Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

A Large Scale Social Web Audit of AI Generated Text Detection Systems

Arka Dutta[1*], Utkarshani Jaimini[2*], Utkarsh Bhatt[3], Sara Shree Muthuselvam[4], Amitava Das[5], Ashiqur Rahman KhudaBukhsh[1]

[1]Rochester Institute of Technology,
[2]Stride Lab, University of Michigan, Dearborn,
[3]Indian Institute of Technology, Kharagpur,
[4]AIISC, University of South Carolina,
[5]BITS Pilani, Goa
ad2688@rit.edu

Abstract

This paper makes three contributions. First, we exploit temporal signals to conduct an in-the-wild audit of a broad suite of AI-generated text detection (AGTD) systems. Our in-the-wild audit reveals that state-of-the-art (SoTA) AGTD systems exhibit considerable false positives. Second, our audit demonstrates that AGTD systems disfavor liberal political discourse and flags them more often as AI-generated as compared to conservative political discourse. Finally, we extend the anti-content sampling approach to robustify existing AGTD systems.

Datasets —
https://github.com/Social-Insights-Lab/AIGD-Dataset

Introduction

Is this piece of text AI-generated? With generative AI transforming the sociotechnical landscape – from scientific peer review systems (Russo Latona et al. 2024), sponsored content (Bertaglia et al. 2024) to law enforcement (Proctor 2024) – identifying AI-generated texts from human-generated texts will have diverse safety and reliability implications. AI-generated text detection (AGTD) has gained considerable research traction to tackle this pressing concern. With growing concerns of an increased presence of AI-generated content in the information ecosystem, going forward, it is safe to assume that social web content moderation systems are likely to rely on AGTD systems if they already haven't started doing so. However, prior literature indicates AI-aided content moderation systems have been far from infallible (Arango, Pérez, and Poblete 2019; Sarkar and KhudaBukhsh 2021) or unbiased (Sap et al. 2019), and often guided by political subjectivity (Sap et al. 2022; Weerasooriya et al. 2023b).

How do AGTD systems fit into the current content-moderation puzzle? Are they ready for real-world deployment? Are they politically neutral? Do they disparately affect certain demographics? Content moderation instruments do not reside in a vacuum and these are important questions with practical consequences. This paper presents a comprehensive in-the-wild audit of AGTD systems leveraging a simple intuition: a social web post authored on or before 1st January, 2019 (before `GPT-2` was made publicly available) is highly unlikely to be AI-generated. At present, there are very few benchmark datasets to evaluate AGTD systems. Utilizing document creation timestamps as an implicit guarantee of the content not being AI-generated and considering social web data which consists of a vast pool of human-generated texts of diverse content and styles allows us to conduct a comprehensive in-the-wild audit at scale.

Our paper investigates the following research questions:

RQ1: Can the existing state of the art AGTD models effectively detect the AI-generated text?

RQ2: How do the political leanings factor into AGTD models' performance?

RQ3: Do these AGTD techniques disproportionately suppress minority voices?

RQ4: Can anti-content sampling (Yoo and KhudaBukhsh 2023) finetune the model performance?

The relationship between US political discourse and web censorship is a fraught one, often marred by accusations of hyper-partisanship. While the Section 230 of a bipartisan Communication Decency Act presented burgeoning internet businesses in late 90's with broad immunity to decide moderation guidelines as they please (Kosseff 2019), algorithmic amplification of political content (Huszár et al. 2022) and differential censorship (Zakrzewski and Lima-Strong 2023) equally concern conservatives and liberals in current polarized US politics. As we are writing this paper, the US political history witnessed a bitterly fought election. During the campaign trail, President-elect Trump has accused Vice President Kamala Harris of cheating and election interference using AI-manipulated images of massive rally crowds (Goldmacher 2024). During this politically turbulent phase, the sociopolitical implications of AGTD systems are amplified. How do we know that the tremendous surge of positive vibe for candidate $\mathcal{A}$ or a sudden increase in social network exhaustion on candidate $\mathcal{B}$ are not AI-manufactured? Type 1 error (human-generated content is misidentified as machine-generated content) has become as

*These authors contributed equally.

Key idea: We leverage a simple assumption to conduct in-the-wild audit of AGTD systems: *before 1st January 2019 (before `GPT-2` was made publicly available), a social media post is highly unlikely to be AI-generated.* Any large scale audit conducted on social web data before 2019 should ideally yield very close to 0% AI-generated text. A high percentage of pre-2019 posts detected as AI-generated content indicates models' unsuitability for real-world deployment. Following (Yoo and KhudaBukhsh 2023), we further tap into the richness and diversity of a vast amount of implicitly labeled (as negative, or non-AI-generated) pool of social media posts (dubbed anticontent) and design an active learning framework. Starting with an annotated AI-generated text dataset, at each step, we train a model on labeled data and evaluate on the pool of implicitly labeled anticontent instances. Instances that are classified as positives (i.e. AI-generated text) with high confidence are added as challenging negatives to augment the train data.

important as the Type II error (machine-generated content is misidentified as human-generated content).

This paper focuses on the type I error - auditing how often human-generated texts get misidentified as machine-generated by cutting-edge AGTD systems. While our particular focus is on political discourse on the social web, our work presents a simple audit framework to test the effectiveness of SoTA AGTD systems on a vast pool of human-generated texts with diverse styles and content. When human-generated texts are misidentified as machine-generated, we run the risk of eliminating legitimate voices that may lead to political unfairness and algorithmic invisibility. Our paper thus contributes to an important understanding of the implications of AGTD systems to the future of the web moderation.

To summarize, our contributions are the following:

Framework: We present a novel framework to conduct in-the-wild of AGTD systems using temporal signals. To our knowledge, this simple-yet-powerful audit design leveraging temporal signals is novel in the context of AGTD systems audit.

Social: Our audit reveals that SoTA AGTD systems are not yet ready for real-world deployment as most of them record false positives at an alarming rate. In addition, these systems are not politically neutral and might marginalize historically disadvantaged groups such as the LGBTQ+ and Black communities potentially leading to algorithmic invisibility.

Method: Follwoing (Yoo and KhudaBukhsh 2023), we extend anticontent sampling, a zero-human-annotation active learning framework relying on implicit labels, and demonstrate a modest performance improvement in AGTD systems.

Datasets

In this work, we focus on the political discourse on Reddit and YouTube. In contrast with Twitter, both Reddit and YouTube put a character limit of 10,000 characters for individual posts or comments allowing room for nuanced political discourse.

For YouTube, we consider a dataset of 60,000 user comments equally sampled from news videos hosted on the official YouTube channels of three major cable news outlets in the US: CNN, Fox News, and MSNBC (SI contains further details). We select this dataset because of the broad participation, topical diversity, and political relevance (Stanley 2012; Bozell 2004; Gil de Zúñiga, Correa, and Valenzuela 2012; Hyun and Moon 2016; Dutta et al. 2022; KhudaBukhsh et al. 2022; Ding, Horning, and Rho 2023). This dataset is considered a comprehensive snapshot of US political discourse (KhudaBukhsh et al. 2021) and has been used for in-the-wild audits of content moderation systems (Weerasooriya et al. 2023a) and misinformation datasets (Yoo and KhudaBukhsh 2023). in addition, to check for *algorithmic chokehold*, we create another separate dataset using comments from these three major cable networks concerning LGBTQ+ individuals. This dataset consists of 1,651 total comments for 2019 and 2024 with token-length exceeding 50[1].

Subreddit	\r Republican	\r Democrats	\r BlackPeopleTwitter
2019	16,325	17,747	289,242
2024	8,729	12,000	110,406

Table 1: The number of comments with more than 50 tokens in each subreddit for the years 2019 and 2024

To further support our hypothesis, we created two Reddit datasets dealing with 1) politics and 2) algorithmic chokehold. We used arctic shift reddit API[2] for the data collection. These datasets, which we collected from subreddits republic, democrats, USPolitics for political views, and BlackPeopleTwitter, are significant in their extensive scope and depth. They include titles, descriptions, and a substantial 454,449 comments across the subreddits. For the experiments in this paper, we consider comments from subreddits *Republican, Democrats and BlackPeopleTwitter* for the years 2019 and 2024. The dataset description is provided in Table 1.

