based methods. Perhaps we could pursue something similar as a continuation of this work.

In summary, due to the rapidly expanding capacities of LLMs, and the many concerning applications of this technology, it would be highly advantageous to develop a means to detect AI-generated text. This is a difficult problem to solve, and it remains to be seen if a robust, durable solution can be developed and effectively implemented. We conducted a series of introductory experiments designed to explore the susceptibility of these detection models, to varying degrees of success. Further study is required to approach this important problem with care and effectiveness.

References

[1] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can ai-generated text be reliably detected?” arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2303.11156

[2] X. Hu, P.-Y. Chen, and T.-Y. Ho, “Radar: Robust ai-text detection via adversarial learning,” arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2307.03838

[3] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh, “From chatbots to phishbots? – preventing phishing scams created using chatgpt, google bard and claude,” 2023.

[4] J. Hazell, “Large language models can be used to effectively scale spear phishing campaigns,” 2023.

[5] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Y. Wang, “On the risk of misinformation pollution with large language models,” 2023.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” 2023.

[8] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, “Release strategies and the social impacts of language models,” 2019.

[9] P. C. Theocharopoulos, P. Anagnostou, A. Tsoukala, S. V. Georgakopoulos, S. K. Tasoulis, and V. P. Plagianakos, “Detection of fake generated scientific abstracts,” in *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, 2023, pp. 33–39.