Catch Me If You GPT: Tutorial on Deepfake Texts

Adaku Uchendu[†] **Saranya Venkatraman** **Thai Le**[♠] **Dongwon Lee**

MIT Lincoln Laboratory, Lexington, MA, USA[†]
The Pennsylvania State University, University Park, PA, USA
Indiana University, Bloomington, IN, USA[♠]

adaku.uchendu@ll.mit.edu[†], {saranyav, dongwon}@psu.edu, leqthai.vn@gmail.com[♠]

Abstract

In recent years, Natural Language Generation (NLG) techniques have greatly advanced, especially in the realm of Large Language Models (LLMs). With respect to the quality of generated texts, it is no longer trivial to tell the difference between human-written and LLM-generated texts (i.e., deepfake texts). While this is a celebratory feat for NLG, it poses new security risks (e.g., the generation of misinformation). To combat this novel challenge, researchers have developed diverse techniques to detect deepfake texts. While this niche field of *deepfake text detection* is growing, the field of NLG is growing at a much faster rate, thus making it difficult to understand the complex interplay between state-of-the-art NLG methods and the detectability of their generated texts. To understand such inter-play, two new computational problems emerge: (1) *Deepfake Text Attribution* (DTA) and (2) *Deepfake Text Obfuscation* (DTO) problems, where the DTA problem is concerned with attributing the authorship of a given text to one of k NLG methods, while the DTO problem is to evade the authorship of a given text by modifying parts of the text. In this **cutting-edge tutorial**, therefore, we call attention to the serious security risk both emerging problems pose and give a comprehensive review of recent literature on the detection and obfuscation of deepfake text authorships. Our tutorial will be 3 hours long with *a mix of lecture and hands-on examples* for interactive audience participation. You can find our tutorial materials here: `https://tinyurl.com/naacl24-tutorial`.

1 Introduction

Since the advent of the Transformer network architecture (Vaswani et al., 2017) in 2018, the field of NLG has exponentially expanded. This architectural design led to the development of GPT-1 (Radford et al., 2018), the first installment in deepfake text generative models that are capable of generating long-coherent texts. Since then there have been several (i.e., GPT-4 (OpenAI, 2023), Flan-T5 (Chung et al., 2022), LlaMA (Meta, 2023), etc). In fact, with each new installment in the world of long-coherent text generation, these generated texts become more and more human-like. Such LLM-generated texts are referred to as **deepfake texts**. While this is a great feat for the field of NLG and has several impactful applications, such text generators pose a security risk. This security risk is the potential inability to distinguish human-written texts from deepfake texts, which allows for malicious users of such NLG models to generate misinformation (Zellers et al., 2019; Uchendu et al., 2020), and propaganda (Varol et al., 2017).

Therefore, we have 2 problems to tackle - (1) distinguish deepfake texts from human-written texts, and (2) detect obfuscated (i.e., a technique to evade detection) deepfake texts. While several researchers are working on these two problems, a few issues with deepfake text generation have been highlighted by other researchers. These issues or limitations include: (1) memorization & plagiarizing of training set (Carlini et al., 2021; Duskin et al., 2021; Lee et al., 2022), (2) generation of toxic & harmful speech (Pavlopoulos et al., 2020; Venkit et al., 2023; Deshpande et al., 2023), (3) generation of hallucinated text (Zhou et al., 2021; Ji et al., 2023), (4) generation of misinformation (Jawahar et al., 2020; Pan et al., 2023; Shevlane et al., 2023), etc.

Such limitations of deepfake text generators, further confirm the need to reliably distinguish human-written and deepfake texts. Thus, in this tutorial, we explore the following: (1) Deepfake Text Attribution (DTA) which involves correctly attributing the authorship of a given text to one of k NLG methods, and (2) Deepfake Text Obfuscation (DTO) that focus on evading the authorship of a given text by modifying parts of the text.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), pages 1–7

2 Target Audience & Prerequisites

The target audience includes graduate students, practitioners, and researchers attending the NAACL conference, coming from different areas of the Machine Learning (ML)/Natural Language Processing (NLP)/Computational Linguistics (CL) field. Basic common knowledge in NLP and ML would be helpful but not required. We plan to make the tutorial as self-contained as possible for a wider audience. We expect about 50-70 participants to attend our tutorial. Lastly, we believe that this tutorial will be most suited to the NAACL 2024 conference.

3 Tutorial Type

Hence, we propose a **cutting-edge** tutorial with hands-on examples that will present the current research on *deepfake text detection*. Our tutorial will be mainly **a mix of lecture and hands-on style**. It will include examples of the generation, detection, and obfuscation of deepfake texts for interactive participation from the audience.