Models

We consider five models for their innovative approaches and SoTA performance in detecting AI-generated text, addressing challenges such as paraphrasing, domain adaptation, and black-box LLMs. These methodologies, including adversarial training, edit-distance invariance, and intrinsic dimensionality, provide a diversity for our evaluation. This selection of models also ensures inclusion of both trainable models (e.g., `RAIDAR`, `Ghostbusters`) and non-trainable models (e.g., `RADAR`, `Binoculars`) for our audit. A brief description of each of these models follows next.

[1]Our preliminary experiments revealed that AGTD systems' performance is even more unreliable for shorter documents.

[2]https://arctic-shift.photon-reddit.com/download-tool

RADAR

RADAR (Hu, Chen, and Ho 2023) uses adversarial learning to jointly train the AI-text generator (paraphraser) and AI-text detector. The paraphraser uses the detector's prediction as a reward and updates its policy using proximal policy optimization (PPO) (Schulman et al. 2017). The detector, in turn, updates its parameters based on the logistic loss function evaluated on both human and AI-generated text. The RADAR model's robustness in handling multiple paraphrasing models makes it a promising tool. RADAR model[3][4] is available as a trained model for evaluation, and it can't be fine-tuned from user-end.

RAIDAR

By prompting LLMs to rewrite the text, RAIDAR (Mao et al. 2024) measures the invariance property—machine-generated text undergoes fewer changes compared to human-written text. This method uses edit distances measures such as Levenshtein distance and n-gram edits to evaluate output stability, noting that LLM-generated text shows less variance upon multiple rewrites. This approach leverages discrete token outputs, making it robust and compatible with both black-box LLMs and open LLMs. This method reaches State-of-the-art accuracy on various benchmark datasets across various LLMs. As this model uses interpretable and trainable estimates to detect machine-generated text, it poses a perfect test case for our experimental design. For our experimental design, we train several ML classifier models to achieve the standard 1% FPR on test set and report the accuracies in the tables.

Intrinsic Dimensionality

Tulchinskii *et al.* (Tulchinskii et al. 2023) focus on the Persistent Homology Dimension (PHD) estimator, which combines local and global dataset properties. They utilize PHD to estimate the dimension of text embeddings, derived from pre-trained Transformer models, as a key feature for detecting AI-generated text. The method involves sampling, linear regression, and averaging to improve the stability and accuracy of ID estimation, ultimately training a logistic regression classifier for AI-generated text detection.

Intrinsic-RAIDAR

As the intrinsic dimension and RAIDAR both use some internal *signals* from the text for classification, we theorize that combining the features to train a final classifier may render more robust results. Based on this assumption, we train several ML classifiers using n-gram distance, Levenshtein distance (following RAIDAR) and PHD, and MLE (following intrinsic dimension). We use the best-performing classifier rendering the consistent baseline of 1% FPR on original validation data.

[3]https://github.com/IBM/RADAR

[4]https://huggingface.co/spaces/TrustSafeAI/RADAR-AI-Text-Detector

Ghostbusters

Ghostbuster (Verma et al. 2024) leverages structured search and token probabilities from weaker language models (e.g., unigram, trigram, and early GPT-3 models) to extract distinguishing features, followed by training a logistic regression classifier on these features. Unlike few SoTA classifier models, it does not require access to token probabilities of the target model, making it effective even for black-box models. Authors show a remarkable 99.0 F1-score, outperforming existing detectors like DetectGPT (Mitchell et al. 2023) and GPTZero (Tian and Cui 2023) by significant margins.

Binoculars

Binoculars (Hans et al. 2024) contrasts perplexity and cross-perplexity metrics derived from two similar LLMs. By leveraging ratios of these statistical signals, the model achieves SoTA detection without requiring training data, enabling it to generalize across different LLMs, including ChatGPT and LLaMA models. The methodology is tested extensively across varied datasets, showing robustness in detecting AI-generated text in multiple domains and outperforming existing commercial and open-source detectors, especially in out-of-domain scenarios.

Related Work

AI Generated Content Detection

In addition to the models we used for our audit, there exist alternative strategies, such as heuristic detectors and ensemble detectors among various methods that achieve strong detection performance. DetectGPT (Mitchell et al. 2023), one of the earliest methods using quantitative methods, employs a zero-shot framework, leveraging the curvature of a language model's log-probability function to distinguish AI-generated text without additional training, achieving strong performance across diverse tasks like such as fake news detection. ConDA (Bhattacharjee et al. 2023) takes a different approach, addressing the task as an unsupervised domain adaptation problem by aligning domain-invariant representations using contrastive learning, delivering robust results even in the absence of labeled target domain data. Expanding on these efforts one of the most recent works, LLM-DetectAIve (Abassy et al. 2024) introduces fine-grained classification to detect nuanced human-machine collaborations, such as polished or humanized AI text, outperforming existing detectors and offering an interactive demo.

AGTD system audit and datasets

Audits. Recent studies have explored the challenges and limitations of detecting AI-generated content, with particular emphasis on fairness, methodology, and real-world applicability. Li *et al.* (2024) introduce the ARIA dataset, comprising over 140,000 real and AI-generated images, to benchmark detection methods and evaluate human performance in distinguishing AI-Art. While their work offers valuable insights, it is limited to controlled experimental settings, overlooking the complexities of social web data, political contexts, and temporal dynamics, which are critical for

real-world applications. Similarly, Skumanich *et al.* (2024) analyze AI-generated disinformation on fringe platforms like Gab and Gettr, focusing on narrative shifts over time. Although their study incorporates temporal signals, it does not address broader issues such as political censorship, minority group impacts, or techniques like anticontent sampling that could enhance detection robustness.

Ramu *et al.* (2023) investigate Generation Z's ability to identify AI-generated text through user studies on Discord, highlighting participants' difficulties, particularly with short-form content. Their findings critique tools such as OpenAI's Text Classifier and `GPTZero` for high false positive rates and vulnerability to paraphrasing attacks. However, the study does not leverage data from mainstream platforms like Reddit or YouTube to evaluate demographic or political biases in detection tools. On a similar note, Elkhatat *et al.* (2023) evaluate five AI content detection tools (OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag) in identifying text from ChatGPT models (3.5 and 4) and human-written content. Using 15 paragraphs each from GPT-3.5, GPT-4, and five human-written control samples, the study measures sensitivity, specificity, and predictive values. Results indicate better performance on `GPT-3.5`-generated text compared to `GPT-4` or human-written content, but the findings are constrained by the small sample size and the limited scope of detectors, underscoring the need for broader evaluations in a rapidly evolving landscape.

Sadasivan *et al.* (2023) assess the reliability of detection methods, including watermarking, neural network-based, zero-shot, and retrieval-based detectors. They introduce a recursive paraphrasing attack that effectively degrades detection accuracy with minimal loss in text quality, as validated through perplexity metrics and human evaluations. Additionally, they demonstrate vulnerabilities to spoofing attacks, where human-written text can be misclassified as AI-generated, potentially causing reputational harm. A theoretical link between detector performance (AUROC) and total variation (TV) distance highlights the growing difficulty of distinguishing human and AI text as models become more advanced. The study emphasizes the susceptibility of current detectors to adversarial attacks and the inherent trade-offs between type-I and type-II errors.

Datasets. The datasets used in AI-generated text detection span diverse domains, reflecting varied linguistic styles and complexities. Creative writing data, sourced from the WritingPrompts subreddit, captures imaginative storytelling, while the Reuters 50-50 dataset (Liu 2006) serves as the foundation for news article authorship. Academic writing is represented through the British Academic Written English corpus (Verma et al. 2024) and the SciXGen dataset (Chen, Takamura, and Nakayama 2021), featuring student essays and scientific abstracts, respectively. Programming tasks are explored using the HumanEval dataset (Chen et al. 2021) with human-written and `GPT-3.5`-generated code. Short-form and opinion-heavy texts are included via Yelp reviews and opinion statements from r/ChangeMyView. Additional datasets such as XSum, TLDR_news, and Wikipedia paragraphs (e.g., SQuAD contexts) cover news summarization and encyclopedic writing. Storytelling and reasoning datasets like ROCStories, ELI5, HellaSwag (Zellers et al. 2019), and WritingPrompts further diversify the collection, ensuring a comprehensive evaluation of detection systems across creative, factual, and logical text generation.

