GPT-who:
An Information Density-based Machine-Generated Text Detector

Anonymous ACL submission

Abstract

The *Uniform Information Density* (UID) principle posits that humans prefer to spread information evenly during language production. We examine if this UID principle can help capture differences between Large Language Models (LLMs)-generated and human-generated texts. We propose GPT-who, the first psycholinguistically-aware multi-class domain-agnostic statistical detector. This detector employs UID-based features to model the unique statistical signature of each LLM and human author for accurate authorship attribution. We evaluate our method using 4 large-scale benchmark datasets and find that GPT-who outperforms state-of-the-art detectors (both statistical- & non-statistical) such as GLTR, GPTZero, DetectGPT, OpenAI detector, and ZeroGPT by over 20% across domains. In addition to superior performance, it is computationally inexpensive and utilizes an interpretable representation of text articles. We find that GPT-who can distinguish texts generated by very sophisticated LLMs, even when the overlying text is indiscernible. UID-based measures for all datasets and code are available at https://anonymous.4open.science/r/gpt-who-03F8/.

1 Introduction

The recent ubiquity of Large Language Models (LLMs) has led to more assessments of their potential risks. These risks include its capability to generate misinformation (Zellers et al., 2019; Uchendu et al., 2020), memorized content (Carlini et al., 2021), plagiarized content (Lee et al., 2023), toxic speech (Deshpande et al., 2023), and hallucinated content (Ji et al., 2023; Shevlane et al., 2023). To mitigate these issues, researchers have proposed automatic and human-based approaches to distinguish LLM-generated texts (i.e., machine-generated) from human-written texts (Zellers et al., 2019; Pu et al., 2022; Uchendu et al., 2023; Mitchell et al., 2023).

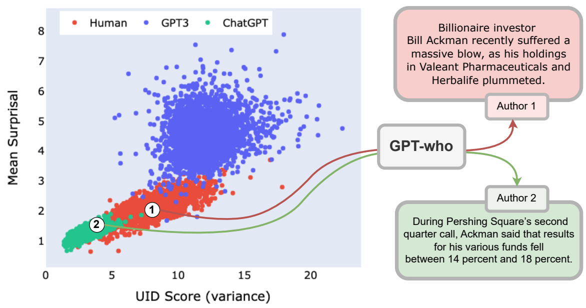


Figure 1: GPT-who leverages psycholinguistically motivated representations that capture authors' information signatures distinctly, even when the corresponding text is indiscernible.

Automatically detecting machine-generated texts occurs in two settings- *Turing Test* (TT) which is the binary detection of human vs. machine; and *Authorship Attribution* (AA) which is the multi-class detection of human vs. several machines (e.g., GPT-3.5 vs. LLaMA vs. Falcon) (Uchendu et al., 2021). While the TT problem is more rigorously studied, due to the wide usage of different LLMs, in the future, it will be imperative to build models for the AA tasks to determine which LLMs are more likely to be misused. This knowledge will be needed by policymakers when they inevitably institute laws to guard the usage of LLMs.

To that end, we propose GPT-who, the first psycholinguistically-aware supervised domain-agnostic task-independent multi-class statistical-based detector. GPT-who calculates interpretable Uniform Information Density (UID) based features from the statistical distribution of a piece of text and automatically learns the threshold (using Logistic Regression) between different authors.

To showcase the detection capabilities of GPT-who, we use 4 large LLM benchmark datasets: TuringBench (Uchendu et al., 2021), GPABenchmark (Liu et al., 2023b), ArguGPT (Liu et al., 2023a), and Deepfake Text in-the-wild (Li et al., 2023). We find that GPT-who remarkably outperforms state-of-the-art statistical detectors and is at par with task and domain-specific fine-tuned LMs for authorship attribution. This performative gain is consistent across benchmark datasets, types of LLMs, writing tasks, and domains.

It is even more remarkable that this performative gain is accompanied by two essential factors: First, GPT-who is computationally inexpensive as it eliminates the need for any LLM fine-tuning. It utilizes a freely available off-the-shelf LM to compute token probabilities, followed by logistic regression using a small set of carefully crafted and theoretically motivated UID features. Second, GPT-who provides a means to interpret and understand its prediction behaviors due to the rich feature space it learns from. UID-based features enable observable distinctions in the surprisal patterns of texts, which help in understanding GPT-who's decision-making on authorship (Figure 1).

We also analyze the UID distributions of different LLMs and human-generated texts across all datasets and find that humans distribute information more unevenly and diversely than models. In addition, UID features are reflective of differences in LLM architectures or families such that models that share architectures have similar UID distributions within but not outside their category. We find that UID-based features are a consistent predictor of authorship. Even when there aren't glaring differences between uniform and non-uniform text, the differences in UID distributions are easily detectable and a powerful predictor of authorship, since they successfully capture patterns that go beyond the lexical, semantic, or syntactic properties of text. Our work indicates that psycholinguistically-inspired tools can hold their ground in the age of LLMs and a simpler theoretically-motivated approach can outperform complex and expensive uninterpretable black-box approaches for machine text detection.

