Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites

Hans W. A. Hanley and Zakir Durumeric

Stanford University
hhanley@cs.stanford.edu, zakir@cs.stanford.edu

Abstract

As large language models (LLMs) like ChatGPT have gained traction, an increasing number of news websites have begun utilizing them to generate articles. However, not only can these language models produce factually inaccurate articles on reputable websites but disreputable news sites can utilize LLMs to mass produce misinformation. To begin to understand this phenomenon, we present one of the first large-scale studies of the prevalence of synthetic articles within online news media. To do this, we train a DeBERTa-based synthetic news detector and classify over 15.46 million articles from 3,074 misinformation and mainstream news websites. We find that between January 1, 2022, and May 1, 2023, the relative number of synthetic news articles increased by 57.3% on mainstream websites while increasing by 474% on misinformation sites. We find that this increase is largely driven by smaller less popular websites. Analyzing the impact of the release of ChatGPT using an interrupted-time-series, we show that while its release resulted in a marked increase in synthetic articles on small sites as well as misinformation news websites, there was not a corresponding increase on large mainstream news websites.

1 Introduction

Since the release of ChatGPT in November 2022, hundreds of millions of Internet users have used the large language model (LLM) to efficiently compose letters, write essays, and ask for advice (Hu 2023). However, LLMs have also been shown to produce factually erroneous text. In one example, CNET, a reputable website that reviews of consumer electronics, published articles generated by OpenAI's ChatGPT that were rife with factual errors (Leffer 2023). Beyond inaccurate text, recent research has shown LLMs can be used to effectively spread misinformation (Tang, Chuang, and Hu 2023). Yet, despite the widespread adoption of LLMs and their potential to accelerate the spread of misinformation, there has not been any study of whether LLMs like ChatGPT *have been* broadly used to produce news articles on mainstream or fringe/unreliable websites.

In this work, we present a large-scale study of the *relative increase* in *machine-generated/synthetic* articles from 3,074 news websites (1,059 misinformation/unreliable websites and 2,015 mainstream/reliable news websites) between January 1, 2022, and May 1, 2023. To do this, we utilize training data from 19 open-source LLMs, as well as adversarial data from article perturbation/re-writes and paraphrases, to train a DeBERTa-based model (He et al. 2021) to detect English-language synthetic news articles. We subsequently benchmark this classifier on eight test sets of machine-generated news articles, including two from real-world companies (Pu et al. 2023) and one containing human-written real-world articles. Across these test datasets, our model, at a false positive rate (FPR) of 1%, achieves an average precision score of 0.992. With this model, we classify over 15.46M articles published between January 1, 2022, and May 1, 2023, from our set of 3,074 news websites.[1]

We find that among reliable/mainstream news websites, synthetic articles increased in prevalence by 57.3% (0.88% of news articles in January 2022 to 1.39% in May 2023) while among unreliable/misinformation websites, the prevalence increased by 474% (0.39% of news articles in January 2022 to 2.22% in May 2023). Examining the content of synthetic articles, we find that while mainstream/reliable news websites have largely utilized synthetic articles to report on financial and business-related news, misinformation/unreliable news websites have reported on a wide range of topics ranging from world affairs (*e.g.*, the Russo-Ukrainian War) to human health (*e.g.*, COVID-19). Examining the impact of ChatGPT on the prevalence of synthetic content, we further find that its release coincided with significant increases in machine-generated articles on misinformation websites and unpopular mainstream news websites.

Our work presents one of the first in-depth analyses of the growth of synthetic articles across the news ecosystem. We show that throughout 2022 and 2023, particularly after the release of ChatGPT, many misinformation websites have rapidly increased the amount of synthetic content on their websites. As misinformation websites increasingly utilize synthetic articles, we hope that our work can serve as the basis for identifying the use of LLMs and for helping enable future studies on the spread of misinformation.

[1] We release the weights of our model and the URLs used in this study at https://github.com/hanshanley/machine-made-media.

2 Background and Related Work

Recent advances in large language models (LLMs) have resulted in impressive performance on a variety of tasks, most notably convincing text generation (Brown et al. 2020; Chowdhery et al. 2023; AI 2022; Zellers et al. 2019). Since 2022, models such as Open AI's ChatGPT, Meta's LLaMa, and Google's Gemini have largely democratized LLMs' use. However, despite their popularity, the widespread availability of LLMs can be problematic. For example, Zeller et al. (2019) showed that even the older GPT-2 LLM can create convincing articles, often with factual errors, that evoke more trust than human-written articles.

Definition: Synthetic Articles Within this work, we consider news articles largely generated by LLMs and other automated software to be *synthetic/machine-generated* (Gagiano et al. 2021). For instance, an article produced by directly prompting the API for OpenAI's `GPT-3.5 davinici` LLM would be considered *synthetic*. We note, however, as shown in prior work (Mitchell et al. 2023; Uchendu et al. 2021), heavily human-edited machine-generates news articles are difficult to detect, often being indistinguishable from human-written news articles. As such, within this work, we further define *synthetic* news articles as those that are largely if not completely generated by LLMs without significant human modification.

Real-World Use of Synthetic News Media. While the large-scale democratization of generative models is new, the use of machine-generated or *synthetic* articles by news websites is not. Since as early as 2019, Bloomberg has used the service Cyborg to automate the creation of nearly one-third of their articles (Peiser 2019). Similarly, since 2019, other reputable news sources including The Associated Press, The Washington Post, and The Los Angeles Times, have used machine-generation services to write articles on topics that range from minor league baseball to earthquakes (Peiser 2019). However, articles that contain machine-generated content from services such as Cyborg, BERTie, or ChatGPT, while reducing the workload of reporters, have also been shown to often contain factual errors (Alba 2023; Leffer 2023). As a result, much research has focused on detecting machine-generated news articles (Zellers et al. 2019; Uchendu et al. 2020; He et al. 2023; Ippolito et al. 2020).

Detecting Machine-Generated Media. Several approaches have been developed to detect machine-generated text. BERT-defense (Ippolito et al. 2020) for instance uses a BERT-based (Devlin et al. 2019) model to identify machine-generated texts. DetectGPT (Mitchell et al. 2023) approximates the probabilistic curvature of specific LLMs for zero-shot detection. Mitchell et al. show that if the specific model used to generate text is known and can be readily queried to obtain the log probabilities of pieces of text, then it is possible to easily differentiate synthetic articles from human-written news articles, with their approach achieving a 0.97 AUROC for the XSum dataset. Zhong et al. (2020) propose a graph-based approach that considers the factual structure of articles to detect machine-generated text.

Our work depends on accurately identifying machine-generated articles across news websites. As shown in previous works, however, many machine learning models trained to detect synthetic texts overfit to their training domain, the token distribution of the model used to generate the synthetic texts, and the topics that they were trained on (Mitchell et al. 2023; Uchendu et al. 2020; Lin, Hilton, and Evans 2022). For example, models trained to detect synthetic news articles, often fail to detect shorter machine-generated tweets. Despite these shortcomings, as illustrated by Pu et al. (2023), classifiers focused on only one domain can often perform exceedingly well on datasets seen "in-the-wild." Adversarially training a RoBERTa (Liu et al. 2019) based classifier, Pu et al. achieve an F_1-classification score of 87.4–91.4 on a test dataset made up of synthetic news articles purchased from AI Forger and Article Forge. Unlike in other domains, such as tweets or comments, news articles tend to be longer, allowing for greater precision in their classification (Pu et al. 2023; Sadasivan et al. 2023).

Reliable and Unreliable News Websites. In this work, we analyze how both reliable/mainstream and unreliable/misinformation news websites have published machine-generated articles throughout 2022 and 2023. Unreliable information can take the form of *misinformation*, *disinformation*, and *propaganda*, among other types (Jack 2017). Within this work, we refer to websites that have been labeled by other researchers as generally spreading *false* or unreliable information as misinformation/unreliable news websites (including both websites labeled as *misinformation* and *disinformation* within this label). As in prior work, we consider reliable/mainstream news websites as "outlets that generally adhere to journalistic norms including attributing authors and correcting errors; altogether publishing mostly true information" (Hounsel et al. 2020).