Across these works, the datasets range from curated benchmarks like ARIA to niche platform data, yet the integration of social web data, political biases, and demographic impacts remains largely unexplored. Also the discussion on the potential of AI detectors to perpetuate discrimination or disproportionately silence marginalized groups, such as Black and LGBTQ+ communities is absent in current literature. Our work presents a cost-effective solution to this fleeting scenario of AGTD systems audit with further emphasis on minimizing *Algorithmic Chokehold* and model biases.

Anticontent Sampling

Yoo *et al.* (2023) introduced anticontent sampling in the context of robustifying COVID-19 misinformation datasets. This approach first identifies an unlabeled set with implicit labels. Next, it sets up an active learning pipeline leveraging this implicit labels. Yoo *et al.* (2023) argued that the social web discourse prior to COVID-19 is unlikely to contain any misinformation about COVID-19. With this assumption, they trained content classifiers on existing COVID-19 misinformation datasets and ran inference pre-COVID-19 social web discourse. The intuition is a social web post classified as COVID-19 misinformation represents a challenging negative (anticontent) that the classifier misclassified as COVID-19 misinformation. Their active learning pipeline leveraged this idea methodology and showed how it enhances misinformation classifiers by leveraging pre-COVID-19 social media posts as implicitly labeled non-misinformation to address shortcut learning and out-of-domain (OOD) generalization issues. An active learning framework iteratively identifies and augments the training set with challenging anticontent misclassified by the model, improving robustness without manual annotation. This approach significantly boosts performance, with F1-scores rising from 87.7% to 95.2%, and drastically reduces pre-COVID-19 false positives from 14.3% to 0.2%.

Based on this idea, we employ a similar technique in our method to robustify AGTD systems where we reasonably assume that pre-GPT2 (2019 and before) there can not exist content that is generated by generative LLMs. With this assumption, in an active-learning setting, we retrain the trainable AGTD systems iteratively through augmenting the training set with mislabeled *challenging* ground truth samples from previous iteration. We continue this till we achieve a 1% FPR and further evaluate on social web datasets.

Disambiguating Bot Content in Social Media

At a philosophical level, the vast bot detection literature (see, e.g., (Cresci 2020; Ng and Carley 2023; Uyheng and Carley 2021)) aligns with the AGTD literature in their shared goal to preserving the integrity of the social web. However, there is a key distinction. Research and platform audits consistently indicate that, prior to 2019, automated accounts ('bots') on social platforms such as Reddit, Twitter, and

YouTube were primarily amplifiers of human-written content, rather than independent content creators. Bots largely shared, reposted, or promoted material originally written by humans (news articles, blog posts, or other users' messages), instead of composing substantial new text themselves. For example, a 2017 Cambridge University study of Twitter behavior found that bot accounts relied "a lot more on retweets and redirecting followers to external websites" – essentially recycling existing content – because "bots aren't that good at creating original Twitter content" (University of Cambridge News Office 2017). In the same vein, a Pew Research Center analysis noted that suspected bots were extremely active in sharing links to news articles and other web content on Twitter (Center 2018), accounting for an estimated 66% of links to popular websites in one study. These findings show that the typical role of bots was to amplify human-generated information (often from news sites or propaganda sources) rather than to generate new persuasive text from scratch. In contrast, this paper focuses on the reliability and robustness of methods that detect whether content is generated using AI rather than who posted the content.

Large-scale generative text production by bots was widely considered implausible before powerful language models like `GPT-2` emerged in 2019. In-the-wild evidence for bots autonomously writing long-form coherent posts at scale was virtually non-existent up to that point. Instead, bad actors tended to use bots as distribution vehicles, for instance, networks of bots that retweet misinformation or paste the same prepared comments, rather than as autonomous authors of novel content (Vosoughi, Roy, and Aral 2018). In short, before recent advances in AI, it was technically unfeasible for bots to mass-produce convincing new content. Although the recent literature has begun to grapple with a combined challenge (ie, social bots that post AI-generated content (Ferrara 2023)), our current study decouples the content from the vehicles that propagate them through the social web.

Experiments

Model audit

Our model audit mainly consist of two major components:

Model Training

- **Pre-Trained Models:** Proprietary pre-trained models such as `RADAR` and `Binoculars` were utilized directly without fine-tuning. These models were evaluated on their out-of-the-box performance on both original and paraphrased datasets.
- **Trainable models:** Non-proprietary models or feature-based model such as `RAIDAR`, `Ghostbusters`, and Intrinsic dimension were trained on the original dataset mentioned in their respective literature, ensuring a consistent 1% FPR on the test split. For the intrinsic dimension model, the training process included use of Persistent Homology Dimension (PHD) (Schweinhart 2021), Manifold-adaptive dimension (MADA) (Farahmand, Szepesvári, and Audibert 2007), and Maximum Likelihood Estimation (MLE) (Levina and Bickel 2004) as features. The training process of `RAIDAR` included n-gram distance and Levenshtein distance as features. `RAIDAR` involves rewriting the ground truth (human-written text) through an LLM and subsequently train the models based on *edit-distance* of the rewritten text. We employ two prominent LLMs: `Llama3` (Grattafiori and team 2024) (an open model) and `GPT-4o` (team 2024) (a proprietary model) for the rewriting and evaluating both separately and in conjunction. We used multiple ML classifiers, trained on these features, to detect AI-generated content, leveraging both global and local text properties. For the intrinsic dimension model, a simple logistic regression model was found to be the best to achieve a 1% FPR while for the RAIDAR model, an XGBoost model was found best-fit.
- **Paraphrasing:** We evaluated the robustness of various models using comments paraphrased through multiple paraphrasers. Specifically, we used `ChatGPTT5`, `Pegasus`, and `Llama` to generate paraphrased versions of the comments. These paraphrased comments were then subjected to a second round of paraphrasing using the same paraphrasers. The twice-paraphrased comments were subsequently analyzed by the models to determine if they could identify them as AI-generated. The models identified the twice-paraphrased comments as AI-generated with 99% confidence. Among the paraphrasers, `ChatGPTT5` demonstrated superior performance occasionally matching or exceeding the performance of `Pegasus` and `Llama` across all models.
- **Dataset Context Length:** We evaluated models on tweets and similar short-form content for comparison. Short-term contexts provided insufficient data for reliable detection, leading to higher FPRs. To address this issue, we focused on Reddit and YouTube user comments with a token length of more than 50. The longer context provided more comprehensive information, resulting in higher detection accuracy and better generalization. Final evaluation prioritized longer context datasets for robustness and reliability.
- **Anticontent Sampling:** We employed an interative anticontent sampling-driven active learning pipeline (Yoo and KhudaBukhsh 2023) to refine AI-generated text detection models. We first present a brief description of anticontent sampling.

  An active learning pipeline guided by anticontent sampling works in the following way. First, it identifies a set $\mathcal{S}$ with an implicit label (e.g., negative). In Yoo and KhudaBukhsh (2023), in the context of COVID-19 misinformation, $\mathcal{S}$ was set to pre-COVID-19 era social media discussions. For a misinformation dataset $\mathcal{D}$, all instances in $\mathcal{S}$ are classified by a model trained on $\mathcal{D}$. Since before COVID-19, no social web post can be a COVID-19 misinformation post, high-confidence positives (misinformation) are in fact challenging negatives (*anticontent*) that the model struggles with. These examples are augmented with $\mathcal{D}$ with the flipped label (negative, not-misinformation) and a new model is trained on the augmented set. This iterative process of identifying challenging negatives and augmenting them with the labeled

dataset continues till a halting criterion is reached. The iterative expansion of the training set guided by model performance mimics active learning (Settles 2009). However, one key advantage is that in contrast with traditional active learning, anticontent sampling uses zero manual annotation as it leverages implicitly labeled $\mathcal{S}$.

