HUMAN-IN-THE-LOOP DETECTION OF AI-GENERATED TEXT VIA GRAMMATICAL PATTERNS

Anonymous authors
Paper under double-blind review

ABSTRACT

The increasing proliferation of large language models (LLMs) has raised significant concerns about the detection of AI-written text. Ideally, the detection method should be accurate (in particular, it should not falsely accuse humans of using AI-generated text), and interpretable (it should provide a decision as to why the text was detected as either human or AI-generated). Existing methods tend to fall short of one or both of these requirements, and recent work has even shown that detection is impossible in the full generality. In this work, we focus on the problem of detecting AI-generated text in a domain where a training dataset of human-written samples is readily available. Our key insight is to learn interpretable grammatical patterns that are highly indicative of human or AI written text. The most useful of these patterns can then be given to humans as part of a human-in-the-loop approach. In our experimental evaluation, we show that the approach can effectively detect AI-written text in a variety of domains and generalize to different language models. Our results in a human trial show an improvement in the detection accuracy from 43% to 86%, demonstrating the effectiveness of the human-in-the-loop approach. We also show that the method is robust to different ways of prompting LLM to generate human-like patterns. Overall, our study demonstrates that AI text can be accurately and interpretably detected using a human-in-the-loop approach.

1 INTRODUCTION

Large language models (LLMs) are demonstrating exceptional capability across diverse domains, including logical reasoning, fluent language usage, and comprehensive factual awareness (Brown et al., 2020; Chowdhery et al., 2022). These capabilities bring new risks such as a tendency to hallucinate new information (Bang et al., 2023), introduce biases (Liang et al., 2021a), violate privacy (Brown et al., 2022), and others. One of the most widely discussed consequences is the lack of ability to distinguish between human and AI written text. This is an important problem as these models can disseminate misinformation at a large scale which can threaten democracy and trust in institutions (Chee, 2023; Juršėnas et al., 2021; Azzimonti & Fernandes, 2023) and propagate biases (Ferrara, 2023; Liang et al., 2021b). This is not just a future concern, as we have already seen the use of AI to generate Amazon product reviews (Palmer, 2023) or write novels for magazines (Hern, 2023). Moreover, educational institutions are also expressing concerns that traditional approaches for measuring student performance are not so effective in the face of new technology.

This underlines the importance of having a reliable and accurate way of identifying text written by AI. However, existing research falls short of this goal, as the proposed methods either only work on smaller models (Mitchell et al., 2023), require integration of detector and text generation(Kirchenbauer et al., 2023), or generally have low accuracy. Importantly, all of these methods do not provide an explanation of why some text has been classified as AI, as the decision is based on probabilities computed using deep neural networks, which are highly non-interpretable. This can have important negative consequences, as low accuracy and non-interpretability mean that a large number of innocent people will be accused of submitting AI-written text while receiving no explanation for this decision.

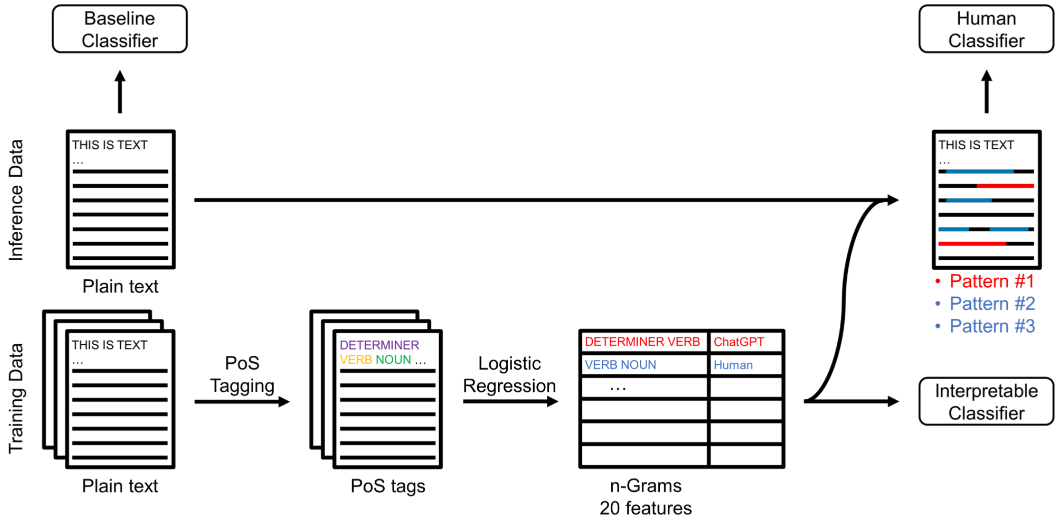


Figure 1: Overview of our approach. We perform PoS tagging on the training dataset and learn the most frequently occurring patterns in both human (blue) and LLM (red) written text. We then use these patterns to either train an interpretable classifier or give them directly to a human to assist in the detection process, thus creating a human-in-the-loop approach.