2 Related Work

2.1 Uniform Information Density (UID)

Shannon's Information Theory states that information exchange is optimized when information travels across the (noisy) channel at a uniform rate (Shannon, 1948). For language production, this uniform rate of information content is the basis of the UID hypothesis that posits that humans prefer to spread information evenly, avoiding sharp and sudden peaks and troughs in the amount of information conveyed per linguistic unit. The information content or **"surprisal"** of a word is inversely proportional to its probability in a given context. Less predictable words have more surprisal while highly predictable words convey lower information.

UID in human language production has been studied by measuring the amount of information content per linguistic unit (sentence length/number of words) or by studying any sudden changes in surprisal at the onset of a word or sentential element (Xu and Reitter, 2016; Jaeger and Levy, 2007). A rich body of work in psycholinguistics has led to the finding that, in language production, humans try to spread information content or surprisal evenly and maintain UID through their lexical, syntactic, phonological, and semantic choices (Frank and Jaeger, 2008; Xu and Reitter, 2018; Jaeger, 2010; Mahowald et al., 2013; Tily and Piantadosi, 2009).

2.2 Machine-Generated Text Detection

Large Language Models (LLMs) such as GPT-3.5, GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), Falcon (Penedo et al., 2023), have the capacity to generate human-like-quality texts, which can be easily construed as human-written (Sadasivan et al., 2023; Chakraborty et al., 2023; Zhao et al., 2023). However, while such LLMs are remarkable, it, therefore, makes them susceptible to malicious use. These include the generation of toxic and harmful content, like misinformation and terrorism recruitment (Shevlane et al., 2023; Zellers et al., 2019; Uchendu et al., 2021). Due to such potential for misuse, we must develop techniques to distinguish human-written texts from LLM-generated ones to mitigate these risks.

To mitigate this potential for misuse of LLMs, researchers have developed several types of automatic detectors. These techniques include supervised (Uchendu et al., 2021; Zellers et al., 2019; Uchendu et al., 2020; Zhong et al., 2020; Kushnareva et al., 2021; Liu et al., 2022) and unsupervised approaches (Gehrmann et al., 2019; Mitchell et al., 2023; Gallé et al., 2021; He et al., 2023; Su et al., 2023). These supervised approaches tend to be stylometric-, deep learning- and ensemble-based models while most unsupervised approaches are statistical-based detectors (Uchendu et al., 2023; Yang et al., 2023).

More recently, due to the increased ubiquity of LLMs, we need more interpretable, and less deep learning-based models. Deep learning models have been shown to be the most susceptible to adversarial perturbations than others (Pu et al., 2022). To that end, we propose the first supervised statistical-based technique, that calculates UID-based features of a given text and uses a classical machine learning model to automatically decide thresholds.

3 Our Proposal: GPT-who

We propose a psycholinguistically-motivated statistical-based machine-generated text detector GPT-who that uses a **GPT**-based language model to predict **who** the author of an article is. GPT-who works by exploiting a densely information-rich feature space motivated by the UID principle. UID-based representations are sensitive to intricate "fluctuations" as well as "smoothness" in the text. Specifically, operationalizations of UID are aimed at capturing the evenness or smoothness of the distribution of surprisal per linguistic unit (tokens, words), as stated by the UID principle. For example, in Figure 2, we show sequences of tokens that correspond to the highest and lowest UID score spans within an article. Here, the differences between the two segments of texts might not be obvious at the linguistic level to a reader, but when mapped to their surprisal distributions, the two segments have noticeably distinct surprisal spreads as can be seen by the peaks and troughs i.e. variance of token surprisals along the y-axis about the mean (dotted line). Most approximations of this notion of "smoothness" of information spread and UID, thus, formulate it as the variance of surprisal or as a measure of the difference of surprisals between consecutive linguistic units (Jain et al., 2018; Meister et al., 2020; Wei et al., 2021; Venkatraman et al., 2023).

In measuring the distribution of surprisal of tokens, UID-based features can capture and amplify subtle information distribution patterns that constitute distinct information profiles of authors. Using just an off-the-shelf language model to calculate UID-based features, GPT-who learns to predict authorship by means of a simple classifier using UID representations. In addition, as these features can be directly mapped to their linguistic token equivalents, GPT-who offers a more interpretable representation of its detection behavior, unlike current black-box statistical detectors, as illustrated in Figure 2. The use of a psycholinguistically motivated representation also enables us to better interpret the resulting representation space. It can capture