In what follows, we outline the key steps of our pipeline in which we extend anticontent sampling to strengthen AGTD systems.

1. **Initial Model Audit:** First, we initialize trainable models using temporal signals, leveraging an equal mix of pre-2019 non-AI-generated content and AI-generated paraphrased versions of the same content (e.g., rewritten by `GPT-4o`, `Llama3`). We then apply the AI-generated text detector to a large corpus of known human-written posts (e.g., social media content from before 2019). Any posts that the detector mistakenly flags as AI-generated (false positives) are strong candidates for *anticontent*. These are human-written examples that the model finds confusing.
2. **Selection of Anticontent:** From those false positives, we select a set of diverse, representative examples (throughout our experiments, the top 50% of highest-confidence false positives). This selection is analogous to choosing *challenging negatives*.
3. **Augmentation and Fine-Tuning:** We then consider the implicit label (human-written, negative class) of these examples and add them to the training data, effectively augmenting the negative class with content that the model previously struggled with. The detector is retrained or fine-tuned on this expanded dataset, which now includes the challenging anticontent. We repeat this process. After retraining, we audit the model again to find new high-confidence false positives (anticontent), add them, and repeat (as resources permit). Prior work (Yoo and KhudaBukhsh 2023) demonstrated that this kind of iterative, self-supervised active learning can substantially improve robustness without any manual labeling effort. This iterative process continued until the model achieved the target FPR of 1%.

Model Evaluation

- **Overall trends:** To assess the overall trend, we evaluated AGTD systems on datasets from both YouTube (CNN, MSNBC, FOX) and Reddit (r/Democrat, r/Republicans, r/Blackpeopletwitter). This provided a comprehensive view of model performance across different platforms and discourse types focusing on social web.
- **Political bias trends:** To analyze potential political bias, we focused on liberal (CNN, MSNBC, r/Democrat) and conservative (FOX, r/Republicans) discourse. The results highlighted disparities in detection rates between these ideological spectrums.
- **Algorithmic Chokehold:** To investigate biases against minority voices, we specifically evaluated AGTD systems on r/Blackpeopletwitter from Reddit and LGBTQ+-related content from CNN, MSNBC, and FOX on YouTube. This allowed us to audit whether these systems disproportionately misclassify marginalized community discourse.

Results

Audit Findings

Finding I: SoTA AGTD systems exhibit high false positive rates.

Our findings demonstrate that SoTA AGTD systems exhibit significant false positive rates, particularly when assessing social media discourse and YouTube comments. As illustrated in Tables 2 and 3, these systems frequently misclassify human-generated content as AI-generated in 2019, a period before the adoption of generative LLMs. Notably, `RAIDAR` detects AI-generated content probabilities as high as 16.9% ± 4.71% for MSNBC comments and 17.9% for r/Democrat comments in 2019, while `RADAR` reports considerably higher probabilities, such as 62.46% for *r/Democrat* and 72.92% for *r/Blackpeopletwitter*, suggesting greater over-sensitivity to linguistic features. In contrast, `Binocular` and `Ghostbusters` show relatively lower, though still inflated, detection rates, with Binocular detecting 15.2% for *r/Democrat* and Ghostbusters reaching 17.28% ± 2.4% for CNN comments. In 2024, a minor increase in detection probabilities is observed across all models, such as `RAIDAR`'s rise to 23.25% for r/Democrat and Ghostbusters' increase to 17.6% ± 3.2% for CNN. However, these increments fail to establish a reliable baseline due to the inflated detection rates in 2019. The differences in model behavior reveal varying degrees of misclassification: `RADAR` consistently overestimates detection rates across all datasets, while `RAIDAR` exhibits more moderate but still erroneous results, and `Binocular` tends to perform conservatively but remains susceptible to false positives. These discrepancies highlight systemic issues in all models, indicating that none can reliably distinguish between human-written and AI-generated content. Hence, the short answer to the question of whether AGTD models are ready for real-world deployment is a resounding no.

Finding II: Liberal discourse is (slightly) more likely to be flagged as AI-generated than conservative discourse.

Analysis of the 2019 data reveals that liberal discourse, as represented by MSNBC and r/Democrat, is marginally more likely to be flagged as AI-generated compared to conservative discourse, represented by FOX and r/Republicans. From Table 2, `RAIDAR` detects AI-generated content probabilities of 16.9% ± 4.71% for MSNBC comments compared to 14.8% ± 1.46% for FOX. Similarly, Table 3 shows `RAIDAR` identifying 17.9% of r/Democrat comments as AI-generated, while detecting only 11.2% for r/Republicans. This trend persists across models, with `RADAR` reporting disproportionately high probabilities for r/Democrat (62.46%) relative to r/Republicans (40%). While the discrepancies are less pronounced for models such as `Binocular` and `Ghostbusters`, the overall pattern remains consistent. `Binocular` shows slight bias against liberal discussion as observed in Table 3 for r/Democrat vs r/Republicans while `Ghostbusters` show the similar bias as observed in Table

Models	Year	CNN	MSNBC	FOX
RAIDAR	2019	13.85% ± 2.34%	16.90% ± 4.71%	14.80% ± 1.46%
	2024	16.74% ± 3.43%	17.48% ± 2.79%	14.34% ± 0.92%
RADAR	2019	66.00% ± 3.80%	69.24% ± 4.05%	65.63% ± 3.25%
	2024	71.23% ± 3.10%	75.15% ± 3.95%	75.53% ± 2.90%
Ghostbusters	2019	17.28% ± 2.40%	21.40% ± 0.85%	18.60% ± 1.68%
	2024	17.60% ± 3.20%	20.82% ± 1.97%	19.26% ± 2.61%
RAIDAR+Intrinsic	2019	12.65% ± 1.87%	16.15% ± 3.10%	13.45% ± 1.85%
	2024	16.68% ± 2.42%	16.78% ± 5.71%	14.88% ± 1.81%
RAIDAR Anticontent Sampled	2019	**11.00% ± 2.20%**	**15.33% ± 2.17%**	**9.26% ± 2.69%**
	2024	16.65% ± 1.13%	16.93% ± 3.36%	13.57% ± 0.72%

Table 2: Probability of AI-generated-content detection on YouTube comments for CNN, MSNBC, and FOX News across 2019 and 2024 (mean ± 95% CI).

Models	Year	r/Democrat	r/Republicans	r/Blackpeopletwitter
RAIDAR	2019	17.90% ± 3.38%	11.20% ± 2.9%	8.66% ± 2.41%
	2024	23.25% ± 3.7%	22.20% ± 3.64%	10.16% ± 2.73%
RADAR	2019	62.46% ± 3.92%	40.00% ± 4.18%	72.92% ± 3.76%
	2024	72.16% ± 3.58%	40.07% ± 4.2%	70.05% ± 3.69%
Binocular	2019	15.20% ± 3.04%	14.60% ± 2.91%	21.42% ± 3.83%
	2024	15.12% ± 2.87%	10.02% ± 2.46%	20.48% ± 3.71%
Ghostbusters	2019	13.34% ± 2.93%	18.12% ± 3.29%	12.50% ± 2.77%
	2024	15.26% ± 3.02%	17.92% ± 3.18%	16.75% ± 3.11%

Table 3: Probability of AI-generated-content detection across subreddits r/Democrat, r/Republicans, and r/Blackpeopletwitter in 2019 and 2024 (mean ± 95% CI).

2 for MSNBC and FOX discussions. These results suggest a systematic bias in AGTD systems, which are more prone to flagging liberal-oriented discourse as AI-generated, potentially due to linguistic or stylistic features more prevalent in these datasets.