This work In this work, we propose a novel method to detect AI-written text using machine learning with a human-in-the-loop approach, which allows us to achieve high accuracy and interpretability. As the problem of detection has been shown to be intractable in the most general setting (Sadasivan et al., 2023), we focus on a case where there is an available dataset of human written text from a certain domain (e.g. arXiv texts or news articles). An overview of our method is shown in Figure 1. We first supplement the dataset with texts produced from a large language model by prompting it to generate texts from the same domain (e.g. abstracts of scientific articles with given titles). While we could train a highly accurate deep neural network classifier (e.g. by fine-tuning the BERT model) for detection, it would still not solve the problem of interpretability. Instead, our key insight is to learn PoS (part of speech) patterns whose occurrence is highly predictive of either human or AI-written text. To achieve this, we extract the number of occurrences of each PoS pattern in the text and then train the logistic regression classifier to discriminate between human and AI written text using these PoS features. Finally, we select the 20 patterns whose features have the highest (indicative of human) or lowest (indicative of AI) weight.

In our experimental evaluation, we show that our approach is effective. When humans are assisted with our patterns, their accuracy increases from 40% (worse than random guessing) to around 86% (close to a non-interpretable ML classifier). We also experimented with changing the prompt to LLM so that it generates more human-like patterns and showed that our approach still performs well, demonstrating that it is quite robust. These results indicate that interpretable AI detection with a human-in-the-loop approach can lead to classification with high accuracy and interpretability.

Main contributions Our main contributions are:

- We propose a novel method for accurate and interpretable detection of AI-generated text using a human-in-the-loop approach.
- In our experimental evaluation we demonstrate that the method generalizes across different state-of-the-art LLMs and text domains and is robust against several evasion strategies.
- We demonstrate the practical utility of our method through a human trial where results indicate that we can enable non-experts to identify machine-generated texts.

2 Related work

Several approaches have emerged to address societal concerns about the increasing difficulty of identifying LLM-generated texts. Current techniques differ in several key aspects. One pertains to output manipulation, where detection is simplified by modifying LLMs to embed characteristic signals in the produced text (Mitchell et al., 2023). Another important distinction involves accessing the model's probability distribution: white-box methods require this access, whereas black-box methods do not (Tang et al., 2023). Moreover, approaches also bifurcate into zero-shot techniques and those reliant on training sets for effective generalization to new contexts. Additionally, detection methods vary in terms of interpretability. These attributes reveal the main strengths and limitations of existing techniques.

Learned classifiers have been trained to distinguish human-authored and LLM-generated texts. In particular, BERT-based architectures like RoBERTa and DistilBERT (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2020) have been fine-tuned to accurately identify scientific abstracts produced by ChatGPT (Guo et al., 2023; Mitrović et al., 2023; Wang et al., 2023b; Yang et al., 2023; Yu et al., 2023; Theocharopoulos et al., 2023). Various training strategies such as contrastive or adversarial learning have also been successfully applied to increase performance (Bhattacharjee et al., 2023; bing Hu et al., 2023; Koike et al., 2023). Moreover, to foster human-AI interaction, post-hoc explanation methods such as SHAP and Polish-Ratio have been studied (Lundberg & Lee, 2017; Mitrović et al., 2023; Yang et al., 2023). Others have suggested reducing model complexity, for instance through gradient boosting tree classifiers reliant on linguistic features (Desaire et al., 2023).

Zero-short methods do not require training data, in contrast to learned classifiers. Recently, statistical tests have been developed to detect texts produced by a specific LLM; often using pivots based on token-wise conditional probabilities such as average token log probability, mean token rank, and predictive entropy (Gehrmann et al., 2019). Similarly, DetectGPT relies on a curvature-based criterion for zero-shot detection (Mitchell et al., 2023). The curvature is estimated through multiple perturbations of the original text, using a partial masking approach. Su et al. (2023) expounds upon similar ideas, utilizing log-rank information and specifically normalized log-rank perturbations.

Watermarking is the most prominent attempt of making detection less challenging (Kirchenbauer et al., 2023) and has recently garnered endorsement from major tech companies as well as the US government as a safeguard against AI-misuse (Bartz & Hu, 2023). This technique partitions output tokens into distinct green- and red-list categories, compelling the model to predominantly generate green-listed tokens. Applying statistical convergence results, one can ensure accurate identification of LLM-generated texts, accompanied by statistical guarantees on the false-positive rate. Moreover, recent works have improved the robustness, information content, and textual integrity of the watermark signal (Zhao et al., 2023; Wang et al., 2023a; Kuditipudi et al., 2023; Christ et al., 2023).

Human ability to recognize LLM-generated texts and methods to improve it has also been investigated. Several works have shown that without guidance, humans struggle to recognize LLM-generated texts (Gehrmann et al., 2019). However, by employing the visualization tool GLTR, peoples' accuracy when identifying GPT-2-generated texts increases from 54% to 72% (Gehrmann et al., 2019). Still, for current state-of-the-art LLMs, the performance is significantly lower (Uchendu et al., 2021). Moreover, mixed-initiative approaches aimed at enhancing experts' ability to recognize LLM-generated content have also shown great potential, with current work analyzing and visualizing syntactical, semantical, and pragmatical features (Weng et al., 2023). Also, collaboration between humans can increase their ability to collectively recognize LLM-generated texts, but without further guidance, the accuracy is still below 70% (Uchendu et al., 2023).