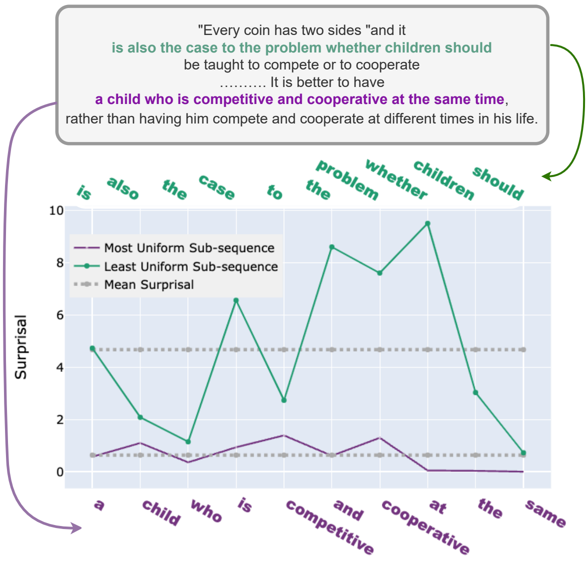


Figure 2: An example of UID span feature extraction that selects the most uniform and non-uniform segments from the token surprisal sequence. As can be seen in this example, two texts that read well can have very different underlying information density distributions in a given context. UID features capture these hidden statistical distinctions that are not apparent in their textual form.

surprisal distributions indicative of and commonly occurring in human-written or machine-generated text. GPT-who is one of the first text detectors that focus on informing a simple classifier with theoretically motivated and intuitive features, as it only requires a fixed-length UID-based representation of length 44 and learns to predict authorship based on just these features, without the need for the full text or any LM fine-tuning in the process (See GPT-who's complete pipeline in Figure 3).

3.1 UID-based features

We use the 3 most widely used measures of UID scores as defined in previous works (Jain et al., 2018; Meister et al., 2020; Wei et al., 2021; Venkatraman et al., 2023) as follows: We first obtain the conditional probability p of each token (y_t) in an article using a pre-trained LM (GPT2-XL). The surprisal (u) of a token y_t is,

$$u(y_t) = -\log(p(y|y < t)), \qquad (1)$$

for $t \geq 1$ where $y_0 = < BOS >$, and t = time step.

The lower the probability of a token, the higher its surprisal and vice-versa. Thus, surprisal indicates how unexpected a token is in a given context.

1. **Mean Surprisal** (μ) of an article (y) defined

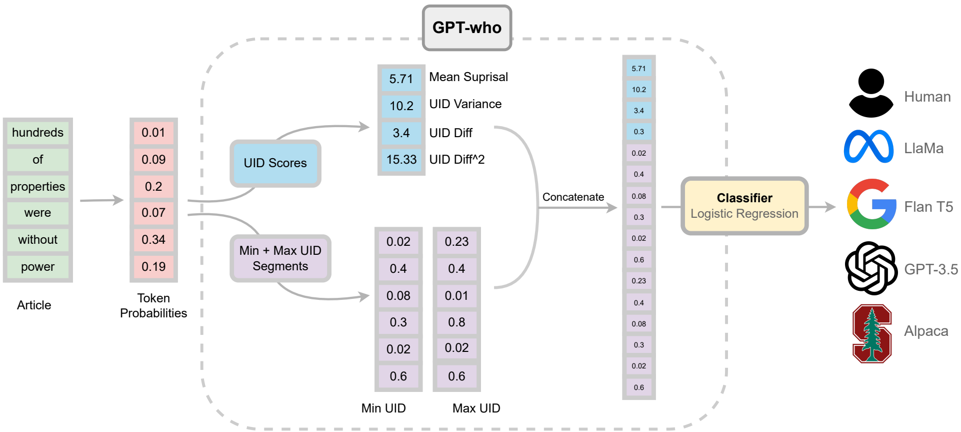


Figure 3: GPT-who uses token probabilities of articles to extract UID-based features. A classifier then learns to map UID features to different authors, and identify the author of a new unseen article.

as follow:

$$\mu(y) = \frac{1}{|y|} \sum_t (u(y_t)) \tag{2}$$

2. **UID** ($Variance$) score or **global** UID score of an article (y) is calculated as the normalized variance of the surprisal:

$$\text{UID}(y) = \frac{1}{|y|} \sum_t (u(y_t) - \mu)^2 \tag{3}$$

   From this formulation, a perfectly uniform article would have the same surprisal at every token and hence 0 UID (variance) score.

3. **UID** ($Difference$) score or **local** UID score of an article (y) is calculated as the average of the difference in surprisals of every two consecutive tokens $\mu(y_{t-1})$ and $\mu(y_t)$:

$$\text{UID}(y) = \frac{1}{|y|-1} \sum_{t=2}^{|y|} |\mu\left(y_t\right) - \mu\left(y_{t-1}\right)| \tag{4}$$

4. **UID** ($Difference^2$) score is defined as the average of the squared difference in surprisals of every two consecutive tokens $\mu(y_{t-1})$ and $\mu(y_t)$:

$$\text{UID}(y) = \frac{1}{|y|-1} \sum_{n=2}^{|y|} (\mu\left(y_t\right) - \mu\left(y_{t-1}\right))^2 \tag{5}$$

   From this formulation, both local measures of UID capture any sudden bursts of unevenness in how information is dispersed in consecutive tokens of the articles.