Finding III: AGTD systems may affect disadvantaged groups: case study on Black and LGBTQ+ community reveals. Our audit reveals potential biases that may disproportionately impact disadvantaged groups, as demonstrated by data from *r/Blackpeopletwitter* and YouTube comments related to the LGBTQ+ community. From Table 3, the 2019 data shows that *r/Blackpeopletwitter* comments are flagged as AI-generated at rates as high as 72.92% by RADAR and 21.42% by Binocular, significantly exceeding the rates for *r/Democrat* (62.46%) and *r/Republicans* (40%). This trend suggests that the language or discourse patterns prevalent in Black community discussions are more likely to trigger false positives, potentially due to linguistic features or cultural expressions that these systems fail to recognize accurately.

Table 4 highlights a similar concern for LGBTQ+ discourse. In 2019, RADAR flagged LGBTQ+-related comments on CNN, MSNBC, and FOX at exceptionally high rates, such as 70.30%, 70.45%, and 69.27%, respectively. Even models with comparatively lower detection rates, such as Ghostbusters, still reported rates exceeding 28% for these comments. This over-detection persists in 2024, with RAIDAR showing an increase from 21.17% to 23.33% for CNN LGBTQ+ comments and RADAR maintaining elevated probabilities like 73.52% for FOX LGBTQ+ comments. These results suggest systemic biases against marginalized groups, as AGTD systems misclassify their discourse at disproportionately high rates.

The elevated false positive rates for Black and LGBTQ+ communities indicate a lack of cultural and linguistic inclusivity in AGTD model training and evaluation. Such biases may lead to further marginalization of these historically oppressed communities in digital spaces leading to algorithmic invisibility.

Towards Robust AGTD Systems

Finding I: Anticontent sampling brings about a modest improvement.

The results in Table 2 indicate that integrating anticontent sampling into AGTD models yields modest yet consistent improvements in detection performance across datasets. For 2019, RAIDAR Anticontent Sampled achieves lower detection rates, such as 11% ± 2.2% for CNN, 15.33% ± 2.17% for MSNBC, and 9.26% ± 2.69% for FOX, compared to 13.85% ± 2.34%, 16.9% ± 4.71%, and 14.8% ± 1.46%, respectively, for the standard RAIDAR model. This trend suggests that anticontent sampling helps mitigate high FPRs to a certain degree. Incorporating anticontent sampling also achieves the best performance surpassing the base SoTA methods in all the instances.

Finding II: Linguistic signal-based trainable models can be calibrated towards enhanced reliablility.

The results in Table 2 demonstrate that Intrinsic+RAIDAR model, leveraging mixed linguistic signals for classification, outperform both the standalone RAIDAR and intrinsic di-

Models	Year	CNN LGBTQ+	MSNBC LGBTQ+	FOX LGBTQ+
RAIDAR	2019	21.17% ± 3.12%	19.24% ± 2.9%	18.80% ± 2.78%
	2024	23.33% ± 3.18%	29.60% ± 3.96%	22.50% ± 3.05%
RADAR	2019	70.30% ± 3.54%	70.45% ± 3.6%	69.27% ± 3.42%
	2024	67.60% ± 3.48%	70.72% ± 3.62%	73.52% ± 3.49%
Binocular	2019	24.98% ± 3.10%	23.91% ± 2.96%	24.18% ± 3.02%
	2024	19.44% ± 2.88%	18.44% ± 2.79%	19.54% ± 2.91%
Ghostbusters	2019	28.17% ± 3.26%	29.70% ± 3.44%	26.83% ± 3.11%
	2024	28.17% ± 3.20%	28.15% ± 3.18%	27.03% ± 3.15%

Table 4: Probability of AI-generated-content detection on LGBTQ+-related YouTube comments for CNN, MSNBC, and FOX News in 2019 and 2024 (mean ± 95% CI).

Timeline	r/Democrat	r/Republicans	r/Blackpeopletwitter	CNN	MSNBC	FOX
2019	3.07 ± 0.95	3.09 ±0.67	2.96 ± 0.7	2.98 ± 0.59	3.01 ± 0.63	3.1 ± 1.63
2024	3 ± 0.62	3.02 ± 0.61	3 ± 0.71	3.07 ± 0.71	2.88 ± 1.16	2.85 ± 2.73

Table 5: The Persistent Homology Dimension (PHD) scores for subreddit and YouTube comments

mension models. For 2019 data, Intrinsic+RAIDAR consistently achieves lower detection probabilities compared to the baseline RAIDAR model, such as 12.65% ± 1.87% for CNN, 16.15% ± 3.1% for MSNBC, and 13.45% ± 1.85% for FOX, compared to 13.85% ± 2.34%, 16.9% ± 4.71%, and 14.8% ± 1.46%, respectively.

Intrinsic dimension models leverage global geometric signals such as PHD and MLE, that capture the broader structural characteristics of the text. RAIDAR, on the other hand, utilizes fine-grained linguistic signals like n-gram distance and Levenshtein distance of a human-written text and an AI-generated text for classification. By combining these complementary feature sets in a unified framework, Intrinsic+RAIDAR creates a more robust representation, effectively reducing false positives and enhancing model performance. This finding suggests that linguistic signal-based trainable methods can be better at adapting to different datasets, and we can further fine-tune them for more robust classification.

Discussions and Conclusion

Among the growing literature (Chaka 2024; Bellini et al. 2024; Wu et al. 2025; He et al. 2024) of auditing AGTD systems, our study marks one of the first large-scale audit of AGTD systems on social web data. Our findings reveal significant limitations in SoTA AGTD systems, particularly their high FPR and systemic biases against specific linguistic and cultural groups. Models like RADAR consistently overestimate the probabilities of a text being AI-generated, especially for liberal discourse and discussions related to marginalized communities, such as Black and LGBTQ+ groups. The other SoTA models also show a similar trend to various degrees. These biases pose a risk of further marginalizing historically oppressed communities in digital spaces. This further solidifies the concerns established in the prior literature about how African-American English is more prone to be negatively attributed by AI systems (Koenecke et al. 2020; Hofmann et al. 2024). In the era where AI is incorporated in almost anything and everything in a maddening rush, our audits suggest a few words of caution for such rapid real-world deployment of AGTD systems in the web content moderation pipeline highlighting the need for rigorous human evaluation and precise operational guidelines while using these tools.

Our experiments also demonstrate that integrating anticontent sampling and leveraging mixed linguistic signal-based models can enhance AGTD performance. Anticontent sampling modestly reduces false positives by exposing models to a broader diversity of non-AI-generated text, while the Intrinsic+RAIDAR framework, combining fine-grained linguistic and global geometric features, achieves consistently better results.

Limitations

Our audit considers human-generated text content from the social web and reveals that even in such informal settings, AGTD systems produce a high number of false positives. Similar studies can be conducted on human-generated texts in more formal settings (e.g., scientific reviews, journalistic articles, political speeches etc.).

Extant literature points to the widening gaps between resource-rich and resource-poor languages (Ahuja et al. 2023). Our audit is limited to AI-generated English text detection systems. We hope this work will open the gates for similar audits in non-English languages.

Our audit considers liberal and conservative discourse in broad brush strokes. Given the strong political dissonance in the social web and frequent word of wars between political camps (Garimella and Weber 2017; Dutta et al. 2019; KhudaBukhsh et al. 2021; Wu and Resnick 2021), it is reasonable to assume that our liberal and conservative subplatforms are essentially liberal-majority and conservative-majority subplatforms. Morever, a user can be fiscally conservative and socially liberal. Our current research does not consider such nuanced political positions.

Finally, our current work focuses primarily on auditing AGTD models from the perspective of high false positive rates and their potential socio-technical consequences

like algorithmic invisibility for certain communities. While we do evaluate models under paraphrasing-based techniques (including multi-pass rewriting using LLMs), we acknowledge that more diverse adversarial attacks such as translation-based perturbations, malicious paraphrasing, and collage attacks are not covered in this version. Incorporating such attacks—especially those designed to evade detection while preserving semantic intent—would be a valuable extension to stress-test the robustness of AGTD systems.