Moreover, as the quality of LLMs continues to improve and their text-generation capabilities approach that of humans, distinguishing between human and LLM-generated texts becomes increasingly challenging. In a recent study (Sadasivan et al., 2023), an upper limit on detection performance was established, which is based on the total variation distance between probability distributions of human authors ($\mathbb{P}_{\mathcal{H}}$) and the investigated LLM ($\mathbb{P}_{\mathcal{M}}$). Specifically, the area under the receiver operating characteristic curve (AUROC) of any classifier is upper bounded by:

$$\text{AUROC} \leq \frac{1}{2} + \text{TV}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\mathcal{H}}) - \frac{1}{2}\text{TV}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\mathcal{H}})^2 \tag{1}$$

Nevertheless, even though detecting machine-generated texts is a challenging problem, it carries immense societal importance.

3 BACKGROUND

In this section, we introduce the necessary background for our work.

Problem formalization AI Detection is a binary classification problem concerned with discerning human-written and LLM-generated texts. Let $\mathbb{P}_{\mathcal{H}}$ and $\mathbb{P}_{\mathcal{M}}$ be the distribution of texts authored by humans and the investigated LLM, respectively. Text-label pairs are then sampled $(t, y) \sim \mathbb{P}$ with $y \sim \text{Unif}(\{0, 1\})$, $\mathbb{P}[\cdot \mid y = 0] = \mathbb{P}_{\mathcal{H}}$ and $\mathbb{P}[\cdot \mid y = 1] = \mathbb{P}_{\mathcal{M}}$. The problem of AI detection is then to construct a classifier $f_\theta : \mathcal{T} \to \{0, 1\}$ that accurately predicts author y (human or AI) given text t where $\mathcal{T}$ is the set of all texts.

Grammatical patterns Grammar specifies which word sequences are contained in a language, and provides syntactic rules for how words can be combined into sentences. In most languages, these are formulated based on parts-of-speech (PoS) or word classes (Kroeger, 2005). Modern English employs nine fundamental word classes, as depicted in Table 1. Furthermore, the problem of assigning the appropriate PoS tag to each word is challenging due to polysemy and context dependency and has been extensively studied in computational linguistics. Modern approaches often employ machine learning and rely on a hidden Markov assumption (Toutanova et al., 2003; Toutanvoa & Manning, 2000; Zewdu & Yitagesu, 2022). Moreover, the resulting sequence of PoS tags contains all grammatical information from the original text. For illustration, we provide an example of the mapping between plain text and the sequence of PoS tags given in Table 1:

$$\text{This is a sentence} \Longrightarrow \text{DETERMINER VERB DETERMINER NOUN}$$

Feature Selection In many machine learning applications feature selection, aimed at identifying the most informative attributes while discarding irrelevant or redundant ones, is a crucial step. Reducing the number of active features can boost model performance and interpretability. In particular, a large feature set limits comprehensive model understanding (Miller, 1956). Moreover, achieving optimal feature selection is known to be NP-hard, even in linear scenarios (Welch, 1982). As a result, heuristic approaches are often employed, generally extracting or combining attributes based on various heuristically motivated criteria (Jolliffe, 2002; Hyvärinen & Oja, 2000; Peng et al., 2003).

In this work, we perform feature selection using Lasso (Tibshirani, 1996). Specifically, sparsity is induced by applying l_1-regularization to the optimization problem

$$\arg\min_{\theta \in \mathbb{R}^n} \mathcal{L}(\theta) + \alpha \|\theta\|_1 \tag{2}$$

where $\mathcal{L}$ is the loss function and α is the regularization parameter. Sparsity generally increases as $\alpha > 0$ grows, and it can be shown that for any feature there exists an upper bound on the value of α such that it is still contained in the active feature set (Henche, 2013; Tibshirani, 1996). Therefore, by adjusting α one can approximate the optimal active feature set.

4 OUR METHOD

In this section, we introduce key parts of our method.

Formalization Currently, humans struggle to recognize LLM-produced texts, often because they don't know which information is relevant. This issue is addressed by the following general framework, allowing for synergistic human-AI cooperation. Using a training set of text-label pairs, $\mathcal{D} = \{(t_i, y_i)\}_{i=1}^n$, we learn a function, h_ϕ, that highlights certain text passages accompanied by an explanation of their relevance regarding the origin of the text. Formally, we construct

$$h_\phi : \mathcal{T} \longrightarrow (\mathcal{T} \times \mathcal{T})^*, t \mapsto \{(p_i, e_i)\}_i \tag{3}$$

where $\{(p_i, e_i)\}_i$ are pairs of highlighted text passages and corresponding justifications. These are provided to a human decision-maker who contextualizes the information and weighs the arguments against each other, before making the final decision.

	Word Class	Definition
Word Classes	NOUN	A reference to a person, place or thing
	VERB	A reference to an action
	ADJECTIVE	A description of a noun's properties
	ADVERB	A description of a verb's properties
	PRONOUN	A substitute for a noun and any words which depend on it
	INTERJECTION	An expression that occurs as an utterance
	PREPOSITION	An description of a relationship in space or time
	CONJUNCTION	A link between different clauses of a sentence
	DETERMINER	A reference to a noun and any words which depend on it
Penn Treebank Extensions	DIGIT	A number or digit
	MODAL	An auxiliary verb expressing necessity, possibility, or permission
	EXISTENTIAL THERE	The word "there" when used to express existence
	FOREIGN WORD	A non-English word
	POSSESSIVE ENDING	The English genitive marker
	INFINITY MARKER	The word "to" when used to mark a verb in infinitive
	PARTICLE	An uninflected word that typically accompanies another word
	QUESTION WORD	A word expressing a question

Table 1: The nine modern English word classes as given in Blake (1988) and the further PoS-tags adopted from the Penn Treebank tag set (Marcus et al., 1993).