3 Detecting Machine-Generated Articles

As described in Section 2, several approaches have been developed for identifying synthetic articles, with some of the most successful being transformer-based methodologies (Pu et al. 2023; Gehrmann, Strobelt, and Rush 2019). However, given that past models were trained to (1) only detect text from *particular* models (Zellers et al. 2019), (2) are deeply vulnerable to adversarial attacks (Pu et al. 2023), (3) or have unreleased weights (Zhong et al. 2020), we design and benchmark our *own* transformer-based machine-learning classifiers to identify synthetic articles in the wild.

In addition to training three transformer architectures (BERT, RoBERTa, DeBERTa) on a baseline training dataset (detailed below), we further train these models on datasets generated by two common adversarial attacks (Krishna et al. 2024; Mitchell et al. 2023). To benchmark and understand the generalization of our approach, we test our new models against datasets of articles generated by two companies, AI Writer and AI Forger provided to us by Pu et al. (2023), the Turing Benchmark (Uchendu et al. 2021), four distinct GPT-3.5 generated datasets (OpenAI 2022), and finally a dataset of human-written articles from 2015 (Corney et al. 2016). We now describe our training and test datasets, the architectures of our models, and finally our models' performances on our benchmarks.

Baseline Training Datasets. To train a classifier to detect machine-generated/synthetic news articles found in the wild, we require a diverse dataset of articles from a wide array of generative models. Thus, for our baseline training dataset, we take training data of machine-generated/synthetic articles from three primary sources: the Turing Benchmark, Grover, and articles generated from GPT-3.5.

Machine-Generated Training Articles: For much of our training data, we utilize the Turing Benchmark (Uchendu et al. 2021), which contains news articles generated by 10 different generative text architectures including GPT-1 (Radford et al. 2019a), GPT-2 (Radford et al. 2019b), GPT-3 (Brown et al. 2020), CTRL (Keskar et al. 2019), XLM (Conneau and Lample 2019), Grover (Zellers et al. 2019), XLNet (Yang et al. 2019), Transformer-XL (Dai et al. 2019), and FAIR/WMT (Ng et al. 2019; Chen et al. 2020). We note that given the different settings and trained weights provided by the authors of these respective works, the Turing Benchmark altogether includes articles generated from 19 different models. We randomly subselect 1000 articles from within the Turing benchmark generated by each of these different models as training data.

In addition to the Turing Benchmark training dataset, we use the training dataset of Zellers et al. (2019), which contains realistic, often long-form articles, that mimic the fashion of popular news websites such as cnn.com, nytimes.com, and the washingtonpost.com. Unlike the Grover-generated articles from the Turing Benchmark dataset, which are generated using a prompt of just the title of potential articles, these Grover articles are generated in an unconditional setting *and* from prompting the Grover model with metadata (*i.e.*, title, author, date, website). As found by Zellers et al., many of the articles produced by their models were convincing to human readers, and we thus include 11,930 machine-generated articles from the base model of Grover (across different Grover decoding settings [*e.g.*, p=1.00, p= 0.96, p=0.92 (nucleus/top-p), k=40 (top-k), *etc...* settings]) in our training dataset.

Finally, given the popularity of the GPT-3.5 model (Hu 2023), with it being the basis of the February 2023 released version of ChatGPT, and GPT-3.5 being one of the most powerful released models, we add 3,516 articles generated from the `GPT-3.5 davinici` model. To create these articles, we prompt the public API of `GPT-3.5 davinici` with the first 10 words of 3,516 real news articles from 2018 (see Section 4; while scraping our news dataset, we acquired several million articles from 2018). For `GPT-3.5 davinici` model, we use a nucleus decoding setting of p=1.00, p=0.96, and p=0.92 (some of the most common (Mitchell et al. 2023; Zellers et al. 2019)).

We finally note that, as found in prior work (Pu et al. 2023; Uchendu et al. 2021; Zellers et al. 2019), machine-generated news articles are often shorter in length than human-written articles. While training, to ensure that our models do not simply distinguish between longer human-written articles and those generated by generative transformers by their different lengths, we ensure that our machine-generated and human-written articles are of similar lengths (median training synthetic article length of 210 words and median training human article length of 224 words). Furthermore, as found by past work, predictions for particularly short texts, tend to be unreliable (Kirchner et al. 2023; Zellers et al. 2019; Pu et al. 2023); conversely, as shown by Sadasivan et al. (2023), as the lengths of texts increase the variance between human and machine-generated texts increases. As such, for our training and our generated test data (GPT 3.5 dataset), we exclude texts shorter than 1,000 characters (140 words) (OpenAI 2022). We note as a result, we do not use every trained model's articles from the Turing Benchmark; given that WMT-20/FAIR articles within this dataset are all shorter than 1000 characters, we do not include them within our training dataset. Altogether our training dataset thus only includes data from 19 different models (18 from Turing Benchmark and `GPT-3.5 davinici`).

Training Dataset	Human Written	Machine Generated
Baseline	33,446	33,446
Pert.	33,446	44,003
Para.	33,446	41,498
Perturb + Para.	33,446	52,055

Table 1: The number of machine-generated and human-written articles within the `Baseline`, `Pert`, `Para`, and `Pert.+Para.` training datasets.

Test Dataset	Human Written	Machine Generated
Turing Benchmark	975	18,076
GPT-3.5	1,000	243
GPT-3.5 w/ Pert.	1,000	241
GPT-3.5 w/ Para.	1,000	118
Article Forger	1,000	1,000
AI Writer	1,000	1,000

Table 2: The number of machine-generated and human-written articles within our test datasets.

Human-Written Training Articles: For our set of human-generated articles, as in Zellers et al. (2019), we utilize news articles published in 2018. Specifically, we use 28,446 articles from 2018 from our set of news websites that we later measure (see Section 4; while scraping our news dataset, we acquired several million articles from 2018), 2,500 articles from the human split of the Grover dataset, and 2,500 articles from the human-train-split within the Turing Benchmark dataset.

We present an overview of our complete `Baseline` dataset in Table 1.

Baseline Test Datasets. For our baseline test datasets (Table 2), we utilize the validation split from the Turing Benchmark (the labels from the test split were unavailable to us), and another test dataset consisting of 243 additional GPT-3.5 articles that we created by again prompting `GPT-3.5 davinici`, and 1000 human-written articles from 2018 (see Section 4; as with our training data, while scraping our news dataset, we acquired several million articles from before 2019). Further, to ensure our models generalize and

handle articles seen in the wild, we utilize the *In-the-Wild* dataset provided to us by Pu et al. (2023). This dataset consists of news articles created using generative LLMs from two independent companies, Article Forger and AI Writer. By testing against these outside datasets, we validate our approach against articles generated by (1) models not within our dataset and (2) by generative news article services available to the public. We provide details in Table 2.

Training and Test Dataset using Perturbations and Paraphrases. Transformer-based classifiers are often particularly susceptible to adversarial attacks, particularly attacks that rewrite sections of the generated article (Mitchell et al. 2023; Pu et al. 2023) and paraphrase attacks (*i.e.*, where a generic model is used to paraphrase the output of a different generative model (Krishna et al. 2024)). To guard against these weaknesses, we take two approaches (1) perturbing our set of synthetic articles by rewriting at least 25% of their content using the generic T5-1.1-XL model[2] and (2) paraphrasing articles with the T5-based Dipper model.[3]

Constructing Perturbed Synthetic Articles. To perturb/rewrite sections of our machine-generated articles, as in Mitchell et al. (2023), we randomly MASK 5-word spans of text in each article until at least 25% of the words in the article are masked. Then, using the text-to-text generative model T5-1.1-XL (Raffel et al. 2020), we fill in these spans, perturbing our original generated articles. As shown by Mitchell et al. (2023), large generic generative models such as T5 can apply perturbations that roughly capture meaningful variations of the original passage rather than arbitrary edits. This enables us to model divergences from the distributions of texts created by our 19 different generative models (18 from Turing Benchmark and GPT-3.5). We thus utilize T5-1.1-XL to perturb a portion of the machine-generated articles of our `Baseline` train dataset. In addition, we create a separate test dataset by perturbing our GPT-3.5 test dataset (Table 2). We note that after perturbing our datasets, we filter to ensure all articles used for training contain at least 1000 characters. We annotate training and test datasets containing synthetic articles perturbed with T5-1.1-XL with the suffix `Pert`. After perturbation we still consider these articles to be synthetic.