4 Tutorial Outline

The materials of our tutorial will mainly contain lecture-based slideshows of this **cutting-edge** niche field. Although we are delivering the tutorial in lecture style, we also include a few quick interactive activities to showcase real-life examples of deepfake texts and their implications. They will be polls, binary/multiple-choice, and group-based questions.

4.1 Introduction and Background (30 minutes)

This section will introduce the topic of NLG and the many improvements it has seen since the incorporation of the Transformer network into Language models. After this introduction, we will briefly discuss deepfake text generation. Next, we motivate the many benefits of deepfake texts as well as the risks they could pose. This will allow us to transition to briefly introducing the main problem - *deepfake text attribution and obfuscation*. Finally, we provide an outline of the tutorial which will include topics that would be covered and not covered in the tutorial.

Then, we will focus on the evolution of deepfake text generation and detection. We will also briefly introduce the history of Deepfake Text Attribution (DTA) and Deepfake Text Obfuscation (DTO). Particularly, we will discuss the different terminologies used to describe deepfake texts (such as artificial texts, synthetic texts, machine-generated texts, etc.). As this is still a relatively new field, there is still no agreed-upon universal term for deepfake text generation. We will also briefly highlight why we use the term *deepfake texts generation*, instead of the other terms.

4.2 Deepfake Text Attribution (40 minutes)

This section will present the following sub-topics:

1. **Interactive Activity.** We start this session by inviting the audience to join in a hands-on activity. The attendees will be asked to detect some examples of human-written v.s. deepfake texts. We will prepare paper handouts for the audience for this activity, which will include all the needed descriptions. In the hybrid setting, we will show the material through the provided video call system (e.g., Zoom). Through this activity, we want the attendees to grasp the difficulty of detecting deepfake texts, due to the challenges of distinguishing them from human-written ones.

2. **Datasets.** We will introduce several relevant publicly available English and multilingual datasets across different domains such as TuringBench dataset (Uchendu et al., 2021), M4 (Wang et al., 2023), Med-MMHL (Sun et al., 2023), DeepfakeTextDetect (Li et al., 2023), etc.

3. **Computational Approaches.** We will present the different ways in which researchers have tackled the problem of deepfake text detection. We will also discuss the limitations of the current computational approaches and potential ways the ML/NLP/CL communities could mitigate or solve such limitations. Some of the current SOTA automated DTA approaches include GPT-2 Output Detector[1], DetectGPT (Mitchell et al., 2023), GPTZero[2], etc.

4. **Human Approaches.** We will present and discuss the several ways in which researchers have attempted to improve human detection

[1] https://huggingface.co/spaces/openai/openai-detector

[2] https://gptzero.me/

(Clark et al., 2021; Ippolito et al., 2020; Dugan et al., 2020; Pillutla et al., 2021; Gehrmann et al., 2019; Dou et al., 2021; Uchendu et al., 2023b; Perkins et al., 2023) of deepfake texts.

4.3 Watermarking LLMs (25 minutes)

In addition, we will discuss several watermarking techniques (Kirchenbauer et al., 2023; Yoo et al., 2023; Zhao et al., 2023), another computational approach to mitigating the potential negative effects of deepfake text generation. Watermarking essentially embeds a hidden pattern into a text such that the pattern enables its detection by deepfake text detectors while being imperceptible to the human eye. This has implications for inhibiting misuse, misattribution and Intellectual Property (IP) infringement of deepfake texts, and is a growing and increasingly crucial line of work for the safe and large-scale deployment of LLMs in real-world settings.

4.4 QUESTIONS (10 minutes)

4.5 BREAK (30 minutes)

4.6 Obfuscation of Deepfake Texts (40 minutes)

This section will present the following sub-topics:

1. **Deepfake Text Obfuscation Techniques.** We first introduce the definitions of DTO task and how it is different from adversarial attacks. Then, we will briefly describe some of the current SOTA DTO algorithms (e.g., (Mahmood et al., 2019; Haroon et al., 2021)) and also some relevant adversarial attack techniques on text. Then, we discuss in detail all the research that has been done in this area to highlight the lack of adversarial robustness of SOTA DTA models for deepfake texts detection (Jun et al., 2022; Crothers et al., 2022; Gagiano et al., 2021; Wolff and Wolff, 2020). Next, we discuss the gaps in the literature, the future direction of problems in this domain, and the ways in which the ML, NLP and CL community could contribute and improve upon the current landscape.