Extraction of grammatical patterns We instantiate our framework by the highlighting function h_ϕ that matches certain grammatical patterns, defined as n-grams of PoS tags. In our experimental setup, we adopt the PoS-tagger introduced in Toutanova et al. (2003), which uses the Penn Treebank tag set (Marcus et al., 1993). As the tag set is too extensive, we reduce complexity by consolidating tags into the categories outlined in Table 1. Moreover, we use $n \in \{1, \dots, 7\}$, resulting in a comprehensive set of 100.004 distinct grammatical features. As similar approaches have previously been successfully applied for authorship attribution (Sidorov et al., 2014), we anticipate these text passages to provide valuable insights into the texts' origin by revealing their grammatical structure.

Selecting predictive patterns Highlighting relevant text passages based on grammar requires understanding which grammatical patterns are informative. This is achieved by training a logistic regression model (Cramer, 2002) with l_1-regularization to induce sparsity, making the model reliant only on the most predictive grammatical patterns. Moreover, Miller's law (Miller, 1956) affirms the capacity of most people to retain maximally 9 items in short-term memory. This principle strongly implies that the number of extracted patterns should not significantly surpass this cognitive threshold if interpretability is a desired property. In our experimental setup, we find that relying on 20 grammatical patterns provides a good trade-off between interpretability and performance, which is achieved by adjusting the regularization parameter α from Equation (2).

Human-in-the-loop When assessing whether any text is LLM-generated, text passages matching the extracted grammatical patterns are highlighted and presented to a human who makes the final decision regarding the origin of the text. In order to associate each pattern with either human-written or LLM-generated texts, we refit the logistic regression model on the extracted patterns and assess the sign of the coefficient vector. The resulting, interpretable model can also be evaluated to understand how predictive the information provided to human users truly is. This approach guides decision-makers by directing their attention to the relevant parts of the text but remains interpretable as the final decision is based on specific, verifiable information that is extracted using our model.

5 Experiments

In this section, we empirically evaluate the efficacy of our method. First, we show that the extracted grammatical patterns are highly informative and can be applied to detect texts produced by the current state-of-the-art LLMs. Similarly, we demonstrate the robustness of our approach against several evasion strategies. Finally, through a human trial, we demonstrate that our patterns improve the ability of non-experts to recognize LLM-generated text, thus resulting in an interpretable and accurate classification procedure.

Datasets & metrics We employ several datasets in our setup: scientific abstracts from arXiv (Kaggle, 2023), social media comments from Reddit (Ethayarajh et al., 2022), CNN news articles (Hermann et al., 2015; See et al., 2017) and Wikipedia entries (Wikimedia Foundation, 2022). In particular, we first obtain human-written samples for each dataset considered. Then, using these as a reference, we query the LLM under consideration to produce similar texts, ensuring alignment in terms of subject matter, literary genres, origin, and length. The specific prompts are given in Appendix B.1. We measure the performance according to the AUROC score, and in Appendix F we additionally report accuracy.

5.1 Interpretable detection of LLM text

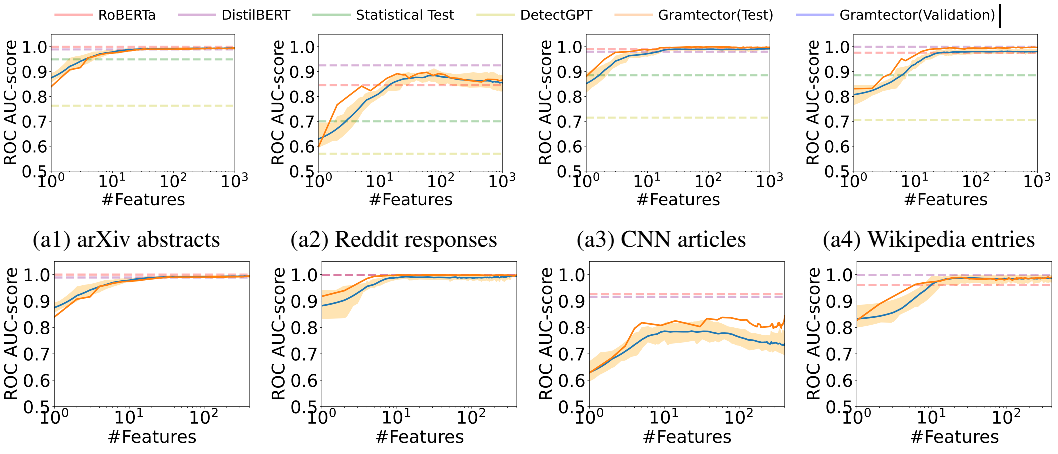
We first experiment with patterns as features for an interpretable classifier (no human-in-the-loop).