Constructing Paraphrased Synthetic Articles. To paraphrase each of the machine-generated articles within our dataset, we use the approach outlined by Krishna et al. (2024). Specifically, as in their work, we utilize Dipper, a version of the T5 generative model fine-tuned on paragraph-level paraphrases, that outputs paraphrased versions of the inputted text. We use the default and recommended parameters[4] as in Krishna et al. to paraphrase a portion of the text within our original training dataset as well as our GPT-3.5 test dataset (Krishna et al. 2024). We note that after paraphrasing our datasets, we again filter to ensure all articles utilized for training contain at least 1000 characters (Table 2). We annotate training and test datasets containing articles paraphrased with Dipper with the suffix `Para`. After paraphrasing we still consider these articles to be synthetic.

Detection Models. Having described our training test sets, we now detail our models and evaluate their performance on our 6 test datasets (Turing Benchmark, GPT-3.5, GPT-3.5 w/ `Pert`, GPT-3.5 w/ `Para`, Article Forger, AI Writer). Specifically, we fine-tune three pre-trained transformers, BERT-base (Devlin et al. 2019), RoBERTa-base (Liu et al. 2019), and DeBERTa-v3-base (He et al. 2021).[5][6][7] For each architecture, we train 4 models to detect machine-generated news articles using our `Baseline`, `Perturb`, `Para`, and `Perturb+Para` training datasets. For each architecture, we build a classifier by training an MLP/binary classification layer on top of the outputted [CLS] token. We use a max token length of 512 (Ippolito et al. 2020; Pu et al. 2023), a batch size of 32, and a learning rate of 1×10^{-5}. Each model took approximately 2 hours to train using an Nvidia RTX A6000 GPU. After training, as in Pu et al. (2023), we determine each model's binary F_1-scores, precision, and recall for each test dataset and rank each model using its average F_1-score. We classify each text based on its outputted softmax probability ($>$0.5 being classified as *synthetic*). For a baseline comparison for our trained models, we further test the Roberta-based classifier released by Open AI in 2019 (Solaiman et al. 2019) on each of our test datasets.

Consistent with prior works (Veselovsky, Ribeiro, and West 2023; Mitchell et al. 2023; Gagiano et al. 2021), due to training our model on synthetic articles from a wide variety of sources, and due to our model's focus on news articles, we observe that all our trained models perform markedly better than Open AI's 2019 released detection model. We present the full table of results in Appendix A in Table 11. We further observe, as aggregated in the `Avg.` F_1-score column, that our set of DeBerta models performs the best in classifying machine-generated/synthetic content, all achieving an average F_1-score greater than 0.959. In particular, we observe that our DeBERTa model trained on a dataset that includes our set of adversarial data `Pert` + `Para`, performs the best at an average F_1-score of 0.977. This particular model further achieves the best respective F_1-scores in classifying the set of articles from Article Forger and AI-writer provided by Pu et al., achieving F_1-scores of 0.968 and 0.979 on the two datasets respectively. We note that our model, in addition to performing better than Open AI's Roberta, also outperforms all models benchmarked by Pu et al. (2023) on the AI Forger and the AI Writer test datasets, which achieved F_1-scores ranging from 1.6 to 94.9. This illustrates that our model can generalize to other types of machine-generated articles from models not included in our dataset.

In addition to testing our models on these six datasets, to further ensure that our approach generalizes well, we test our models in two additional settings: (1) a setting where ChatGPT is utilized to rewrite a given human-written article, (2) a setting that includes articles not from the year of train-

[2]https://huggingface.co/google/t5-v1_1-large

[3]https://huggingface.co/kalpeshk2011/dipper-paraphraser-xxl

[4]We use a lexical diversity parameter of 60. For more details on the Dipper model see Krishna et al. (2024)

[5]https://huggingface.co/bert-base-uncased

[6]https://huggingface.co/roberta-base

[7]https://huggingface.co/microsoft/deberta-v3-base

	ChatGPT Rewrite			Signal Art.
	F_1	Prec.	Recall	Accuracy
OpenAI Roberta	0.002	0.200	0.001	0.997
BERT+Para	0.905	0.978	0.842	0.766
RoBERTa+Pert.+Para.	0.937	0.964	0.912	0.820
DeBERTa+Pert.+Para.	0.892	0.979	0.820	0.942

Table 3: We benchmark our BERT `+Para`, RoBERTa`+Pert+Para`, and DeBERTa `+Pert+Para` models, and the OpenAI RoBERTa model on a dataset of 1,000 articles from 2018 rewritten by ChatGPT (along with the original 1,000 human-written articles) and a dataset of 10,000 human-written articles from 2015 chosen randomly from the Signal article dataset.

ing (2018) and from websites not in our original dataset. As such, we finally test the OpenAI Roberta classifier as well as the best BERT, RoBERTa, and DeBERTa models on (1) a set of 1,000 articles from our dataset of 2018 news articles that were rewritten[8] by ChatGPT (OpenAI 2022) as well as the corresponding set original news articles, and (2) 10,000 randomly selected human-written articles from 2015 from the Signal Media news article dataset. As seen in Table 3, our DeBERTa`+Pert+Para` model achieved the highest accuracy on the Signal dataset and the second highest precision on the ChatGPT Rewrite dataset, with scores of 94.2% accuracy and 97.9% precision respectively.

Selecting a classification threshold for *synthetic* articles. Given its performance across all eight of our datasets, we use our `DeBERTa+Pert+Para` trained model as our detection model for the rest of this work. However, as noted in prior research (Krishna et al. 2024), a realistic low false positive rate (FPR) would be near 1%. Given our model only achieves an average FPR of 5.8% on our Signal article dataset at a softmax probability threshold of 0.50, when classifying articles within this work, we raise our softmax probability classification threshold to 0.98, allowing us to achieve a 1% FPR/accuracy on the Signal article dataset. At this threshold, our model achieves a 0.993/0.972 precision/recall on our original six datasets with an FPR of 0.7%. Similarly, at this threshold, our model reaches a precision of 0.989 on our ChatGPT rewrite test set at the expense of only reaching a 0.639 recall. We thus find that by increasing our threshold to 0.98, we can achieve a realistic FPR at the expense of recall. For the rest of this work, we utilize a softmax probability threshold of 0.98. Our work thus likely represents a conservative estimate of the amount of *synthetic* articles online.

4 News Dataset and Classification Pipeline

Having described the DeBERTa-based model that we use to identify machine-generated/synthetic articles, we now describe our datasets of scraped news articles.

[8]We had ChatGPT rewrite each article by supplying the prompt "Rewrite the following news article in your own words:" followed by the article.

Website List. Between January 1, 2022, and May 1, 2023, we gather all articles published from 3,074 news websites.[9] Our list of websites consists of domains labeled as "news" by Media Bias Fact Check[10] and by prior work (Hanley, Kumar, and Durumeric 2023). Within our list of news sites, we differentiate between "unreliable news websites" and "reliable news websites." Our list of unreliable news websites includes 1,059 domains labeled as "conspiracy/pseudoscience" by mediabiasfactcheck.com as well as those labeled as "unreliable news", misinformation, or disinformation by prior work (Hanley, Kumar, and Durumeric 2023; Barret Golding 2022; Szpakowski 2020). Our set of "unreliable" or misinformation news websites includes websites like realjewnews.com, davidduke.com, thegatewaypundit.com, and breitbart.com. We note that despite being labeled unreliable every article from each of these websites *is not* necessarily misinformation.