Maximum and minimum UID spans In addition to previously used approximations of UID, we also craft a new set of features using the most and least uniform segments of an article. Our intuition for this feature is to focus on the extremities of the UID distribution in an article, as the most and least uniform spans would be the most expressive and distinct sequences from a UID perspective. All other spans or segments in an article necessarily lie in between these two extremities. Thus taking account of these two spans would ensure coverage of the whole range of surprisal fluctuations within an article. Thus, for each article, we calculate UID (variance) scores for all spans of consecutive tokens of a fixed length using a sliding window approach. We tuned this window size and found that a window size of 20 tokens per span sufficiently represented an article's UID range. We also experimented with randomly drawn and re-ordered spans and found that random features did not contribute to task performance (see Table 1 for ablation study results). We use the surprisal values corresponding to the highest and lowest UID scoring span as additional features and obtain fixed length UID features of length 44 for each article.

4 Empirical Validation

We use Meister et al. (2021)'s implementation of UID-based scores[1] and use the publicly available off-the-shelf pre-trained GPT2-XL language model[2] to obtain conditional probabilities. For all our experiments, we calculate the UID features for the publically released train and test splits of

[1] https://github.com/rycolab/revisiting-uid/tree/main

[2] https://huggingface.co/gpt2-xl

Span Length (N)	Random UID spans	No Spans	Min + Max UID spans				
			N=4	N=10	N=15	N=20	N=30
GPT-1	0.75	0.76	0.99	0.99	0.98	**1.00**	0.99
GPT-2_small	0.62	0.64	0.75	0.82	**0.88**	**0.88**	0.85
GPT-2_medium	0.63	0.63	0.73	0.80	**0.88**	0.87	0.84
GPT-2_large	0.65	0.62	0.73	0.79	**0.88**	**0.88**	0.83
GPT-2_xl	0.65	0.61	0.72	0.80	0.88	**0.89**	0.85
GPT-2_PyTorch	0.55	0.64	0.83	0.84	**0.87**	0.85	0.86
GPT-3	0.63	0.69	0.71	0.73	0.77	**0.84**	0.74
GROVER_base	0.63	0.65	0.76	0.77	0.79	**0.81**	0.78
GROVER_large	0.59	0.60	0.71	0.71	0.73	**0.75**	0.72
GROVER_mega	0.55	0.56	0.67	0.67	0.68	**0.72**	0.67
CTRL	0.79	0.83	**0.99**	0.98	0.98	**0.99**	0.98
XLM	0.62	0.69	0.96	0.96	0.96	**0.99**	0.96
XLNET_base	0.62	0.71	0.95	0.97	0.98	0.98	**0.99**
XLNET_large	0.49	0.70	0.99	0.99	0.99	**1.00**	0.99
FAIR_wmt19	0.54	0.57	0.74	0.75	**0.78**	0.74	0.76
Fair_wmt20	0.62	0.63	0.72	0.75	0.88	**1.00**	0.89
TRANSFO_XL	0.70	0.70	0.79	0.80	0.83	0.79	**0.84**
PPLM_distil	0.57	0.62	0.92	0.91	0.93	**0.95**	0.93
PPLM_gpt2	0.54	0.58	0.88	0.88	**0.90**	0.89	0.88
TuringBench (Avg F1)	0.62	0.65	0.82	0.84	0.87	**0.88**	0.86
InTheWild (Avg F1)	0.72	0.75	0.79	0.83	0.86	**0.88**	0.87

Table 1: Max. & Min. UID spans ablation study: Setting a span length of N=20 tokens maximized performance across large-scale datasets (N>30 leads to subsequently lower and eventually consistent performance). It can be seen that our min/max features tremendously impact performance against randomly sampled or no span features at all.

all datasets. We train a logistic regression model[3] using these features on the train splits and report performance on the test splits. We replicate all the original evaluation settings and metrics for each of the datasets (except one setting from the ArguGPT (Liu et al., 2023a) dataset that required access to unreleased human evaluation data). We do this to be able to directly compare the performance of GPT-who with current state-of-the-art detection methods reported so far.

4.1 Datasets

To test the applicability of GPT-who across text detection tasks, we run all experiments across 4 large-scale and very recent datasets that span over 15 domains and 35 recent LMs.

TuringBench Benchmark (Uchendu et al., 2021) dataset is the largest multi-class authorship attribution dataset that contains over 168k news articles generated by 19 neural text generators using 10K prompts from CNN and the Washington Post.

[3] https://scikit-learn.org/stable/

GPABenchmark (Liu et al., 2023b) or GPT Corpus for Academia is a multi-domain (Computer Science (CS), Humanities and Social Sciences (HSS) and Physics (PHX)) academic articles dataset aimed at helping detection of LLM use or misuse in academic writing. It contains 150k human and 450k ChatGPT-generated articles for 3 task settings (completion, writing, and polishing).

ArguGPT (Liu et al., 2023a) is a prompt-balanced dataset of argumentative essays containing over 4k human-written essays and 4k articles generated by 7 recent LLMs (including many variants of ChatGPT) using prompts from English datasets such as TOEFL11 (Blanchard et al., 2013) and WECCL (Wen et al., 2005) datasets.