2. **Interactive Game.** We will demonstrate a demo for adversarially perturbing the deepfake texts in real-time to mislead the deepfake texts DTA detectors to misclassify. For this demonstration, we will utilize the ChatGPT Detectors - GPTZero, and ZeroGPT.

4.7 Applications and Implications (15 minutes)

We will use this session to encourage the audience to ponder how deepfake texts will influence their sub-discipline community. In particular, we will discuss how improvements in *DTA and DTO tasks* could be applied to similar problems like *fake news detection, hallucinated text detection, chatbot detection, hate speech detection*, etc. We will also briefly discuss conversational AI models, such as ChatGPT under the context of the tutorial (e.g., distinguishing between human and automated conversational agents via DTA). Finally, we will then focus our talk on discussing one to two specific scenarios where deepfake texts can be utilized for both good and malicious purposes. We will encourage the audience to engage in the discussion via live polling.

4.8 Future Direction (10 minutes)

In this section, we will present and discuss the future directions in this field and potential ways the ML, NLP, and CL communities can both benefit and assist. These directions include the building of (1) *explainable & Intuitive DTA models for deepfake text detection*, (2) *robust style-based classifier*, and (3) *deepfake text obfuscation for* $k > 2$ *authors*.

4.9 QUESTIONS (15 minutes)

5 Reading List

The references included in this tutorial proposal are relevant references to help the audience get more acquainted with the topic. Also, this NAACL 2024 tutorial will be largely drawn from the authors' recent survey paper (Uchendu et al., 2023a).

6 Tutors

Presenters of this tutorial include a diverse group of researchers. See below for their brief biographies.

- **Adaku Uchendu**[3] is a Technical Staff member (AI researcher) at MIT Lincoln Lab. She recently earned her Ph.D. in Information Sciences and Technology from Penn State University. She was a Sloan scholarship fellow, an NSF CyberCorps SFS scholar, and a Button-Waller fellow. Her dissertation is titled *Reverse Turing Test in the Age of Deepfake Texts*. She has authored several papers

[3]Main Contact

in deepfake text detection at top-tier conferences & journals - EMNLP, KDD Exploration, Web Conference, AAAI HCOMP, NAACL, etc. In addition, she led two similar Tutorials titled, *Tutorial on Artificial Text Detection* (Uchendu et al., 2022) at the INLG conference in July 2022 and *Catch Me If You GAN: Generation, Detection, and Obfuscation of Deepfake Texts* (Fionda et al., 2023) at the Web conference in April 2023. Also, she will give a similar tutorial (with the same title as this proposal) at the 2023 NSF Cybersecurity Summit in October 2023. More details of her research can be found at: https://adauchendu.github.io/.
E-mail: adaku.uchendu@ll.mit.edu

- **Saranya Venkatraman** is a Ph.D. student at Penn State University, working under the guidance of Dr. Dongwon Lee in the College of Information Sciences and Technology. Her research focuses on using psycholinguistics theories and theories of human cognition to inform natural language processing techniques, with a focus on deepfake text detection and deepfake text obfuscation. She also contributed to and presented a *Tutorial on Artificial Text Detection* (Uchendu et al., 2022) at the INLG conference, in July 2022 and has published in top-tier conferences like AAAI, EMNLP, EACL, NAACL, and CHI. More details of her research can be found at: https://saranya-venkatraman.github.io/.
E-mail: saranyav@psu.edu

- **Thai Le** is joining the Department of Computer Science at Indiana University as an Assistant Professor. He has been an Assistant Professor at the University of Mississippi, worked at Amazon Alexa, and obtained his doctorate from The Pennsylvania State University. He has published several relevant works at top-tier conferences such as KDD, ICDM, ACL, EMNLP, and Web Conference. He is also one of the Instructors in a similar Tutorial presented at the Web conference in April 2023 and the 2023 NSF Cybersecurity Summit in October 2023. In general, he researches the trustworthiness of machine learning and AI, with a focus on explainability and adversarial robustness of machine learning models. More details of his research can be found at: https://lethaiq.github.io/tql3.
E-mail: leqthai.vn@gmail.com