Our set of "reliable"/mainstream news websites consists of the news websites that were labeled as belonging to the "center", "center-left", or "center-right" by Media Bias Fact Check as well as websites labeled as "reliable" or "mainstream" by other works (Hanley, Kumar, and Durumeric 2023; Barret Golding 2022; Szpakowski 2020). This set of "reliable news websites" includes websites like washingtonpost.com, reuters.com, apnews.com, cnn.com, and foxnews.com. Altogether after removing duplicates and unavailable websites, we scraped 2,015 "reliable news" or mainstream websites.

We note that to later understand how websites of varying popularity/size have used machine-generated articles on their websites, we striate our list of websites by their popularity using ranking data provided by the Google Chrome User Report (CrUX) (Ruth et al. 2022). We note that the CrUX dataset, rather than providing individual popularity ranks for each website, instead provides rank order magnitude buckets (*e.g.*, top 10K, 100K, 1M, 10M websites). As such, we analyze our set of websites in the following buckets: Rank $<$ 10K (125 websites), 10K $<$ Rank $<$ 100K (511 websites), 100K $<$ Rank $<$ 1M (1,164 websites), 1M $<$ Rank $<$ 10M (802 websites), and finally Rank $>$ 10M+ (472 websites).

Article Collection. To collect the articles published by our set of news websites, we queried each website's RSS feeds (if available) and crawled the homepages of each website daily from January 1, 2022, to May 1, 2023. Upon identifying newly published articles, we subsequently scraped websites using Colly[11] and Headless Chrome, orchestrated with Python Selenium. To extract the article text and publication date from each HTML page, we parsed the scraped HTML using the Python libraries `newspaper3k` and `htmldate`.

Given that many of our websites (*e.g.*, cnn.com) have multilingual options, we use the Python `langdetect` library to filter out all non-English articles. To prepare data

[9]We note that while this study focuses on the release of ChatGPT as a possible focal point, our data collection for this project actually began in January 2022.

[10]https://mediabiasfactcheck.com/

[11]https://github.com/gocolly/colly

for classification, we remove boilerplate language using the Python `justext` library and then remove URLs, emojis, and HTML tags. Further, to ensure the reliability of our classifications, we only classify news articles that are at least 1000 characters (approximately 140 words) long. Altogether, from our selection of 3,074 websites, we gathered 15.46M articles (12.06M from mainstream websites and 3.39M from misinformation websites) that were published between January 1, 2022, and May 1, 2023. Finally, we utilize our `DeBERTa+Pert+Para` model at a softmax classification threshold of 0.98 to classify each article as either human-written or machine-generated. Classifying all 15.46M articles took approximately 65.8 hours using an Nvidia RTX A6000 GPU.

Ethical Considerations. With the rise of LLMs, many companies have widely scraped and gathered data from websites to fuel their models (Schappert 2023). As a result, websites ranging from Twitter to Reddit have begun to set up restrictions to ensure the privacy of their users and to protect their content from being used in other private companies' generative models. While we do not train a generative model that could artificially produce convincing and seemingly unique reproductions of the texts that we utilize, we note the concern that our work raises.

Our work, however, only *studies* the texts of our set of 15.46 million articles and classifies them as machine-generated or human-written. We do not seek to generate summaries or artificial rewrites of this content. In terms of web crawling for this data, as noted elsewhere (Singrodia, Mitra, and Paul 2019; Hanley, Kumar, and Durumeric 2023; Smith et al. 2013), website crawling and scraping remain pivotal for understanding and documenting what occurs on the Internet. Without scraping, understanding trends and how the Internet could potentially affect real life becomes impossible. As decided in Van Burn v. United States, publically accessible information can be legally scraped as long as it is done ethically and does not harm the site (Emily R. Lowe and Katrina Slack 2022). As such, we collect only publicly available data from our set of websites and follow the best practices for web crawling as in Acar et al. (2014). We limit the load that each news site experiences by checking for new articles daily at a maximum rate of one request every 10 seconds. The hosts that we scan from are identifiable through WHOIS, reverse DNS, and an HTTP landing page explaining how to reach us if they would like to be removed from the study. During our crawling period, we received no requests from websites to opt out.

5 The Rise of Machine-Generated Media

Having described our detection model and datasets, in this section, we analyze the *relative change* in the levels of synthetic content across our set of websites between January 1, 2022, and May 1, 2023. Specifically, we determine (1) whether there has been an increase in the use of synthetic articles, (2) if there has been an increase in their use, which sets of websites are driving this increase, (3) what synthetic articles are topically about, and (4) whether the introduction of ChatGPT has changed the prevalence of synthetic articles.

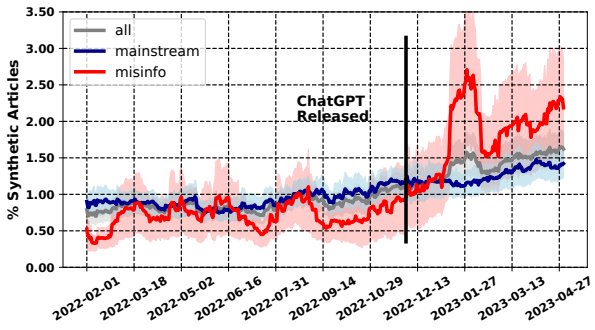


Figure 1: The average percentage of synthetic articles for all, misinformation, and mainstream websites. We provide 95% Normal confidence intervals.

Snippet from Reuters The S&P 500 (.SPX) and Nasdaq (.IXIC) added to losses, while the Dow (.DJI) turned negative on Wednesday after the release of the latest FOMC meeting minutes showed that officials said the central bank may need to raise interest rates sooner than expected and reduce asset holdings quickly.

Figure 2: Example first paragraph of an article classified by our system as machine-generated/synthetic.

Large-Scale Trends in Machine-Generated Media. To begin, we plot the average percentage of synthetic news articles per website across our dataset between January 1, 2022, and May 1, 2023, in Figure 1. In aggregate, across all 3,074 sites, we see that 1.07% of all articles published in January 2022 (12,984 of 1,213,983 articles) were synthetically generated. However, by May 2023, the fraction of synthetic articles nearly went up to 1.78% (25,561 of 1,439,812 articles), a 66.0% relative increase (nearly doubling in raw amount).

We observe that our set of reliable/mainstream websites typically had a greater percentage of synthetic articles at the beginning of 2022 compared with misinformation/unreliable news websites. While only 0.39% of articles on average per domain from our set of misinformation websites were classified as machine-generated in January 2022, 0.88% of articles on average from our set of mainstream/reliable websites were classified as machine-generated. This result is consistent with prior observations that many news websites have begun to use automated services to write quick, often financial-related articles (Section 2). For example, the beginning of one of the articles from Reuters (Figure 2) classified by our system as being machine-generated simply contained simple information about the direction of particular markets and funds.

However, despite reliable/mainstream websites initially having higher levels of synthetic text, misinformation websites had marked increases in levels of machine-generated content during 2022 and 2023 (Figure 1). While between January 1, 2022, and May 1, 2023, reliable/mainstream news websites had a 57.3% relative increase (0.51% absolute percentage increase) in their levels of synthetic content, misinformation websites had a 474% relative increase (1.85% absolute percentage increase). Starting from a lower base,

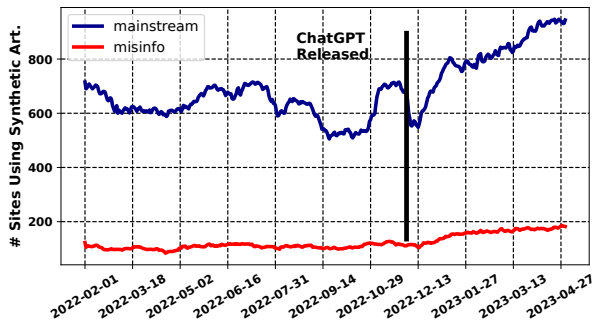


Figure 3: The number of websites that published at least one synthetic article over a 30-day time span.