"InTheWild" Deepfake Text Detection in the Wild (Li et al., 2023) dataset is, to our knowledge, the largest text detection dataset consisting of over 447k human-written and machine-generated texts from 10 tasks such as story generation, news article writing, and academic writing. They use 27 recent LLMs such as GPT-3.5, FLAN-

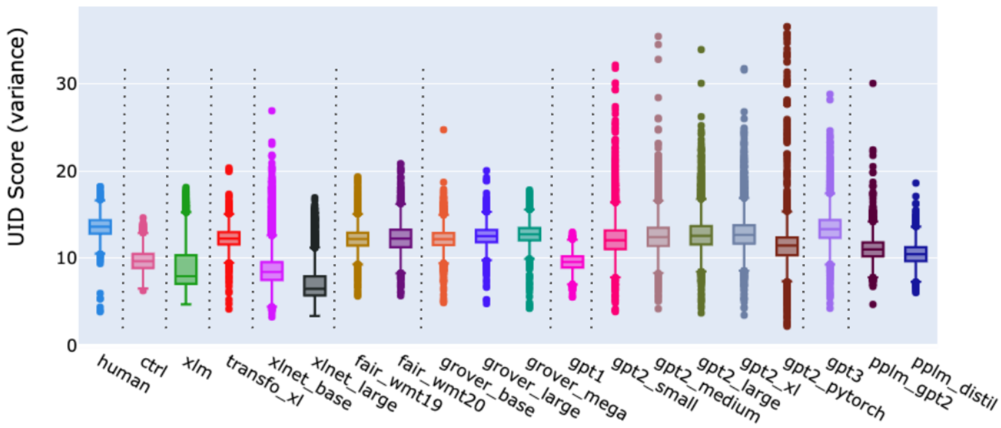


Figure 4: Distribution of UID Scores of 20 authors from the TuringBench dataset grouped (dotted line) by architecture type. LMs that share architectures tend to distribute UID scores similarly.

T5, and LLaMA. We refer to this dataset as the **"InTheWild"** dataset going forward for brevity.

4.2 Baselines & Detectors

We compare our proposed method against the following: DetectGPT [4] (Mitchell et al., 2023), GLTR[5] (Gehrmann et al., 2019), an open-source implementation[6] of GPTZero (Tian and Cui, 2023), ZeroGPT (zer, 2023), OpenAI's detector (Solaiman et al., 2019), Li et al. (2023)'s LongFormer-based detector[7] tuned for the InTheWild benchmark (we refer to this method as **"ITW"**), a stylometric detector[8] (Abbasi and Chen, 2008) and fine-tuned BERT[9] (Kenton and Toutanova, 2019). We are unable to report results for exhaustively all methods across all datasets due to inherent inapplicability in certain task settings. For example, most SOTA text detectors cannot be applied to the ArguGPT dataset as it only contains text written by multiple machines, while most text detectors are designed to differentiate between human-written and machine-generated texts. Beyond such limitations, we have utilized all applicable methods for 4 benchmark datasets.

4.3 UID Signatures of Authors

Given that humans tend to optimize UID, we study if different models spread surprisal in ways that are distinguishable from each other and human-written text and if we can observe unique UID signatures of different LM families. To this end, we plot the UID score distributions of different text generators across (see Figures 4, 5a, and 5b). We observe that, generally, the UID scores of human-written text have a higher mean and larger standard deviation than most machine-written text across writing task types, domains, and datasets. This implies that human-written text tends to be more non-uniform and diverse in comparison to machine-generated text. Hence, machines seem to be spreading information more evenly or smoothly than humans who are more likely to have fluctuations in their surprisal distributions. Going a step further, if we compare models to other models, we see that models that belong to the same LM family by architecture tend to follow similar UID distribution. For example, in Figure 4, the dotted lines separate LMs by their architecture type and it can be seen, for example, that all GPT-2 based models have similar UID distributions, all Grover-based models have similarities, but these groups are distinct from each other. This indicates that UID-based features can capture differences in text generated by not only humans and models but also one step further to capture differences between individual and multiple models and LM families. To our knowledge, this is the first large-scale UID-based analysis of recent machine and human-generated text across writing tasks and domains.

4.4 Machine Text Detection Performance

Overall, GPT-who outperforms other statistical-based detectors and is at par with transformers-based fine-tuned methods for 2 out of 4 benchmarks. For **GPABenchmark** (Table 2), across all

[4] https://github.com/eric-mitchell/detect-gpt
[5] https://github.com/HendrikStrobelt/detecting-fake-text
[6] https://github.com/BurhanUlTayyab/GPTZero
[7] https://github.com/yafuly/DeepfakeTextDetect
[8] https://github.com/shaoormunir/writeprints
[9] https://huggingface.co/docs/transformers/training

Task Type	Domain	GPTZero	ZeroGPT	OpenAI Detector	DetectGPT	BERT	ITW	GPT-who
Task 1	CS	0.30	0.67	0.81	0.58	**0.99**	0.98	**0.99**
	PHX	0.25	0.68	0.70	0.54	**0.99**	0.98	0.98
	HSS	0.72	0.92	0.63	0.57	**0.99**	0.96	0.98
Task 2	CS	0.17	0.25	0.64	0.16	**0.99**	0.81	0.84
	PHX	0.06	0.10	0.24	0.17	**0.96**	0.76	0.90
	HSS	0.44	0.62	0.27	0.20	**0.97**	0.29	0.80
Task 3	CS	0.02	0.03	0.06	0.03	**0.97**	0.38	0.63
	PHX	0.02	0.03	0.04	0.05	**0.97**	0.31	0.75
	HSS	0.20	0.25	0.06	0.06	**0.99**	0.08	0.62
Average F1		0.24	0.40	0.38	0.26	**0.98**	0.62	0.83

Table 2: Test Set Performance (F1 Scores) of different machine text detectors on the GPA Benchmark. Best performance are in bold, and second best underlined.