Ethical Statement

We consider datasets procured through publicly available APIs and well-known AGTD models found in the literature. Moreover, we consider aggregate analyses without focusing on individual users. We thus do not see any major ethical concern. That said, any research on content filter can be used for malicious purposes. That said, our findings reveal that SOTA AGTD models often fail to account for various linguistic features, especially related to diverse ethnic groups. Our results demonstrate the models are more likely to disproportionately classify the African-American and LGBTQ+ related text as AI generated with an exception of `Binocular` and `Ghostbusters` models. The potential misuse of these AGTD models raises significant concerns as they could be used to unjustly suppress the voices of marginalized communities.

References

Abassy, M.; Elozeiri, K.; Aziz, A.; Ta, M. N.; Tomar, R. V.; Adhikari, B.; Ahmed, S. E. D.; Wang, Y.; Afzal, O. M.; Xie, Z.; et al. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection. *arXiv preprint arXiv:2408.04284*.

Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; Bali, K.; and Sitaram, S. 2023. MEGA: Multilingual Evaluation of Generative AI. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4232–4267. Singapore: Association for Computational Linguistics.

Arango, A.; Pérez, J.; and Poblete, B. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

Bellini, V.; Semeraro, F.; Montomoli, J.; Cascella, M.; and Bignami, E. 2024. Between human and AI: assessing the reliability of AI text detection tools. *Current Medical Research and Opinion*, 40(3): 353–358.

Bertaglia, T.; Heisig, L.; Kaushal, R.; and Iamnitchi, A. 2024. InstaSynth: Opportunities and Challenges in Generating Synthetic Instagram Data with ChatGPT for Sponsored Content Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 139–151.

Bhattacharjee, A.; Kumarage, T.; Moraffah, R.; and Liu, H. 2023. Conda: Contrastive domain adaptation for ai-generated text detection. *arXiv preprint arXiv:2309.03992*.

Bozell, L. B. 2004. *Weapons of mass distortion: The coming meltdown of the liberal media*. National Review.

Center, P. R. 2018. The news that bots share on Twitter tends not to focus on politics. https://www.pewresearch.org/short-reads/2018/06/21/the-news-that-bots-share-on-twitter-tends-not-to-focus-on-politics/.

Chaka, C. 2024. Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, 7(1).

Chen, H.; Takamura, H.; and Nakayama, H. 2021. SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1483–1492. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H. P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder, P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code.

Cresci, S. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10): 72–83.

Ding, X.; Horning, M.; and Rho, E. H. 2023. Same Words, Different Meanings: Semantic Polarization in Broadcast Media Language Forecasts Polarity in Online Public Discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1): 161–172.

Dutta, S.; Das, D.; Kaur, G.; Mongia, S.; Mukherjee, A.; and Chakraborty, T. 2019. Into the Battlefield: Quantifying and Modeling Intra-community Conflicts in Online Discussion. In Zhu, W.; Tao, D.; Cheng, X.; Cui, P.; Rundensteiner, E. A.; Carmel, D.; He, Q.; and Yu, J. X., eds., *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, 1271–1280. ACM.

Dutta, S.; Li, B.; Nagin, D. S.; and KhudaBukhsh, A. R. 2022. A Murder and Protests, the Capitol Riot, and the Chauvin Trial: Estimating Disparate News Media Stance. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5059–5065. AI for Good.

Elkhatat, A. M.; Elsaid, K.; and Almeer, S. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1): 17.

Farahmand, A. M.; Szepesvári, C.; and Audibert, J.-Y. 2007. Manifold-adaptive dimension estimation. In *Proceedings*

of the 24th international conference on Machine learning, 265–272.

Ferrara, E. 2023. Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*.

Garimella, V. R. K.; and Weber, I. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and social media*, volume 11, 528–531.

Gil de Zúñiga, H.; Correa, T.; and Valenzuela, S. 2012. Selective exposure to cable news and immigration in the US: The relationship between FOX News, CNN, and attitudes toward Mexican immigrants. *Journal of Broadcasting & Electronic Media*, 56(4): 597–615.

Goldmacher, S. 2024. Trump Falsely Claims That the Crowds Seen at Harris Rallies Are Fake. New York Times.

Grattafiori, A.; and team, M. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hans, A.; Schwarzschild, A.; Cherepanova, V.; Kazemi, H.; Saha, A.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

He, X.; Shen, X.; Chen, Z.; Backes, M.; and Zhang, Y. 2024. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2251–2265.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.

Hu, X.; Chen, P.-Y.; and Ho, T.-Y. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36: 15077–15095.

Huszár, F.; Ktena, S. I.; O'Brien, C.; Belli, L.; Schlaikjer, A.; and Hardt, M. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1): e2025334119.

Hyun, K. D.; and Moon, S. J. 2016. Agenda setting in the partisan TV news context: Attribute agenda setting and polarized evaluation of presidential candidates among viewers of NBC, CNN, and Fox News. *Journalism & Mass Communication Quarterly*, 93(3): 509–529.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2021. We Don't Speak the Same Language: Interpreting Polarization through Machine Translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 14893–14901. AAAI Press.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2022. Fringe News Networks: Dynamics of US News Viewership following the 2020 Presidential Election. In *WebSci '22: 14th ACM Web Science Conference 2022*, 269–278. ACM.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14): 7684–7689.

Kosseff, J. 2019. *The twenty-six words that created the Internet*. Cornell University Press.

Levina, E.; and Bickel, P. 2004. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17.

Li, Y.; Liu, Z.; Zhao, J.; Ren, L.; Li, F.; Luo, J.; and Luo, B. 2024. The Adversarial AI-Art: Understanding, Generation, Detection, and Benchmarking. In *European Symposium on Research in Computer Security*, 311–331. Springer.

Liu, Z. 2006. Reuter-50-50. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DS42.

Mao, C.; Vondrick, C.; Wang, H.; and Yang, J. 2024. Raidar: geneRative AI Detection viA Rewriting. arXiv:2401.12970.

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, L. H. X.; and Carley, K. M. 2023. Botbuster: Multi-platform bot detection using a mixture of experts. In *Proceedings of the international AAAI conference on web and social media*, volume 17, 686–697.

Proctor, J. 2024. B.C. lawyer reprimanded for citing fake cases invented by ChatGPT. CBC News.

Ramu, D.; Jain, R.; and Jain, A. 2023. Generation Z's Ability to Discriminate Between AI-generated and Human-Authored Text on Discord. *arXiv preprint arXiv:2401.04120*.

Russo Latona, G.; Horta Ribeiro, M.; Davidson, T. R.; Veselovsky, V.; and West, R. 2024. The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates. *arXiv e-prints*, arXiv–2405.

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*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. Seattle, United States: Association for Computational Linguistics.

Sarkar, R.; and KhudaBukhsh, A. R. 2021. Are chess discussions racist? an adversarial hate speech data set (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15881–15882.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schweinhart, B. 2021. Persistent homology and the upper box dimension. *Discrete & Computational Geometry*, 65(2): 331–364.

Settles, B. 2009. Active learning literature survey.

Skumanich, A.; and Kim, H. K. 2024. Modes of Analyzing Disinformation Narratives With AI/ML/Text Mining to Assist in Mitigating the Weaponization of Social Media. *arXiv preprint arXiv:2405.15987*.

Stanley, A. 2012. How MSNBC Became Fox's Liberal Evil Twin. Online; accessed 01-May-2024.

team, O. 2024. GPT-4o System Card. arXiv:2410.21276.

Tian, E.; and Cui, A. 2023. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods".

Tulchinskii, E.; Kuznetsov, K.; Kushnareva, L.; Cherniavskii, D.; Barannikov, S.; Piontkovskaya, I.; Nikolenko, S.; and Burnaev, E. 2023. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. arXiv:2306.04723.

University of Cambridge News Office. 2017. Celebrity Twitter accounts display bot-like behaviour. https://www.cam.ac.uk/research/news/celebrity-twitter-accounts-display-bot-like-behaviour. University of Cambridge.

Uyheng, J.; and Carley, K. M. 2021. Computational analysis of bot activity in the Asia-Pacific: A comparative study of four national elections. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 727–738.