Detecting different text types We evaluate Gramtector's ability to identify different types of ChatGPT-generated texts. As seen in Figure 2, Gramtector performs on par with non-interpretable approaches. On all but one dataset, we attain an AUROC score close to 1 when at least 20 features are used; Gramtector even outperforms the RoBERTa and DistilBERT benchmarks on the Wikipedia and arXiv datasets, respectively. Even though we observe a performance decrease on the Reddit dataset, as we show in Appendix E, text length strongly influences performance and for longer Reddit responses Gramtector notably outperforms all non-interpretable benchmarks.

Detecting different LLMs We also study Gramtector's performance on texts produced by different state-of-the-art LLMs, in particular, ChatGPT, GPT-4, BARD, and LLAMA-2-70B. For each model, we construct a dataset of arXiv abstracts (Kaggle, 2023). Similar to our results on various textual domains, Gramtector's performance generalizes across LLMs. For ChatGPT, GPT-4, and LLAMA, we obtain AUROC scores close to 1, even outperforming some of the DNN-based methods. The outlier is the dataset containing abstracts produced by BARD where all models perform significantly worse, with Gramtector lagging behind the DNN-based benchmarks. It is possible that BARD-produced texts better resemble their human-written counterparts or that the model uses a more diverse language, making it harder to detect. Nonetheless, we can conclude that Gramtector generalizes to most practical scenarios; to almost all state-of-the-art LLMs and textual domains.

Robustness. To evaluate the robustness of Gramtector, we study several common evasion strategies. Specifically, we limit our ablation study to two realistic scenarios where a malicious actor tries to alter the linguistic expression of an LLM either by prompt engineering or paraphrasing sentences containing characteristics associated with the model. More details of the prompts are given in Appendix B.3, which resemble the attacks studied by Sadasivan et al. (2023). As baselines, we employ our framework instantiated with vocabulary or stylometric features, and we limit our investigation to separating human-written and ChatGPT-generated abstracts from arXiv.

In Table 2, we report the accuracy, AUROC score, and true positive ratio (TPR) of all models on the original dataset as well as the datasets containing adversarially constructed LLM samples. Although the model reliant on vocabulary features performs slightly better on the original dataset, Gramtector is significantly more robust in the adversarial setting. Its performance is unchanged under adversarial prompting and only marginally affected by paraphrasing. However, both other detection methods can be trivially evaded using these strategies; paraphrasing is especially effective, reducing both models' TPR from 98% to 4%. It therefore seems like grammatical sentence structure is an intrinsic characteristic of the LLM that is challenging to alter.



(a1) arXiv abstracts (a2) Reddit responses (a3) CNN articles (a4) Wikipedia entries

(b1) ChatGPT (b2) GPT-4 (b3) BARD (b4) LLAMA-2-70B

Figure 2: AUROC score dependent on the number of active features. The solid lines indicate the validation and test performance of Gramtector, while the stipulated refer to the benchmarks' utility on the test set. The shaded area marks the 10% and 90% quantiles of Gramtector's AUROC score when employing 10-fold cross-validation.

	Original Model			Adversarial Prompting			Adversarial Paraphrasing		
	Acc.	AUROC	TPR	Acc.	AUROC	TPR	Acc.	AUROC	TPR
Vocabulary	**0.970**	**0.997**	**0.980**	0.900	0.983	0.840	0.500	0.546	0.040
Stylometric	0.895	0.966	0.980	0.815	0.9111	0.820	0.425	0.588	0.040
Gramtector	0.955	0.984	0.950	**0.955**	**0.984**	**0.950**	**0.940**	**0.977**	**0.920**

Table 2: The efficacy of our framework when instantiated with vocabulary, stylometric, and grammatical (Gramtector) features. The models' accuracy, AUROC score, and true positive ratio (TPR) on the original as well as adversarially constructed datasets are shown.

5.2 Human Trial

We now describe a human trial where we used our patterns to assist human labelers in the detection.

Setup. We assess the efficacy of our approach by replicating plausible scenarios in which non-experts might encounter LLM-generated texts. Specifically, we study the research questions:

Q1 *Can insights extracted from Gramtector help non-experts recognize LLM-generated texts?*

Q2 *Which level of AI guidance is most suited to support human decision-making?*

We engage participants in an online study. Each individual is given 10 abstracts which they are asked to classify according to perceived origin. We reveal that 5 are human-authored and 5 are ChatGPT-generated. In the baseline study, this is all the information participants are given. In subsequent experiments, we provide increasingly easier access to the grammatical patterns extracted from Gramtector. Specifically, we employ a tiered approach with three levels:

1. **PoS tagging**: Participants are explained the grammatical patterns and we color each word with its corresponding PoS tag. However, individuals still need to manually search for any matches, which requires comprehending the provided information.

2. **Matched patterns**: All pattern matches are highlighted, but users still have to manually look up in the table whether each pattern is indicative of human or ChatGPT-generated text.

3. **Matched and classified patterns**: Pattern matches associated with human-written texts are highlighted in blue whereas ChatGPT-patterns are colored red. Interpreting this information is then similar to assessing a black-box probability score.