Human v.	GROVER	GTLR	GPTZero	DetectGPT	RoBERTa	BERT	ITW	Stylometry	GPT-who
GPT-1	0.58	0.47	0.47	0.51	0.98	0.95	0.92	0.99	**1.00**
GPT-2_small	0.57	0.51	0.51	0.51	0.71	0.75	0.47	0.75	**0.88**
GPT-2_medium	0.56	0.49	0.50	0.52	0.75	0.65	0.47	0.72	**0.87**
GPT-2_large	0.55	0.46	0.49	0.51	0.79	0.73	0.46	0.72	**0.88**
GPT-2_xl	0.55	0.45	0.51	0.51	0.78	0.79	0.45	0.73	**0.89**
GPT-2_PyTorch	0.57	0.72	0.50	0.52	0.84	**0.99**	0.47	0.83	0.85
GPT-3	0.57	0.35	0.47	0.52	0.52	0.79	0.48	0.72	**0.84**
GROVER_base	0.58	0.39	0.52	0.51	**0.99**	0.98	0.49	0.76	0.81
GROVER_large	0.54	0.41	0.47	0.52	**0.99**	0.98	0.52	0.71	0.75
GROVER_mega	0.51	0.42	0.42	0.51	0.94	**0.97**	0.53	0.68	0.72
CTRL	0.49	0.88	0.67	0.67	**1.00**	**1.00**	0.91	0.99	0.99
XLM	0.50	0.89	0.67	0.67	0.58	**1.00**	0.92	0.96	0.99
XLNET_base	0.58	0.75	0.51	0.67	0.79	**0.99**	0.84	0.95	0.98
XLNET_large	0.58	0.88	0.67	0.52	**1.00**	**1.00**	0.93	**1.00**	**1.00**
FAIR_wmt19	0.56	0.56	0.56	0.51	0.84	**0.93**	0.49	0.74	0.74
Fair_wmt20	0.58	0.49	0.50	0.51	0.45	0.47	0.47	0.73	**1.00**
TRANSFO_XL	0.58	0.35	0.49	0.52	0.96	**0.97**	0.81	0.79	0.79
PPLM_distil	0.59	0.64	0.52	0.67	0.90	0.88	0.51	0.92	**0.95**
PPLM_gpt2	0.58	0.68	0.51	0.51	**0.90**	0.89	0.49	0.88	0.89
Average F1	0.56	0.57	0.52	0.55	**0.88**	0.61	**0.88**	0.82	**0.88**

Table 3: Test Set Performance (F1 score) for TuringBench dataset. Overall, GPT-who outperforms both statistical and supervised detectors, and is at part with BERT.

Detection Setting	Testbed Type	GPTZero	GLTR	DetectGPT	BERT	ITW	GPT-who
In-distribution	Domain-specific Model-specific	0.65	0.94	0.92	**0.98**	0.97	0.93
	Cross-domains Model-specific	0.63	0.84	0.6	**0.98**	0.97	0.88
	Domain-specific Cross-models	0.57	0.8	0.57	0.49	**0.87**	0.86
	Cross-domains Cross-models	0.57	0.74	0.57	0.49	0.78	**0.86**
Out-of-distribution	Unseen Models	0.58	0.65	0.6	**0.84**	0.79	0.74
	Unseen Domains	0.57	0.72	0.57	0.68	**0.8**	0.77
	Average F1	0.60	0.78	0.64	0.74	**0.86**	0.84

Table 4: Test Set Performance (F1 score) for InTheWild dataset. ITW refers to the LongFormer-based detector trained by Li et al. (2023) specifically for this benchmark.

Author	Experts*	Stylometry	BERT	GPT-who
text-babbage-001	0.47	0.45	0.84	**0.85**
text-curie-001	0.47	0.45	0.83	**0.84**
text-davinci-003	0.66	0.59	**0.95**	0.77
gpt-3.5-turbo	0.63	0.69	**0.96**	0.84
gpt2-xl	0.37	0.49	**0.95**	0.91
Average F1	0.52	0.53	**0.91**	0.84

Table 5: Test Set Performance (F1 score) for ArguGPT dataset.* denotes results reported in Liu et al. (2023a).

task types and domains, GPT-who outperforms GPTZero, ZeroGPT, DetectGPT and, OpenAI's detector by over **40%**. The machine-generated texts for this task are from 7 very recent and highly sophisticated LLMs (including GPT3.5, GPT3 variants), making the detection of machine-generated text a much more challenging task on which GPT-who outperforms other detectors exceedingly.