- **Dongwon Lee** is a Full Professor in the College of Information Sciences and Technology (a.k.a. iSchool) at Penn State University, and also an ACM Distinguished Scientist and Fulbright Cyber Security Scholar. Before starting at Penn State, he worked at AT&T Bell Labs and obtained his Ph.D. in Computer Science from UCLA. From 2015 to 2017, he also served as a Program Director at National Science Foundation (NSF), co-managing cybersecurity education and research programs and contributing to the development of national research priorities. In general, he researches problems in the areas of data science, machine learning, and cybersecurity. Since 2017, in particular, he has led the SysFake project at Penn State, investigating computational and socio-technical solutions to better combat fake news. More details of his research can be found at: http://pike.psu.edu. Previously, he has given nine tutorials at various venues, including WWW, AAAI, CIKM, SDM, ICDE, and WebSci.
E-mail: dongwon@psu.edu

7 Previous Tutorials

Adaku, Thai, and Dongwon presented a similar tutorial at the *ACM Web conference* in April 2023, titled "Catch Me If You GAN: Generation, Detection, and Obfuscation of Deepfake Texts"[4]. Furthermore, the tutors led the same tutorial at the *2023 NSF Cybersecurity Summit* in October 2023. However, due to the growing interest in *deepfake text detection*, and the emerging strategies to ascertain the authorship of deepfake texts, we introduce another tutor, Saranya Venkatraman for the NAACL Tutorial to include latest developments, such as watermarking strategies of deepfake texts both in theory and practical applications.

8 Ethics Statement

While we highlight the potential negative applications of LLMs to motivate the creation of solutions to mitigate their effects, we understand that malevolent actors could use such knowledge maliciously.

[4] https://adauchendu.github.io/Tutorials/

However, since the focus of our tutorial is on mitigation, we believe that the benefits of this tutorial outweigh the risks. Additionally, our tutorial will also include strategies, like watermarking that can be used by creators of LLMs to further mitigate the potential negative exploitation of LLMs.

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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)*, pages 7282–7296.

Evan Crothers, Nathalie Japkowicz, Herna Viktor, and Paula Branco. 2022. Adversarial robustness of neural-statistical features in detection of generative transformers. *arXiv preprint arXiv:2203.07983*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

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*.

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*, pages 189–196.

Kayla Duskin, Shivam Sharma, Ji Young Yun, Emily Saldanha, and Dustin Arendt. 2021. Evaluating and explaining natural language generation with genx. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 70–78.

Valeria Fionda, Olaf Hartig, Reyhaneh Abdolazimi, Sihem Amer-Yahia, Hongzhi Chen, Xiao Chen, Peng Cui, Jeffrey Dalton, Xin Luna Dong, Lisette Espin-Noboa, et al. 2023. Tutorials at the web conference 2023. In *Companion Proceedings of the ACM Web Conference 2023*, pages 648–658.

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Jenny Zhang, and Jennifer Biggs. 2021. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 119–127.

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*, pages 111–116.

Muhammad Haroon, Fareed Zaffar, Padmini Srinivasan, and Zubair Shafiq. 2021. Avengers ensemble! improving transferability of authorship obfuscation. *arXiv preprint arXiv:2109.07028*.

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*, pages 1808–1822.

Ganesh Jawahar, Muhammad Abdul-Mageed, and VS Laks Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Evan Lim Hong Jun, Chong Wen Haw, and Chieu Hai Leong. 2022. Robustness analysis of neural text detectors.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2022. Do language models plagiarize? *arXiv preprint arXiv:2203.07618*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild. *arXiv preprint arXiv:2305.13242*.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proc. Priv. Enhancing Technol.*, 2019(4):54–71.

AI Meta. 2023. Introducing llama: A foundational, 65-billion-parameter large language model. *Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

OpenAI. 2023. Gpt-4 technical report.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Mike Perkins, Jasper Roe, Darius Postma, James McGaughran, and Don Hickerson. 2023. Game of tones: Faculty detection of gpt-4 generated content in university assessments. *arXiv preprint arXiv:2305.18081*.

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*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. 2023. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv preprint arXiv:2306.08871*.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023a. Attribution and obfuscation of neural text authorship: A data mining perspective. *SIGKDD Explorations*, page vol. 25.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Ting-Hao 'Kenneth' Huang, and Dongwon Lee. 2023b. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? *The Eleventh AAAI Conference on Human Computation and Crowdsourcing*.

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*, pages 2001–2016.

Adaku Uchendu, Vladislav Mikhailov, Jooyoung Lee, Saranya Venkatraman, Tatiana Shavrina, and Ekaterina Artemova. 2022. Tutorial on artificial text detection. *15th International Conference on Natural Language Generation (INLG): Tutorial*.

Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 280–289.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Max Wolff and Stuart Wolff. 2020. Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust multi-bit natural language watermarking through invariant features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2115.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. *arXiv preprint arXiv:2302.03162*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP Findings)*, Virtual.