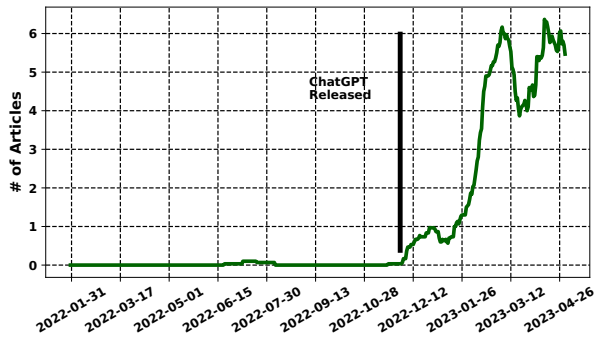


Figure 4: The number of articles that contained a common ChatGPT error message over time.

we thus see a substantial increase in the prevalence of synthetic articles on unreliable/misinformation websites. Furthermore, as seen in Figure 3, we further observe that an increasing number of news outlets published at least one synthetic article within any given 30-day time frame. Across our period of study, the number of mainstream websites that published at least one synthetic article increased from 697 (34.6% of mainstream websites) in January 2022 to 940 (46.6%) in April 2023. Similarly, the number of misinformation websites that published at least one synthetic article increased from 110 (10.4% of misinformation websites) to 179 (16.9%).

To confirm these initial findings, we further examine the increase in common idiosyncratic error messages often returned by ChatGPT. Specifically using a list of error messages including *"my cutoff date in September 2021", "as an AI language model", and "I cannot complete this prompt"* that the company News Guard (Sadeghi and Arvanitis 2023) has used to detect AI-generated websites, we gather every article among our 15.46 million articles that utilized such message: altogether 570 articles from 280 domains. Amongst these websites, the top domains of these articles included forbes.com (32 articles), dailymail.co.uk (29), fairobserver.com (19), theregister.com (13), and patheos.com (13). As seen in Figure 4, we find that while at the beginning of 2022, there were seemingly no such error messages within our set of articles, by the end of April 2023, there were nearly six of these articles each day. We note that this graph also mirrors the behavior of the percentage of machine-generated articles that our DeBERTa detector found amongst all of our websites. Together, these results confirm that there *has* been a noted increase in the use of synthetic content generation by our set of news websites in 2022 and 2023.

We finally note that we observe a small but noticeable dip in the percentage and amount of synthetic content in Figures 1 and 4 (particularly among misinformation websites) between February and March 2023. We find, as seen in Figure 1, that unreliable websites such as foreignpolicyi.org, prophecynewswatch.com, and awarenessact.com, in particular, drove the initial increase in machine-generated content in January and February 2022, before dramatically decreasing their amount of synthetic content in the following month. Examining Google trends data, we also observe that ChatGPT experienced a noticeable dip/decline (from 87% of peak search traffic on February 5 to 75% peak search traffic

	Misinformation		Mainstream	
Rank	Abs. % Inc	Rel. % Inc	Abs % Inc	Rel. % Inc
All	1.85%	474%	0.51%	57.3%
Rk < 10K	0.70%	175%	0.36%	27.0%
10K < Rk < 100K	1.77%	221%	0.26%	27.3%
100K < Rk < 1M	1.38%	349%	0.23%	30.1%
1M < Rk < 10M	1.55%	646%	0.14%	15.4%
Rk > 10M+	3.42%	736%	2.13%	423%

Table 4: Estimated absolute percentage increase in machine-generated/synthetic articles between January 1, 2022 and May 1, 2023.

Jan.2022	**% Syn.**	**CrUX Rank**
opensecrets.org	42.5%	<100K
theodysseyonline.com	26.2%	<1M
logically.ai	17.2%	<10M
china.org.cn	16.3%	<1M
globaltimes.cn	16.0%	<100K
egypttoday.com	15.0%	<1M+
sourcewatch.org	14.4%	<1M
bleacherreport.com	9.84%	<100K
thequint.com	9.81%	<10K
africanews.com	9.64%	<1M

Table 5: Websites with the largest percentage of synthetic content (with at least 100 articles in that month) in January 2022.

April 2023	**% Syn.**	**CrUX Rank**
china.org.cn	34.9%	<1M
globaltimes.cn	26.3%	<100K
thelist.comm	26.0%	<1M
bjreview.com	26.0%	>10M+
thefrisky.com	23.6%	<1M
northkoreatimes.com	23.1%	>10M+
egypttoday.com	21.0%	<1M
waynedupree.com	20.1%	<1M
ancient-origins.net	15.3%	<100K
entrepreneur.com	15.0%	<100K

Table 6: Websites with the largest percentage of synthetic content (with at least 100 articles in that month) in April 2023.

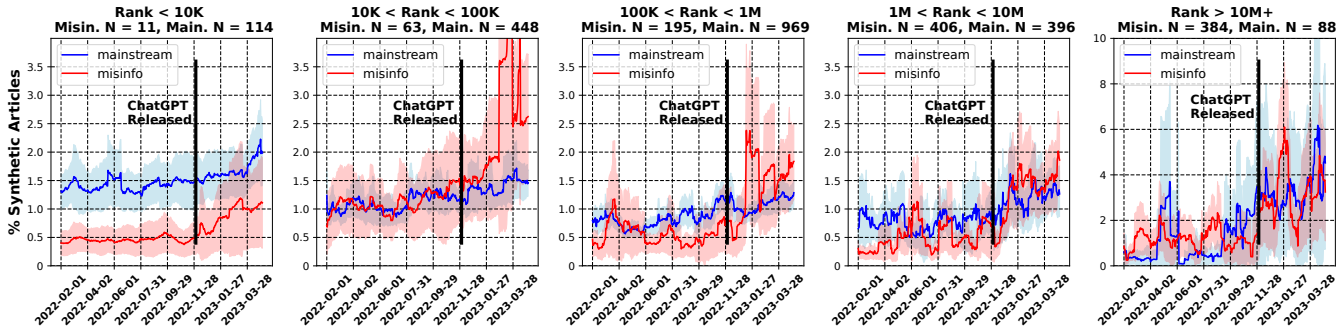


Figure 5: The average percentage of machine-generated/synthetic articles for misinformation/unreliable and mainstream/reliable news websites at different striations of popularity according to Google Chrome User Report (CrUX) from October 2022. All striations of misinformation websites experienced a small uptick of machine-generated content around November 30, 2022, the release date of OpenAI's ChatGPT. We note that the scale of synthetic content is much larger for websites with popularity rank >10M+.

on March 5) in popularity in the United States during this period perhaps (but not definitively) explaining this small decline.

Trends Among Popular and Unpopular Websites. To understand how popularity and website size correlated with the near doubling of machine-generated content in 2022 and 2023, we plot the percentage of machine-generated/synthetic articles over time in Figure 5 for websites within different rank buckets and striated by whether they are considered unreliable/misinformation or reliable/-mainstream. As seen in Figure 5, there is a general upward trend in the amount of machine-generated articles across every popularity stratum.

Examining these increases within particular brackets of popularity, we see (as pictured in Figure 5 and calculated in Table 4) that the least popular websites saw the largest percentage increase in the use of synthetic articles. For both unreliable/misinformation and reliable/mainstream categories, we observe that for websites that rank >10M+ in popularity, the percentage of their articles that were synthetic increased by 3.42% (736% relative increase) and 2.13% (423%) on average, respectively. By contrast, among the most popular misinformation/unreliable websites (*e.g.*, breitbart.com, zerohedge.com) and mainstream/reliable websites (*e.g.*, cnn.com, foxnews.com), synthetic articles had a smaller 0.70% (175% relative increase) and a 0.36% (27.0%) increase overall. Indeed, calculating the websites with the most machine-generated content, we again observe in Tables 5 and 6 that the websites that had the largest amounts of synthetic content were all fairly small or unpopular small.