For **TuringBench** (Table 3), GPT-who significantly outperforms GLTR by **0.32 F1** points, and at par with BERT fine-tuned for the task. The **InTheWild** dataset contains 6 testbeds with varying levels of detection difficulties, such as out-of-domain, out-of-distribution, and unseen-task test sets. We used all 6 testbeds to analyze the performance of GPT-who in detecting machine-generated texts across increasing levels of 'wildness' and find that overall, GPT-who outperforms all other methods except the one specifically tuned to the task (ITW) across all testbeds. More importantly, GPT-who performs tremendously even for the most challenging or 'wildest' testbed settings of unseen model and unseen domain distributions (see Table 4). For the **ArguGPT** dataset (Table 5), we find that GPT-who outperforms human experts and stylometry in predicting authorship by **0.31 F1** points, but is outperformed by fine-tuned BERT. Although unable to perform as well as BERT, GPT-who is one of the only statistical-based detectors that can handle distinctions between machine-only texts. We were unable to evaluate other detectors as their human-generated texts were not publicly released, and they only work in human v/s machine settings.

5 Discussion

We turn to the UID principle, which states that *humans prefer to spread information evenly in language*, to automatically extract features that measure the spread and flow of information content or surprisal in texts. Our UID-based features are formulated to capture how surprisal is distributed in an article as they measure the local and global variance, mean, and most uniform and non-uniform segments of a text. This rich and succinct representation space drives the predictive capability of our proposed detector and the interpretability of its representations. Analysis of this feature space reveals that **human-written text tends to be more non-uniform in comparison to machine-generated text**. Hence, machines seem to be spreading information more evenly or smoothly than humans who are more likely to have fluctuations in their surprisal distributions. We also find that UID-based features can capture differences between text generated by not only humans and models but also capture differences between multiple models and LM families. Our main contribution is a novel psycholinguistically-aware domain-agnostic multi-class statistical-based machine-generated text detector, GPT-who, that:

- Outperforms statistical approaches across 4 large-scale benchmark datasets that include texts from over 35 LLMs across more than 10 domains.
- Generalizes better to out-of-distribution datasets than SOTA detectors.
- Computationally more efficient than other supervised detectors as it does not require the fine-tuning or training of any LLMs.
- Intuitively interpretable due to its psycholinguistically motivated UID-based feature space.

While our detector may not significantly outperform fine-tuned transformers-based models, it is essential to highlight its independence from fine-tuning, offering nearly comparable performance at significantly lower computational costs and remains one of the only statistical-based detectors that can operate in multi-author settings beyond the Turing Test. These findings indicate that approaches rooted in psycholinguistic theories that delineate indicators of "human-like" language use hold enormous and untapped potential in tackling the fast catapulting and ever-changing LLM landscape. This work has implications for cognitively plausible and explainable solutions to complex challenges arising from ever-growing automated text generators.

Limitations

In our pursuit of a comprehensive examination of texts produced by recent large language models, we encountered limitations arising from resource constraints and the availability of publicly accessible datasets. These factors constrained our ability to encompass a more diverse array of models and tasks, including summarization and question-answering. Furthermore, our study did not delve into whether UID-based methods extend their utility beyond detecting machine-generated text to identify potential issues such as misinformation and plagiarism. We acknowledge these constraints as part of our ongoing commitment to refining and expanding our efforts in future research endeavors.

Ethical Statement

It is important to note that there are inherent limitations of AI-based tools and automated machine text detectors such as in this work. Acknowledging the fallibility of these detectors, particularly in generating false positives, we note that there is still a crucial need for human oversight and discretion in the usage of such detectors in real-world settings. For example, ethical concerns surrounding over-vigilance in scrutinizing student-written text are an important consideration for striking a balance between the convenience of automated detection and the preservation of academic integrity. By advocating for responsible development and implementation, we hope to contribute to a landscape where ethical considerations guide the integration of automatic text detection systems in educational settings, safeguarding against undue reliance and promoting fairness, equity, and respect for individual expression.

References

2023. Zerogpt.

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

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

Austin F Frank and T Florain Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society*, volume 30.

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2018. Uniform Information Density effects on syntactic choice in Hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48, Santa Fe, New-Mexico. Association for Computational Linguistics.

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

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647.

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

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Yu Lan, and Chao Shen. 2022. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023a. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023b. Check me if you can: Detecting chatgpt-generated academic writing using checkgpt. *arXiv preprint arXiv:2306.05524*.

Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980.

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

OpenAI. 2023. Gpt-4 technical report.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2022. Deepfake text detection: Limitations and opportunities. *arXiv preprint arXiv:2210.09421*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods.

Harry Tily and Steven Piantadosi. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the gap between Computational and Empirical approaches to Reference*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 25(1):1–18.

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

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. Turingbench: A benchmark environment for turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016.

Saranya Venkatraman, He He, and David Reitter. 2023. How do decoding algorithms distribute information in dialogue responses? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 923–932.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5191–5202.