Figure 3 shows the setup for Level 1, PoS tagging, while more details about the setup for all levels are given in Appendix C. Moreover, we ask participants to justify their decisions, allowing us to gauge their level of engagement as expounded on in Appendix D. Specifically, we employ three categories: unengaged responses, engaged responses, and engaged responses employing the provided grammatical patterns. We thus address the trend of using to LLMs complete online surveys (Veselovsky et al., 2023); allowing us to separate hastily completed and thoughtful responses.

We review the developments of QCD multipole expansion and its applications to hadronic transitions and some radiative decays of heavy quarkonia. Theoretical predictions are compsred with updated experimental results.

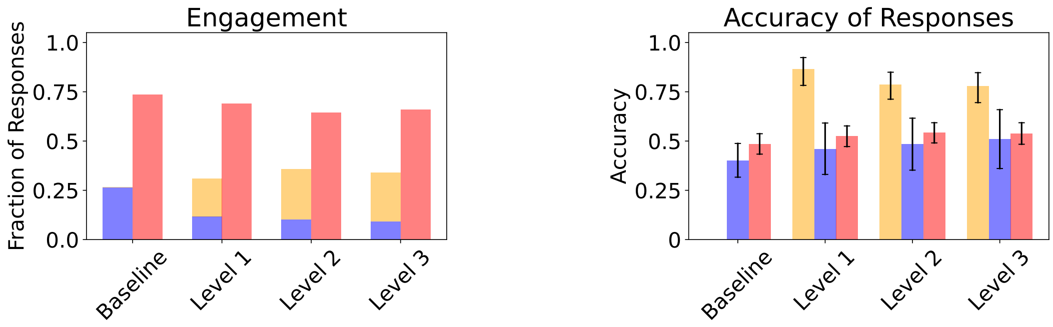
(a) Text to classify

ChatPGT	Human
DETERMINER NOUN VERB this paper presents, this paper investigates, this research contributes	DIGIT two, 2, one
PREPOSITION VERB by employing, to advancing, by considering	VERB ADVERB is not, do not, are not
PRONOUN NOUN our understanding, our findings, our approach	VERB VERB PREPOSITION is based on, is shown that, is related to
NOUN CONJUNCTION properties and, accuracy and, behavior and	PREPOSITION DETERMINER of the, in the, to the
DETERMINER NOUN VERB DETERMINER this paper presents a, this paper investigates the, this paper explores the	MODAL can, may, will

(b) Grammatical patterns from Gramtector

Figure 3: The information presented to participants in the human trial at Level 1, PoS tagging. Test takers are given a text (a) to classify according to origin: human-written or ChatGPT-generated. To inform their decision they are given access to the grammatical patterns extracted from Gramtector (b). Furthermore, each word in the text is accentuated with the corresponding PoS tag.

Human-in-the-loop with Gramtector patterns We observed in Figure 4a and Table 4 that among engaged participants, most actively employ the provided grammatical patterns, indicating that even non-experts find these insights useful. Notably, also at Level 1, engaged individuals make active use of the information we provide; requiring them to understand the patterns to find matches. Furthermore, as seen in Figure 4b, participants who actively employ grammatical characteristics to detect LLM-produced abstracts, significantly outperform the baseline; increasing the accuracy from 40% to 86% at Level 1. Also at subsequent levels, participants employing the insights extracted from Gramtector better detect LLM-generated texts compared to unengaged participants as well as the baseline, though slightly worse than at Level 1. When studying unengaged participants, their performance steadily increases with easier access to the grammatical characteristics, indicating that they implicitly make use of this information. From Table 3, we observe that the performance of the entire population increases together with the access to the grammatical patterns. Therefore, it indeed seems that the insights encapsulated in Gramtector can be transferred to humans, empowering them to better recognize LLM-generated texts.



(a) Distribution of engagement among responses for different levels of AI guidance.

(b) Accuracy of responses conditioned their engagement for different levels of AI guidance.

Figure 4: Trial results by engagement: unengaged (🟥), engaged (🟦), and engaged responses referencing the grammatical patterns (🟨). Error-bars: 5% Clopper & Pearson (1934) confidence interval.

	Baseline	Level 1	Level 2	Level3
$\hat{p}$	0.463	0.583	0.598	0.594
$\hat{p}_{\text{Low}}$	0.419	0.540	0.557	0.550
$\hat{p}_{\text{High}}$	0.507	0.625	0.638	0.637
True Positive Rate	0.573	0.600	0.610	0.659
False Positive Rate	0.537	0.417	0.402	0.406
p-value	N/A	0.685%	0.169%	0.338%
Correct Responses	236	309	353	303
Total Responses	510	530	590	510

Table 3: Estimated accuracy among all participants. $[\hat{p}_{\text{Low}}, \hat{p}_{\text{High}}]$ provides a 5% Clopper & Pearson (1934) confidence interval. The p-value assesses whether $\hat{p}$ is larger at the given level compared to the baseline. LLM-generated texts are considered positive samples for true and false positive rates.