Topics Addressed by Synthetic Articles. While misinformation websites and less popular websites have seen the largest increase in the use of synthetic articles, many reliable and large news websites also heavily use synthetic articles. However, as noted in Section 2, many reliable news sites have acknowledged their use of these machine-generated articles and utilize them in a benign manner. To understand different websites' use of synthetic articles, in this section,

Topic	Odds Ratio	Topic	Odds Ratio
Entertainment	0.68	Science	1.58
Business	0.61	Sports	0.23
Health	2.06	Technology	0.21
Nation	0.77	World	1.56

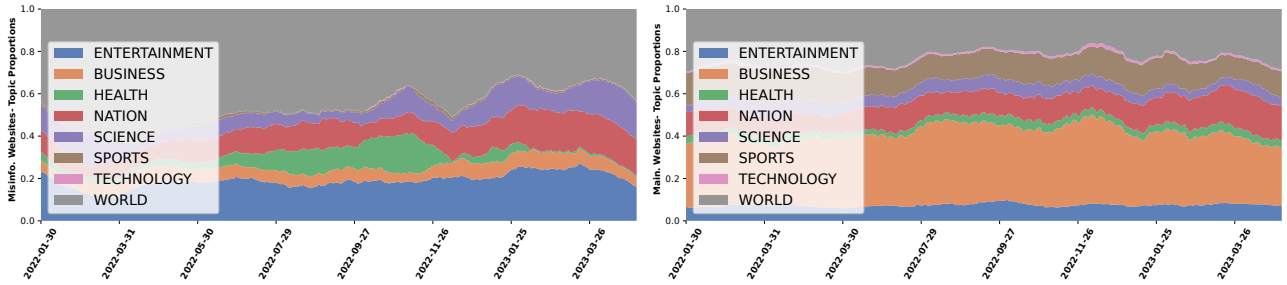
Table 7: Odds Ratio for the amounts of synthetic articles and human-written articles from misinformation websites for each topic category.

we analyze the topics addressed by synthetic articles among different types of websites and how this has changed between January 2022 and May 2023.

To identify the topics within our identified set of machine-generated articles, we train a DeBERTa-based classifier to identify the topic of an article based on its text. As training data, we utilize the News Catcher Topic Labelled dataset,[12] which contains topic labels for 106,395 different articles as belonging to 8 different categories {*Business, Entertainment, Health, US/Nation, Science, Sports, Technology, and World*}. We note while the original dataset only contained the title of each article, the dataset also included the original URL. As such, using the method outlined in section 4, we gather the set of articles listed in the dataset and subsequently train a DeBERTa-based classifier to correctly label articles based on their content. We note that a significant portion of these URLs were not available; as a result, we trained our model on a subset of 79,000 articles from the original dataset, further removing articles that were less than 1000 characters. Keeping out a 10% of this dataset as a test dataset, upon training, we achieve a 0.819 F_1 score an average of 0.819 precision across the eight categories. Once trained, we finally categorize the topic of each of the 15.46 million articles within our dataset.

Plotting the proportion of each topic amongst synthetic articles from misinformation websites, as seen in Figure 6a, a significant portion of synthetic articles from misinformation

[12]https://www.kaggle.com/datasets/kotartemiy/topic-labeled-news-dataset



(a) Misinfo Synthetic Topics

(b) Mainstream Synthetic Topics

Figure 6: The plurality of synthetic articles from mainstream/reliable websites is related to the *Business* topic. In contrast, the majority of synthetic articles from misinformation/unreliable websites are related to *Entertainment*.*World* affairs, and *US/Nation* current events.

Topic	Odds Ratio	Topic	Odds Ratio
Entertainment	0.59	Science	1.48
Business	1.53	Sports	0.85
Health	0.89	Technology	0.66
Nation	1.14	World	0.80

Table 8: Odds Ratio for the amounts of synthetic articles and human-written articles from mainstream websites for each topic category.

Topic	Odds Ratio	Topic	Odds Ratio
Entertainment	2.96	Science	1.91
Business	0.13	Sports	0.06
Health	1.67	Technology	0.11
Nation	1.02	World	2.75

Table 9: Odds Ratio for the amounts of synthetic articles between misinformation and mainstream websites for each topic category. As seen above, misinformation websites are more likely to have synthetic articles about *Entertainment*, *Health*, *Science*, and *World*-related topics compared to mainstream websites. We observe similar proportions of US/Nation topics between misinformation and mainstream websites.

websites concerned *World* affairs, *Nation/US*-current events, *Science*, and *Entertainment*. For example, among our set of synthetic articles from misinformation websites, we identify a variety of articles about concerns about tensions between Russia and Ukraine, COVID-19 vaccines, and updates about the love life of Ed Sheeran. Calculating the odds ratio between the number of synthetic and human-written articles for each of our topic categories, as seen in Table 7, among our selection of misinformation websites, relative to their own topic proportions, misinformation websites were most likely to utilize synthetic articles for *Health* and *Science* related topics. This suggests that misinformation websites *have* proportionally utilized synthetic articles for both mundane topics like Entertainment *and* more serious topics such as Health (Peiser 2019).

Plotting the proportion of each topic amongst synthetic articles from mainstream websites, as seen in Figure 6b, the plurality of synthetic articles concern *Business*. Indeed, as discussed previously, websites ranging from Bloomberg to Reuters have utilized synthetic articles to give updates on financial markets (Figure 2). Furthermore, again calculating the odds ratio between the number of synthetic and human-written articles for each of our topic categories, as seen in Table 8, among our set of mainstream websites, relative to their own topic proportions, mainstream websites are most likely to utilize synthetic articles for *Science* and *Business* topics. This again reinforces prior reporting about the use of synthetic articles among mainstream websites.

Finally, calculating the odds ratio (Table 9) between the rates of usage of synthetic articles per category between mainstream and misinformation websites, we further observe that misinformation news and mainstream websites, throughout our period of study were more likely to utilize synthetic articles on topics related to *Entertainment*, *Health*, and *Science*, and *World* affairs. In contrast, mainstream websites were more likely to utilize synthetic articles for *Business*, *Technology* (very small proportion), and *Sports*. We observe similar proportions of *US/Nation* topics between misinformation and mainstream websites.

Estimating the Impact of ChatGPT. As seen in the previous sections, misinformation websites and less popular websites saw the largest increases in the use of synthetic articles. In order to estimate how the introduction of ChatGPT specifically may have affected the levels of synthetic content on news websites, we now utilize an ARIMA model (Zhang 2003) to perform an *interrrupted-time-series* analysis. Namely, we examine whether there was a direct jump in the number of synthetic articles above expectation following the release of ChatGPT on November 30, 2022 (OpenAI 2022).

As seen in Table 10, after the release of ChatGPT on November 30, 2022, we observe a noted jump (0.50%) above expectation in the number of synthetic articles from misinformation websites. Many of the popularity ranking brackets of misinformation websites saw a statistically significant increase in the absolute percentage of their articles that were synthetic, with misinformation websites in the Rank >10M+ popularity bracket seeing the highest jump of

	Misinformation		Mainstream	
Rank	Abs. % Inc.	Trend % Inc.	Abs. % Inc.	Trend % Inc.
All	$0.50\%^{***}$	$0.006\%^{**}$	0.04%	0.001%
Rk $<$ 10K	$0.10\%^{***}$	$0.004\%^{***}$	0.03%	$0.003\%^{***}$
10K $<$ Rk $<$ 100K	0.10%	0.01%	0.03%	0.001%
100K $<$ Rk $<$ 1M	$0.41\%^{***}$	$0.007\%^{*}$	0.007%	0.0005%
1M $<$ Rk $<$ 10M	0.12%	$0.006\%^{*}$	$0.19\%^{***}$	$0.002^{*}\%$
Rk $>$ 10M+	$1.68\%^{***}$	0.004%	$0.79\%^{***}$	$0.004\%^{***}$

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 10: Estimated absolute percentage increase immediately following the release of ChatGPT on November 30, 2022, in machine-generated articles (determined using an ARIMA-based interrupted time series analysis).

1.68%. This was visually seen in Figure 5. We similarly observe that websites in every popularity bracket except those with Rank >10M+ saw the rate at which the percentage of synthetic articles increases, also increase (*i.e.*, increase in the rate of increase).