Qiufang Wen, Lifei Wang, and Maocheng Liang. 2005. Spoken and written english corpus of chinese learners. *Foreign Language Teaching and Research Press*.

Yang Xu and David Reitter. 2016. Entropy converges between dialogue participants: Explanations from an information-theoretic perspective. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:537–546.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, (170):147–163.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

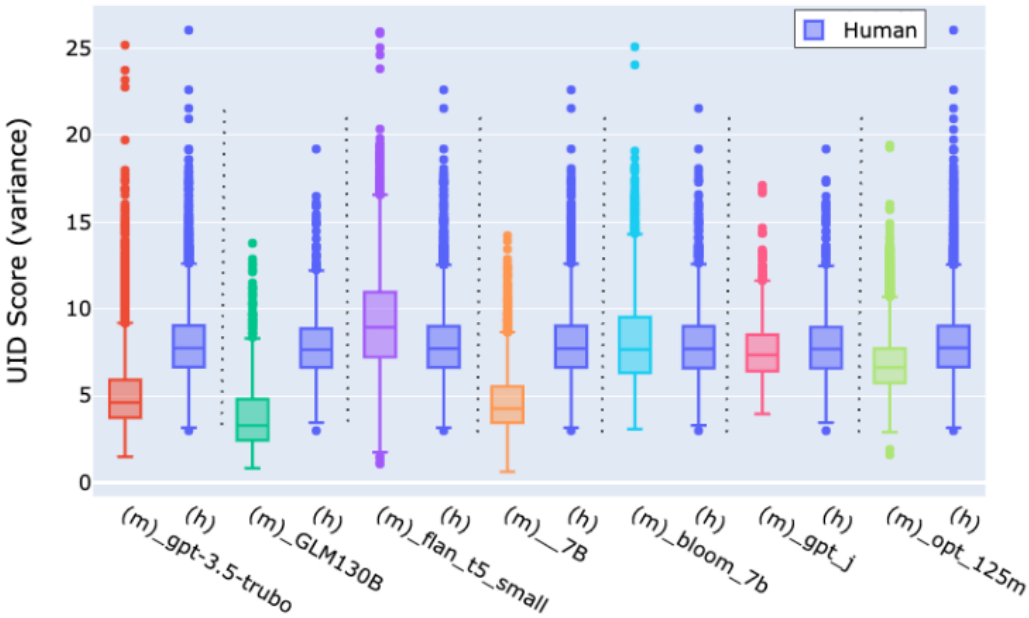
Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2461–2470.

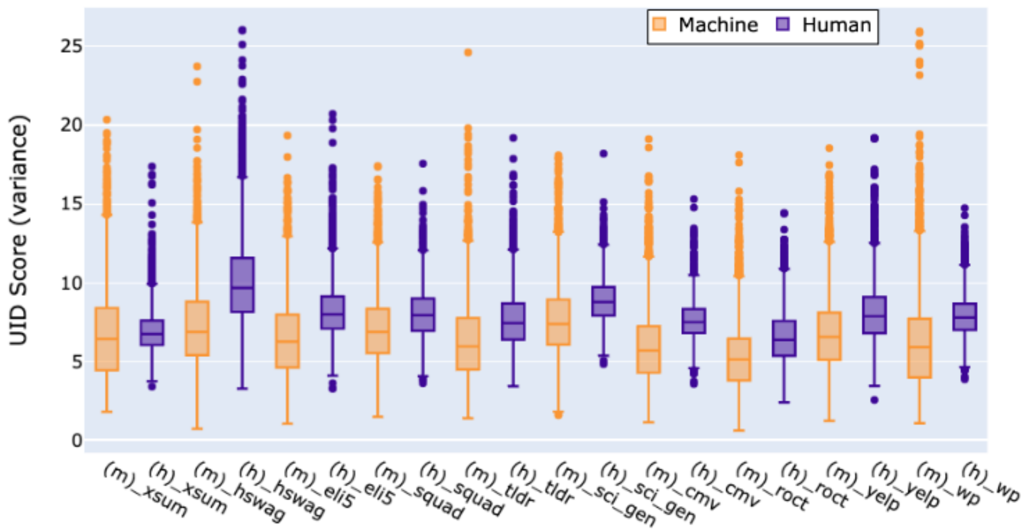
A Appendix

A.1 UID Score distributions of authors

We see that for most cases, humans have a higher UID (variance) score than machines, as can be seen by the higher means of their scores in the box plots. This holds when comparing human-written texts with multiple machine-generated texts over shared tasks (Figure 5a), and also when comparing their differences between tasks (Figure 5b).



(a) Pairwise comparisons of human and different machine-generated texts for shared tasks: Distribution of UID Scores of 8 authors (7 models + human) from the InTheWild dataset. (m) indicates machine and (h) indicates human written texts. This is followed by the model name along the x-axis labels to indicate the different authors.



(b) Pairwise comparisons of human and different machine-generated texts for different tasks: Distribution of UID Scores of humans v.s. machines per task type. (m) indicates machine and (h) indicates human written texts. This is followed by the writing task type along the x-axis labels to indicate the different tasks.