AI Guidance. Furthermore, optimal performance among individuals employing the grammatical pattern is attained at Level 1, resulting in a paradoxical situation: increased access to information results in lowered performance. To understand this result, we assess how participants treat the provided information; do they apply the patterns in a black-box fashion, merely counting if they are mostly associated with human-written or LLM-generated texts, or do they contextualize the information? The specific procedure to detect black-box usage is explained in Appendix D. As seen in Table 4, stronger AI guidance correlates with more black-box usage. Therefore, it seems that if the model's predictions are all but directly presented to the user, individuals become overly reliant on AI guidance and do not comprehend and contextualize the provided information. This results in a noticeable performance decrease. Consequently, optimal results seem to be attained in a setting that requires cooperation between humans and AI.

	Baseline	Level 1	Level 2	Level 3
Black-Box References to Patterns	0	21	47	87
Engaged Responses Referencing Patterns	0	103	150	126
Engaged Responses	135	164	210	173
Total Responses	510	530	590	510
Pattern Utilization	N/A	62.8%	71.4 %	72.8%
Black-Box Pattern Utilization	N/A	20.4%	31.3 %	69.0%

Table 4: Engagement metrics across the various levels of AI guidance. Pattern utilization is the fraction of engaged responses that reference the provided grammatical patterns. Black-box pattern utilization is the fraction of responses that reference the provided grammatical patterns which do this in a black-box manner.

6 Conclusion

We introduced Gramtector, a framework for accurately and interpretably detecting LLM-generated texts. Our key insight is to learn grammatical patterns associated with texts produced by an LLM, which can subsequently be employed as identifiable markers of the model. Our experimental evaluation on datasets containing various types of text produced by leading-edge LLMs demonstrated that Gramtector performs on par with state-of-the-art non-interpretable detection methods. Moreover, the method appeared robust against several evasion strategies. A major advantage over prior work is Gramtector's inherent interpretability, allowing its insights to be transferred to humans. Through a human trial, we demonstrated that access to these insights significantly increases human decision-makers' ability to recognize LLM-produced texts, raising their accuracy from 40% to 86%. Our work thereby addresses several key concerns, contributing to the responsible deployment of LLMs.

Ethics Statement

The proliferation of LLMs raises several societal concerns due to the difficulty of discerning human-authored and LLM-produced texts. Our work aims to mitigate these issues by developing a robust framework for accurately and interpretably detecting LLM-generated texts. Although our framework does not completely remove the issue of false positives, its inherent interpretability allows individuals to understand and refute allegations of LLM use, unlike non-interpretable methods. We also demonstrated that the insights from our method could be transferred to humans, allowing them to better recognize LLM-generated texts. Ethical approval for our study was granted by an independent ethics commission (further details withheld due to double-blind review). Additionally, participants received fair compensation for their contributions, and stringent measures were in place to prevent exposure to harmful content. Overall, our work addresses important societal concerns regarding the widespread use of LLMs in a responsible manner.

References

Marina Azzimonti and Marcos Fernandes. Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76:102256, 2023. ISSN 0176-2680. doi: https://doi.org/10.1016/j.ejpoleco.2022.102256. URL https://www.sciencedirect.com/science/article/pii/S0176268022000623.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Diane Bartz and Krystal Hu. Openai, google, others pledge to watermark ai content for safety, white house says, Jul 2023. URL https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. Conda: Contrastive domain adaptation for ai-generated text detection. *ArXiv*, abs/2309.03992, 2023. URL https://api.semanticscholar.org/CorpusID:261660497.

Xiao bing Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *ArXiv*, abs/2307.03838, 2023. URL https://api.semanticscholar.org/CorpusID:259501842.

N. F. Blake. *Review of Word Classes*, pp. 14–28. Macmillan Education UK, London, 1988. ISBN 978-1-349-19006-5. doi: 10.1007/978-1-349-19006-5_2. URL https://doi.org/10.1007/978-1-349-19006-5_2.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2280–2292, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Foo Yun Chee. Ai generated content should be labelled, eu commissioner jourova says, Jun 2023. URL https://www.reuters.com/technology/ai-generated-content-should-be-labelled-eu-commissioner-jourova-says-2023-06-05/.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James

Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *ArXiv*, abs/2306.09194, 2023. URL https://api.semanticscholar.org/CorpusID:259092330.

C. J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934. URL https://api.semanticscholar.org/CorpusID:121902459.

J. S. Cramer. The origins of logistic regression. *Econometrics eJournal*, 2002. URL https://api.semanticscholar.org/CorpusID:129379279.

Department of Enterprise, Trade and Employment. National minimum wage increase on 1 january 2023, Dec 2022. URL https://www.gov.ie/en/publication/1786c-national-minimum-wage-increase-1-january/#.

Heather Desaire, Aleesa E. Chua, Madeline Isom, Romana Jarosova, and David C. Hua. Chatgpt or academic scientist? distinguishing authorship with over 99% accuracy using off-the-shelf machine learning tools. *ArXiv*, abs/2303.16352, 2023. URL https://api.semanticscholar.org/CorpusID:257804998.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Employment New Zealand, Apr 2023. URL https://www.employment.govt.nz/hours-and-wages/pay/minimum-wage/minimum-wage-rates/.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models, 2023.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text, 2019.

Government of Canada. Current and forthcoming general minimum wage rates in canada, Aug 2017. URL https://srv116.services.gc.ca/dimt-wid/sm-mw/rpt1.aspx?lang=eng.