We further find that the groups of mainstream websites with popularity ranks >1M saw a marked increase in synthetic articles immediately following the release of ChatGPT on November 30, 2022. In addition, we observe a trend increase for several mainstream website popularity brackets. Combined with our misinformation website results, this suggests that smaller, less popular, and otherwise less monitored websites were the ones that saw the biggest increase in synthetic articles following the release of ChatGPT. Indeed, the number of synthetic articles among *all* groups of websites has been increasing and was at its highest levels on May 1, 2023 (Figure 5). We see this mirrored in the overall increase in the trend of mainstream websites' use of synthetic articles (increase in the rate of increase) in Table 10. We note that while this analysis is *not* causal, it illustrates the noticeable increase in the percentage of synthetic articles among misinformation websites immediately following the release of ChatGPT.

6 Discussion and Conclusion

In this work, we implement a DeBERTa-based model to classify 15.46 million articles from 3,074 news websites as *human-written* or *synthetic*. We find that between January 1, 2022, and May 1, 2023, the percentage of synthetic articles produced by mainstream/reliable news increased by 57.3% while the percentage produced by misinformation/unreliable news websites increased by 474%. Estimating the effect of ChatGPT, we observe a noticeable jump in the percentage of synthetic articles from misinformation websites and unpopular mainstream news around its release. We now discuss several limitations and implications of this work.

Limitations. We note that while we sampled our dataset from a large set of 3,074 news websites and gathered over 15.46M articles, we did not gather articles from *every* news website and focused on English-language media. As such, our results largely do not apply to non-English media. Similarly, because we used pre-defined lists of misinformation websites, our work largely misses the *probable* existence of new misinformation websites that appeared since the launch of ChatGPT.

Because we take a conservative approach to our estimation of machine-*generated/synthetic* texts and due to our removal of articles with characters lengths less than 1000 characters, the absolute numbers presented in this paper are only rough estimates of the percentage of articles on a given website that are machine-generated. As illustrated by Sadasivan et al. (2023), reliable detection of these short texts is near impossible/largely impractical as large language models become more complex. As shown by Sadasivan et al. (2023), as LLMs come to more closely match the distribution of written human language, the distinction between human-written and machine-generate texts disappears. As such, we note that while we manage to create a somewhat reliable detector in this work for longer articles for several released and public models, as more advanced and powerful models are developed, effective detection will be more difficult. Similarly, it has been shown that heavily human-edited machine-generated similarly are very difficult to detect as machine-generated (Mitchell et al. 2023) and in this work, we do not seek to detect these instances. As such, due to our conservative approach, our absolute percentage estimates are likely underestimates.

Furthermore, due to the limitations of our approach in building a model to *estimate* the relative increase in machine-generated texts on news websites, our models are not universal classifiers for *synthetic* texts. Most newspapers and outlets (as of early 2023), are not trying to purposefully evade AI detectors. Our models, which were trained on newspaper data from a given set of websites, are built for a particular context and cannot serve to universally detect synthetic texts.

Detection of Machine-Generated Media. We find that by training on data from a wide variety of generative models, we were able to outperform Open AI's released RoBERTa detector as well as several other released detectors (Pu et al. 2023). Furthermore, we find, as in prior works (Gagiano et al. 2021; Pu et al. 2023), that including data from common attacks can increase overall detection accuracy. We argue that future detectors applied to real-world data should account for these techniques.

Small Websites and Synthetic Articles. As seen throughout this work, while larger more popular websites have been slower to adopt the use of AI-generated and *synthetic* content, smaller less popular websites in particular have shown the greatest relative increase in the use of synthetic (736% increase among the least popular misinformation websites and 423% increase among the least popular mainstream websites). We thus find that to fully understand the influence of synthetic media, as similarly argued by News Guard (Sadeghi and Arvanitis 2023), researchers must document and study these less popular websites rather than just concentrating on the top and most frequently visited domains.

The Rise of Synthetic Misinformation. We found that throughout 2022 and 2023, as LLMs became more widely accessible, the percentage of machine-generated content on

misinformation sites had a 474% relative increase. While at the beginning of 2022, a lower percentage of misinformation/unreliable news websites' content was synthetic (0.39% vs. 0.88%), we find that by May 2023, across all popularity brackets examined, misinformation websites had closed this gap (2.22% vs. 1.39%). Unlike popular mainstream websites, misinformation websites and unpopular mainstream websites experienced a noticeable jump in synthetic content after the release of ChatGPT (as determined by our *interrupted-time-series* analysis). Furthermore, as shown by our topic analysis, misinformation websites have utilized these synthetic articles to address world affairs and health-related news more often than mainstream websites. While not every article posted on an unreliable/misinformation news website is necessarily misinformation, the rapid adoption of synthetic methods by misinformation websites for articles addressing world affairs and health news by these websites could have downstream negative effects. As such given the rapid adoption of the use of synthetic articles by misinformation and unpopular websites, in particular, we argue for future studies of how misinformation websites have utilized these technologies and how the content of these types of articles spread to social media and the broader Internet.

References

Acar, G.; Eubank, C.; Englehardt, S.; Juarez, M.; Narayanan, A.; and Diaz, C. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *ACM Conference on Computer and Communications Security*.

AI, O. 2022. ChatGPT: Optimizing Language Models for Dialogue. http://web.archive.org/web/20230109000707/https://openai.com/blog/chatgpt/. Accessed: 2023-05-01.

Alba, D. 2023. AI Chatbots Have Been Used to Create Dozens of News Content Farms - Bloomberg. https://www.bloomberg.com/news/articles/2023-05-01/ai-chatbots-have-been-used-to-create-dozens-of-news-content-farms. Accessed: 2023-05-01.

Barret Golding. 2022. Iffy Index of Unreliable Sources. https://iffy.news/index/. Accessed: 2023-05-01.

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

Chen, P.-J.; Lee, A.; Wang, C.; Goyal, N.; Fan, A.; Williamson, M.; and Gu, J. 2020. Facebook AI's WMT20 News Translation Task Submission. In *Proceedings of the Fifth Conference on Machine Translation*, 113–125.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Corney, D.; Albakour, D.; Martinez, M.; and Moussa, S. 2016. What do a Million News Articles Look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, 42–47.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J. G.; Le, Q.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *57th Annual Meeting of the Assoc. for Computational Linguistics*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Emily R. Lowe and Katrina Slack. 2022. Data Scraping Deemed Legal in Certain Circumstances. https://www.morganlewis.com/blogs/sourcingatmorganlewis/2022/04/data-scraping-deemed-legal-in-certain-circumstances. Accessed: 2023-05-01.

Gagiano, R.; Kim, M. M.-H.; Zhang, X. J.; and Biggs, J. 2021. Robustness analysis of grover for machine-generated news detection. In *19th Annual Workshop of the Australasian Language Technology Association*.

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

Hanley, H. W.; Kumar, D.; and Durumeric, Z. 2023. A Golden Age: Conspiracy Theories' Relationship with Misinformation Outlets, News Media, and the Wider Internet. *ACM Computer-Supported Cooperative Work And Social Computing*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

He, X.; Shen, X.; Chen, Z.; Backes, M.; and Zhang, Y. 2023. MGTBench: Benchmarking Machine-Generated Text Detection. *arXiv preprint arXiv:2303.14822*.

Hounsel, A.; Holland, J.; Kaiser, B.; Borgolte, K.; Feamster, N.; and Mayer, J. 2020. Identifying Disinformation Websites Using Infrastructure Features. In *USENIX Workshop on Free and Open Communications on the Internet*.

Hu, K. 2023. ChatGPT sets record for fastest-growing user base - analyst note — Reuters. Accessed: 2023-05-01.

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

Jack, C. 2017. Lexicon of lies: Terms for problematic information. *Data & Society*, 3(22): 1094–1096.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Kirchner, J. H.; Ahmad, L.; Aaronson, S.; and Leike, J. 2023. New AI classifier for indicating AI-written text. *OpenAI blog*.

Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; and Iyyer, M. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Leffer, L. 2023. CNET's AI-Written Articles Are Riddled With Errors. https://gizmodo.com/cnet-ai-chatgpt-news-robot-1849996151. Accessed: 2023-05-01.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *60th Annual Meeting of the Assoc. for Computational Linguistics*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 24950–24962. PMLR.