Verma, V.; Fleisig, E.; Tomlin, N.; and Klein, D. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. arXiv:2305.15047.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Nature Communications*, 9(1): 1–9.

Weerasooriya, T.; Dutta, S.; Ranasinghe, T.; Zampieri, M.; Homan, C.; and KhudaBukhsh, A. 2023a. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11648–11668. Association for Computational Linguistics.

Weerasooriya, T. C.; Dutta, S.; Ranasinghe, T.; Zamperi, M.; Homan, C. M.; and KhudaBukhsh, A. R. 2023b. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *EMNLP 2023*, 11648–11668.

Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Chao, L. S.; and Wong, D. F. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 1–65.

Wu, S.; and Resnick, P. 2021. Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives. In *ICWSM 2021*, 15: 808–819.

Yoo, C. H.; and KhudaBukhsh, A. R. 2023. Auditing and Robustifying COVID-19 Misinformation Datasets via Anticontent Sampling. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 15260–15268. AAAI Press.

Zakrzewski, C.; and Lima-Strong, C. 2023. GOP lawmakers allege Big Tech conspiracy, even as ex-Twitter employees rebut them. Washington Post.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Paper Checklist

1. 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
2. Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
3. Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
5. Did you describe the limitations of your work? Yes
6. Did you discuss any potential negative societal impacts of your work? Yes
7. Did you discuss any potential misuse of your work? Yes
8. 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
9. Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
10. 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
11. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in Github
12. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, in Github
13. 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, in Github
14. Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
15. Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
16. If your work uses existing assets, did you cite the creators? Yes
17. Did you mention the license of the assets? Yes

18. Did you include any new assets in the supplemental material or as a URL? Yes
19. Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
20. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
21. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes
22. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset Yes, in Github

Appendices

A Misclassified examples

Here we present a broad analysis of false positives through various linguistic lenses. We find that general features—such as context length, syntactic complexity, and token diversity—often do not differ substantially between correctly and incorrectly classified examples. As we generate the AI-rewriting based on the original text, these basic features do not deviate largely from the original human-written text. Furthermore, advanced model-specific metrics such as Levenshtein distance and unigram distributional differences (used in models like RAIDAR and Ghostbusters) (Figure 1, 2, 3, 4) or Persistent Homology Dimension (Table 5, 6, 7, 8, 9, 10) show wide variability and sometimes close proximity between misclassified human texts and actual AI-generated samples. We provide illustrative plots in this section highlighting the feature overlap across classes, as well as a sample table (Table 6) containing representative examples of false positives and false negatives. These examples demonstrate the inconsistencies and vulnerabilities of current detection systems when confronted with nuanced or contextually rich human writing.

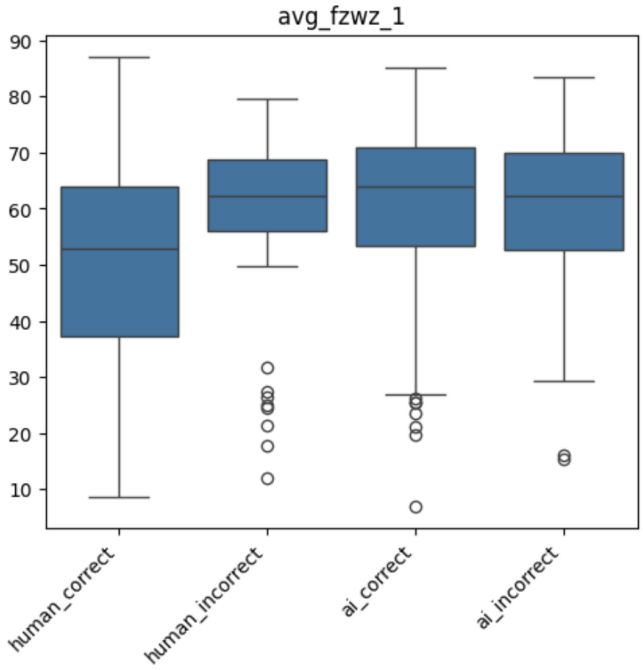

Figure 1: The fuzzywuzzy feature scores for the RAIDAR model in case of TP, TN, FP and FN

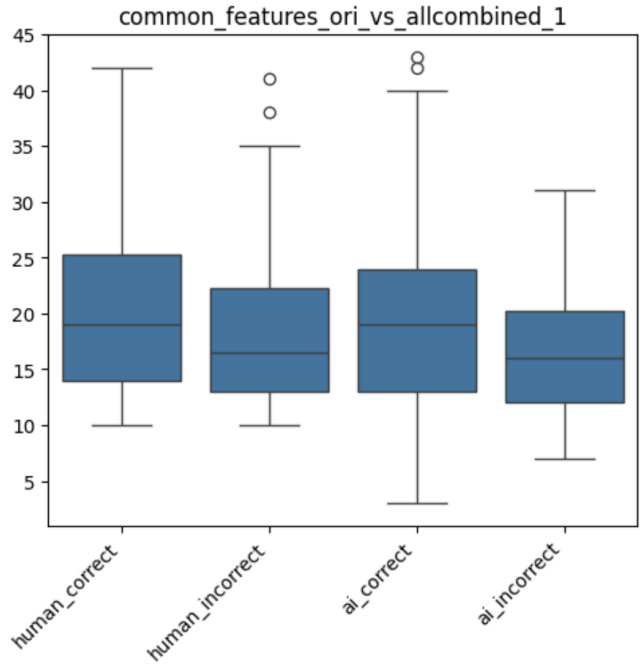

Figure 2: The combined feature scores for the RAIDAR model in case of TP, TN, FP and FN

Text	Original Label	Prediction
can you imagine if the main stream would have reported on Ron Paul's even huger crowds? no neither can I	Human (0)	AI (1)
It's amusing to see people suddenly labeling CNN as "fake news" only after Trump popularized the term. It's a reminder of how easily influenced and gullible people can be.	AI (1)	Human (0)
SOON WILL BE IN 'MOP BOSS MODE.' WILL NEED TO MOP UP ALL THE MESS HE CREATED IN CRIMES AND BRIBES!!!	Human (0)	AI (1)
Disappointed in the USA's government and leader, cancelled trip and boycotting USA and Trump products.	AI (1)	Human (0)
Thank you Mr. Reppenhagen for speaking out with intelligence, truth and compassion about the horrors of war, unwrapped from pretty patriotic propaganda. Godspeed.	Human (0)	AI (1)

Table 6: Examples of misclassifications by AGTD systems, including both false positives (Human → AI) and false negatives (AI → Human).

Paraphraser	Text	r/Democrat	r/Republicans	r/Blackpeopletwitter	CNN	MSNBC	FOX
ChatGPTT5	Comments	3.07 ± 0.95	3.09 ±0.67	2.96 ± 0.7	2.98 ± 0.59	3.01 ± 0.63	3.1 ± 1.63
	Comment paraphrased once	3.43 ± 19.45	2.54 ± 5.46	2.51 ± 15.99	3.11 ± 5.78	5.05 ± 55.53	3.57 ± 23.22
	Comment paraphrased twice	2.31 ± 5.31	4.02 ± 20.96	7.36 ± 97.96	4.96 ± 51.11	2.47 ± 10.88	3.39 ± 27.13
Pegasus	Comments	3.07 ± 0.6	3.01 ± 2.17	3.03 ± 0.7	3.06 ± 1.08	3 ± 0.66	3 ± 0.67
	Comment paraphrased once	1.71 ± 1.93	1.51 ± 2.9	2.57 ± 13.51	1.17 ± 6.67	1.77 ± 4.32	1.47 ± 9.09
	Comment paraphrased twice	1.65 ± 3.06	1.66 ± 2.92	0.75 ± 18.8	1.08 ± 7.79	0.99 ± 14.11	1.31 ± 3.99
Llama	Comments	3.03 ± 0.61	3.02 ± 0.61	3 ± 0.94	2.92 ± 0.58	3.02 ± 1.33	2.96 ± 0.7
	Comment paraphrased once	1.51 ± 15.11	2.24 ± 11.05	3.29 ± 9.41	4.9 ± 44.4	1.53 ± 65.36	1.9 ± 27.87
	Comment paraphrased twice	2.4 ± 7.61	3.06 ± 11.6	3.1 ± 0.48	0.55 ± 41.13	2.33 ± 3.17	2.05 ± 9.98