Government of the United Kingdom. National minimum wage and national living wage rates, 2023. URL https://www.gov.uk/national-minimum-wage-rates.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

Diego Vidaurre Henche. Regularization for sparsity in statistical analysis and machine learning, 2013. URL https://api.semanticscholar.org/CorpusID:129668048.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pp. 1693–1701, 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

Alex Hern. This article is more than 6 months old sci-fi publisher clarkesworld halts pitches amid deluge of ai-generated stories. 2023. URL https://www.theguardian.com/technology/2023/feb/21/sci-fi-publisher-clarkesworld-halts-pitches-amid-deluge-of-ai-generated-stories.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13 4-5:411–30, 2000. URL https://api.semanticscholar.org/CorpusID:11959218.

Ian T. Jolliffe. Principal component analysis. In *International Encyclopedia of Statistical Science*, 2002. URL https://api.semanticscholar.org/CorpusID:27917863.

Juršėnas, Karlauskas, Ledinauskas, Maskeliūnas, and Ruseckas. The double-edged sword of ai: Enabler of disinformation, 2021.

Kaggle. arxiv dataset, 2023. URL https://www.kaggle.com/dsv/6293375.

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

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *ArXiv*, abs/2307.11729, 2023. URL https://api.semanticscholar.org/CorpusID:260091573.

Paul R. Kroeger. *Analyzing Grammar: An Introduction*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511801679.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *ArXiv*, abs/2307.15593, 2023. URL https://api.semanticscholar.org/CorpusID:260315804.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. URL https://api.semanticscholar.org/CorpusID:60827152.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pp. 6565–6576. PMLR, 2021a.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models, 2021b.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19:313–330, 1993. URL https://api.semanticscholar.org/CorpusID:252796.

George A. Miller. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63 2:81–97, 1956. URL https://api.semanticscholar.org/CorpusID:15654531.

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

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text, 2023.

Annie Palmer. Amazon is using generative a.i. to summarize product reviews. 2023. URL https://www.cnbc.com/2023/06/12/amazon-is-using-generative-ai-to-summarize-product-reviews.html.

Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962. URL https://api.semanticscholar.org/CorpusID:122932724.

Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2003. URL https://api.semanticscholar.org/CorpusID:206764015.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956. URL https://api.semanticscholar.org/CorpusID:16643156.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic n-grams as machine learning features for natural language processing. *Expert Syst. Appl.*, 41:853–860, 2014. URL https://api.semanticscholar.org/CorpusID:207738654.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *ArXiv*, abs/2306.05540, 2023. URL https://api.semanticscholar.org/CorpusID:259129463.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts, 2023.

The Office of the Fair Work Ombudsman. Minimum wages increase from 1 july 2023, 2023. URL https://www.fairwork.gov.au/newsroom/news/awr-2023#:~:text=From%201%20July%202023%2C%20the,or%20after%201%20July%202023.

Panagiotis C. Theocharopoulos, Panagiotis Anagnostou, Anastasia Tsoukala, Spiros V. Georgakopoulos, Sotiris K. Tasoulis, and Vassilis P. Plagianakos. Detection of fake generated scientific abstracts. *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 33–39, 2023. URL https://api.semanticscholar.org/CorpusID:258108316.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series b-methodological*, 58:267–288, 1996. URL https://api.semanticscholar.org/CorpusID:16162039.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *North American Chapter of the Association for Computational Linguistics*, 2003. URL https://api.semanticscholar.org/CorpusID:14835360.

Kristina Toutanvoa and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Conference on Empirical Methods in Natural Language Processing*, 2000. URL https://api.semanticscholar.org/CorpusID:10807721.

U. S. Department of Labour. State minimum wage laws, Jul 2023. URL https://www.dol.gov/agencies/whd/minimum-wage/state.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID:237589233.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Ting-Hao 'Kenneth' Huang, and Dongwon Lee. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? 2023. URL https://api.semanticscholar.org/CorpusID:257913864.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks, 2023.

Lean Wang, Wenkai Yang, Deli Chen, Haozhe Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards codable text watermarking for large language models. *ArXiv*, abs/2307.15992, 2023a. URL https://api.semanticscholar.org/CorpusID:260334887.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt, 2023b.

William J. Welch. Algorithmic complexity: three np- hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15:17–25, 1982. URL https://api.semanticscholar.org/CorpusID:122967135.

Luoxuan Weng, Minfeng Zhu, Kamkwai Wong, Siyi Liu, Jiashun Sun, Hang Zhu, Dongming Han, and Wei Chen. Towards an understanding and explanation for mixed-initiative artificial scientific text detection. *ArXiv*, abs/2304.05011, 2023. URL https://api.semanticscholar.org/CorpusID:258059731.

Wikimedia Foundation. Wikimedia downloads, 2022. URL https://dumps.wikimedia.org.

Lingyi Yang, Feng Jiang, and Haizhou Li. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text, 2023.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. Cheat: A large-scale dataset for detecting chatgpt-written abstracts, 2023.

Alebachew Zewdu and Betselot Yitagesu. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9, 01 2022. doi: 10.1186/s40537-022-00561-y.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu xiang Wang. Provable robust watermarking for ai-generated text. *ArXiv*, abs/2306.17439, 2023. URL https://api.semanticscholar.org/CorpusID:259308864.