Ng, N.; Yee, K.; Baevski, A.; Ott, M.; Auli, M.; and Edunov, S. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Fourth Conference on Machine Translation*.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-05-01.

Peiser, J. 2019. The Rise of the Robot Reporter - The New York Times. https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html. Accessed: 2023-05-01.

Pu, J.; Sarwar, Z.; Abdullah, S. M.; Rehman, A.; Kim, Y.; Bhattacharya, P.; Javed, M.; and Viswanath, B. 2023. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*, 1613–1630. IEEE.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2019a. Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Ruth, K.; Kumar, D.; Wang, B.; Valenta, L.; and Durumeric, Z. 2022. Toppling top lists: Evaluating the accuracy of popular website lists. In *ACM Internet Measurement Conference*.

Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2023. Can AI-Generated Text be Reliably Detected? *arXiv preprint arXiv:2303.11156*.

Sadeghi, M.; and Arvanitis, L. 2023. Rise of the Newsbots: AI-Generated News Websites Proliferating Online. https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/. Accessed: 2023-05-01.

Schappert, S. 2023. Twitter blocks non-users. https://cybernews.com/news/twitter-blocks-non-users-reading-tweets-ai-scraping/. Accessed: 2023-05-01.

Singrodia, V.; Mitra, A.; and Paul, S. 2019. A review on web scrapping and its applications. In *2019 international conference on computer communication and informatics (ICCCI)*, 1–6. IEEE.

Smith, J. R.; Saint-Amand, H.; Plamada, M.; Koehn, P.; Callison-Burch, C.; and Lopez, A. 2013. Dirt cheap web-scale parallel text from the common crawl. Association for Computational Linguistics.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Szpakowski, M. 2020. Fake News Corpus. https://github.com/several27/FakeNewsCorpus/. Accessed: 2023-05-01.

Tang, R.; Chuang, Y.-N.; and Hu, X. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Uchendu, A.; Le, T.; Shu, K.; and Lee, D. 2020. Authorship attribution for neural text generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Uchendu, A.; Ma, Z.; Le, T.; Zhang, R.; and Lee, D. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP*.

Veselovsky, V.; Ribeiro, M. H.; and West, R. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv preprint arXiv:2306.07899*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending Against Neural Fake News. *Advances in Neural Information Processing Systems*, 32.

Zhang, G. P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50.

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

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our work largely focuses on uncovering the rate of usage of synthetic

articles. Our work does not invade the privacy of individuals, collects only public data, and does not focus on any particular culture.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. Our abstract is largely reflective of our work.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Given our need to understand the use of synthetic articles on news websites, in Section 3 we outline how our methodology precisely identifies synthetic articles.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we note that specific data peculiar to given websites may be part of the data collected. We note that this data is all public, however, and does not reveal the personal data of any given individual.

(e) Did you describe the limitations of your work? Yes, and we detail them in the Discussion.

(f) Did you discuss any potential negative societal impacts of your work? No, we do not believe that there are any immediate negative social repercussions

(g) Did you discuss any potential misuse of your work? No; our work seeks to detect synthetic articles at a large scale. While potentially our system could be utilized to overly penalize those that publish synthetic articles, we have noted in our work that we focus on trends in the use of synthetic articles, not on predicting whether any specific article is synthetic

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, for the articles scraped we only release their URLs and not their texts. We also release our model to GitHub to allow for reproducibility.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes and we have made sure that it is clear.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes, and where appropriate we have interpreted what these results mean.

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, in our limitations section, we discuss several alternative explanations for our data.

(e) Did you address potential biases or limitations in your theoretical framework? Yes, we addressed the potential biases and limitations of our classifier in Section 4.

(f) Have you related your theoretical results to the existing literature in social science? Yes, see Section 2.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, we discussed how research should inform future AI detection systems and discussed the need to understand the growth of websites that primarily publish synthetic data.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we have included a GitHub link.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we have outlined these details in Section 3.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we benchmark and test our models across multiple datasets to better indicate the robustness of our results. We further include error bars for our results.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we have outlined these details in Section 3

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we have outlined these details in Section 3.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we have outlined these details in Section 3.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes, we have cited Pu et. al's work (Pu et al. 2023).

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we have included a GitHub link to the URLs used in this project.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we have cited Pu et. al's work (Pu et al. 2023) and outlined how we have obtained data from them.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, our dataset is findable, accessible, interoperable,

and reusable as we post in on GitHub and it consists of a list of URLs.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, we outline in our GitHub page which domains are included within our dataset, (the methodology for collection is listed in the paper), and we document how it is supposed to be used. We note that our dataset just consists of a list of URLs as we do not release the news article contents.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

	Turing Benchmark			GPT-3.5			GPT-3.5 w/ Pert			GPT-3.5 w/ Para			Article Forger			AI Writer			Avg.
	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
OpenAI Roberta	0.717	0.997	0.560	0.092	0.684	0.049	0.022	0.375	0.011	0.309	0.950	0.185	0.750	1.000	0.600	0.881	1.000	0.787	0.462
BERT	0.988	0.998	0.978	0.941	0.911	0.973	0.901	0.905	0.898	0.931	0.889	0.976	0.855	0.809	0.905	0.779	0.929	0.670	0.899
BERT+ Pert.	0.995	0.992	0.998	0.898	0.817	0.996	0.896	0.816	0.992	0.872	0.777	0.995	0.808	0.681	0.994	0.892	0.771	0.995	0.894
BERT + Para.	0.995	0.992	0.998	0.937	0.896	0.981	0.915	0.892	0.939	0.930	0.822	0.995	0.839	0.763	0.931	0.856	0.888	0.826	0.912
BERT+Pert.+Para.	0.995	0.994	0.997	0.939	0.897	0.985	0.937	0.896	0.981	0.925	0.871	0.985	0.854	0.903	0.913	0.809	0.909	0.729	0.910
RoBERTa	**0.998**	0.997	0.998	0.956	0.929	0.985	0.952	0.928	0.977	0.949	0.911	0.990	0.856	0.857	0.872	0.951	0.933	0.971	0.943
RoBERTa + Pert.	0.993	**1.000**	0.986	**0.979**	0.981	0.977	**0.975**	**0.981**	0.977	0.968	**0.975**	0.961	0.748	**0.977**	0.606	0.820	**0.997**	0.696	0.914
RoBERTa + Para	**0.998**	0.998	0.998	0.940	0.902	0.981	0.932	0.901	0.966	0.934	0.880	0.995	0.903	0.849	0.965	0.958	0.927	0.991	0.944
RoBERTa+Pert.+Para.	0.995	0.991	0.999	0.956	0.923	0.992	0.960	0.923	1.000	0.947	0.903	0.995	0.912	0.859	0.972	0.951	0.913	0.991	0.954
DeBERTa	0.995	0.997	0.993	0.961	0.935	0.989	0.952	0.934	0.970	0.958	0.920	**1.000**	0.959	0.951	0.968	0.986	0.982	0.989	0.969
DeBERTa + Pert.	0.996	0.995	0.996	0.943	0.895	0.996	0.945	0.895	**1.000**	0.930	0.869	**1.000**	0.956	0.927	0.987	0.985	0.975	0.996	0.959
DeBERTa + Para.	0.996	0.993	0.999	0.941	0.892	0.996	0.943	0.892	**1.000**	0.928	0.866	1.000	0.965	0.940	**0.991**	0.983	0.967	1.000	0.959
DeBERTa+Pert.+Para.	0.995	0.994	0.996	0.970	0.949	0.992	0.972	0.949	0.996	0.967	0.936	**1.000**	**0.968**	0.948	0.989	**0.990**	0.979	**1.000**	**0.977**

Table 11: Binary F_1-Score/Precision/Recall of our models on various benchmarks (*machine-generated/synthetic* being positive). We bold the best score in each column. As seen, our set of DeBERTa models performs the best across many of the test datasets, with `DeBERTa+Pert+Para` having the highest average F-1 score across all six datasets.

A Performance of Classifiers