Table 7: The Persistent Homology Dimension (PHD) scores for 500 comments from the three subreddit and three youtube news channel in the year 2019 across paraphrasers

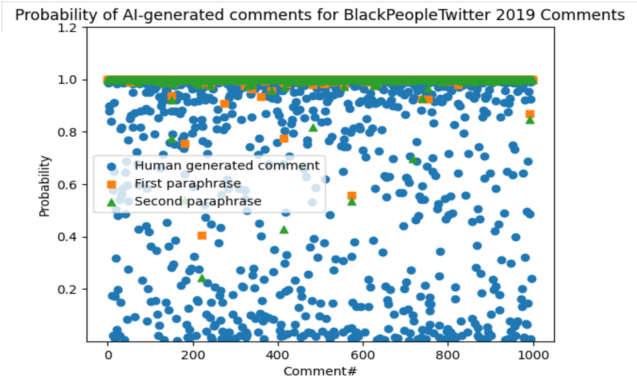

Figure 3: The probability of the comment being AI generated using RADAR for black people twitter subreddit for year 2019.

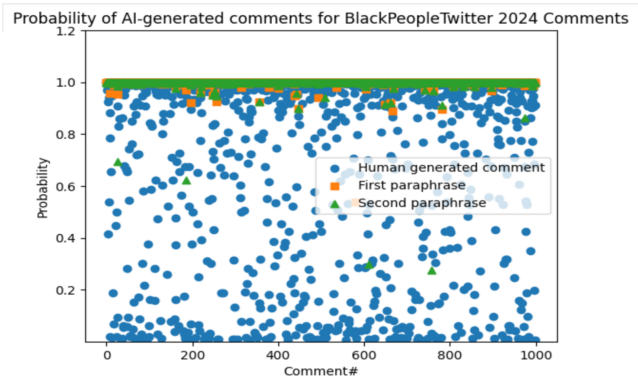

Figure 4: The probability of the comment being AI generated using RADAR for black people twitter subreddit for year 2024.

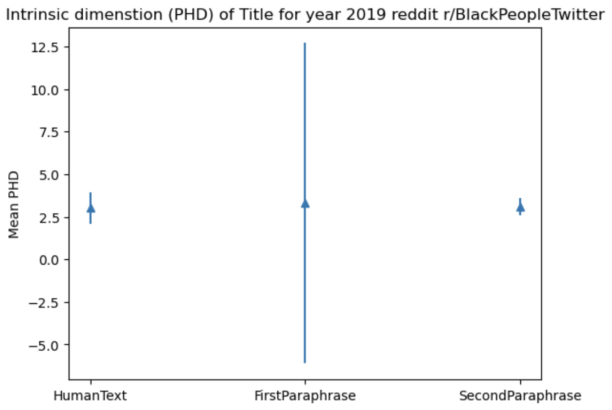

Figure 5: The Persistent Homology Dimension (PHD) scores for the black people twitter subreddit for year 2019

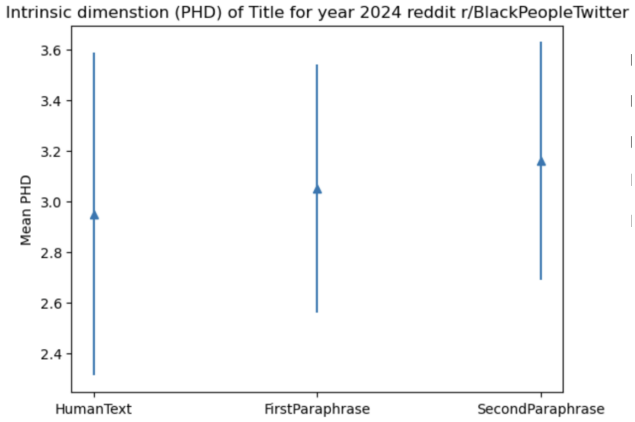

Figure 6: The Persistent Homology Dimension (PHD) scores for the black people twitter subreddit for year 2024

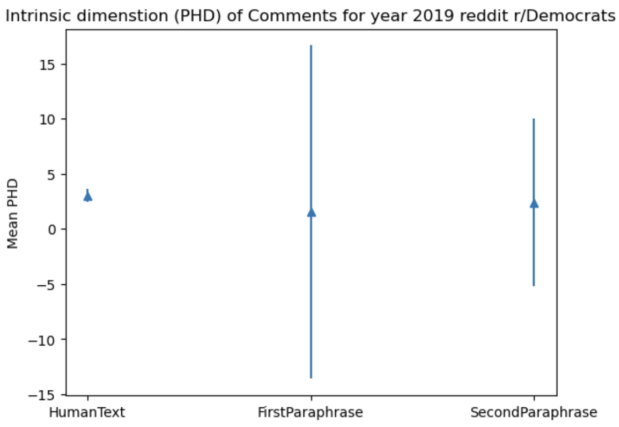

Figure 7: The Persistent Homology Dimension (PHD) scores for the Democrats subreddit for year 2019

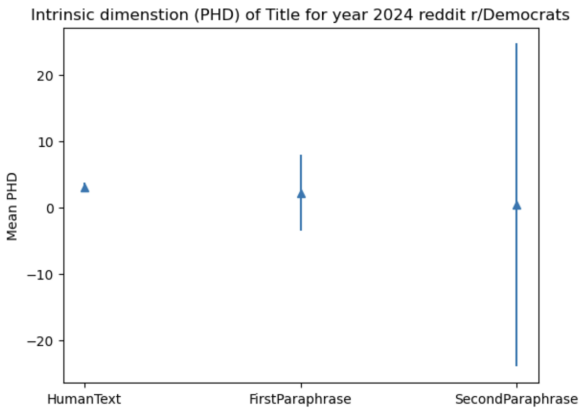

Figure 8: The Persistent Homology Dimension (PHD) scores for the Democrats subreddit for year 2024

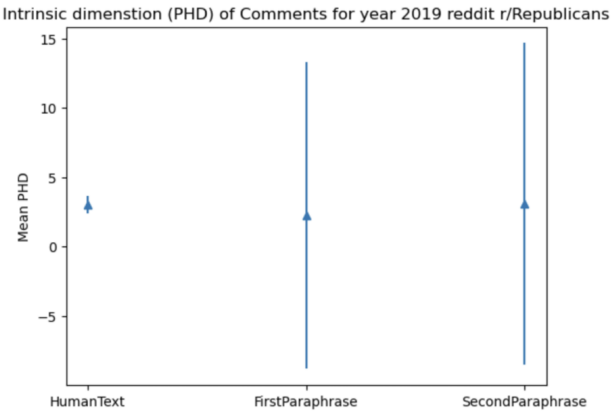

Figure 9: The Persistent Homology Dimension (PHD) scores for the Republicans subreddit for year 2019

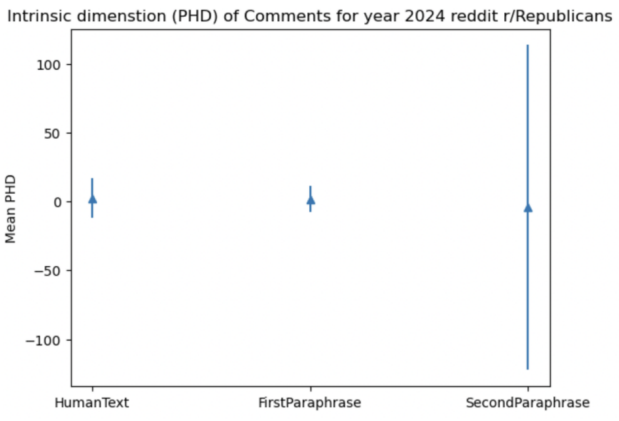

Figure 10: The Persistent Homology Dimension (PHD) scores for the Republicans subreddit for year 2024