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

Saranya Venkatraman
The Pennsylvania State University
saranyav@psu.edu

Adaku Uchendu
MIT Lincoln Laboratory
adaku.uchendu@ll.mit.edu

Dongwon Lee
The Pennsylvania State University
donwon@psu.edu

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-inspired domain-agnostic statistical detector. This detector employs UID-based features to model the unique statistical signature of each LLM and human author for accurate detection. 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 better 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://github.com/saranya-venkatraman/gpt-who.

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 (Carrini 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).

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-inspired 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. Such features are calculated using the surprisals of tokens in a text, for example, the variance of surprisals, the mean, and so on (elaborated in Section 3.1). To showcase the detection capabilities of GPT-who, we use 4 large LLM benchmark datasets: Turing-
Bench (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 outperforms state-of-the-art statistical detectors and is at par with task and domain-specific fine-tuned LMs. This performative gain is consistent across benchmark datasets, types of LLMs, writing tasks, and domains.

In addition to improved detection performance, 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. GPT-who also 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 feature distributions of different LLMs and human-generated texts for the 4 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 are no glaring differences between uniform and non-uniform text, the differences in UID distributions are easily detectable and a powerful predictor of authorship, since they 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 (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 a 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 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 \( p(y | t) \) of each token \((y_t)\) in an article using a pre-trained LM (GPT2-XL). The surprisal \((u)\) of a token \(y_t\) is,

\[
    u(y_t) = - \log(p(y | t)),
\]

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** \((\mu)\) of an article \((y)\) of length
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.

lyl in number of tokens is defined as follows:

\[
\mu(y) = \frac{1}{|y|} \sum_{t} (u(y_t))
\]

(2)

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

\[
UID(y) = \frac{1}{|y|} \sum_{t} (u(y_t) - \mu)^2
\]

(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 \(\mu(y_t)\):

\[
UID(y) = \frac{1}{|y| - 1} \sum_{t=2}^{\left| y \right|} \text{abs}(\mu(y_t) - \mu(y_{t-1}))
\]

(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 \(\mu(y_t)\):

\[
UID(y) = \frac{1}{|y| - 1} \sum_{n=2}^{\left| y \right|} (\mu(y_t) - \mu(y_{t-1}))^2
\]

(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 all

\(^1\)https://github.com/rycolab/revisiting-uid/tree/main
\(^2\)https://huggingface.co/gpt2-xl
datasets of each of the 4 benchmarks as they were released by the dataset developers. We train a logistic regression model\(^3\) using these features on the train splits and report performance on the test splits. We averaged performance over 3 different random seeds and set the number of maximum iterations hyperparameter to 10k after testing a set of values. 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 recent datasets that span over 15 domains and 35 recent LMs.

#### TuringBench Benchmark (Uchendu et al., 2021)

The TuringBench Benchmark 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.

#### GPABenchmark (Liu et al., 2023b) or GPT Corpus for Academia

The 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)

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 WECL (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, FLANT5, 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: DetectGPT4 (Mitchell et al., 2023), GLTR5 (Gehrmann et al., 2019), an open-source implementation6 of GPTZero (Tian and Cui, 2023), ZeroGPT (ZeroGPT, 2023), OpenAI’s detector (Solaiman et al., 2019), Li et al. (2023)’s LongFormer-based detector7 tuned for the InTheWild benchmark (we refer to this method as “ITW”), a stylometric detector8 (Abbasi and Chen, 2008) and fine-tuned BERT9 (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 different 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.
<table>
<thead>
<tr>
<th>Task Type</th>
<th>Domain</th>
<th>GPTZero</th>
<th>ZeroGPT</th>
<th>OpenAI Detector</th>
<th>DetectGPT</th>
<th>BERT</th>
<th>ITW</th>
<th>GPT-who</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>CS</td>
<td>0.30</td>
<td>0.67</td>
<td>0.81</td>
<td>0.58</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>PHX</td>
<td>0.25</td>
<td>0.68</td>
<td>0.70</td>
<td>0.54</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>HSS</td>
<td>0.72</td>
<td>0.92</td>
<td>0.63</td>
<td>0.57</td>
<td>0.99</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>Task 2</td>
<td>CS</td>
<td>0.17</td>
<td>0.25</td>
<td>0.64</td>
<td>0.16</td>
<td>0.99</td>
<td>0.81</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>PHX</td>
<td>0.06</td>
<td>0.10</td>
<td>0.24</td>
<td>0.17</td>
<td>0.96</td>
<td>0.76</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>HSS</td>
<td>0.44</td>
<td>0.62</td>
<td>0.27</td>
<td>0.20</td>
<td>0.97</td>
<td>0.29</td>
<td>0.80</td>
</tr>
<tr>
<td>Task 3</td>
<td>CS</td>
<td>0.02</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.97</td>
<td>0.38</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>PHX</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.97</td>
<td>0.31</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>HSS</td>
<td>0.20</td>
<td>0.25</td>
<td>0.06</td>
<td>0.06</td>
<td>0.99</td>
<td>0.08</td>
<td>0.62</td>
</tr>
<tr>
<td>Average F1</td>
<td></td>
<td>0.24</td>
<td>0.40</td>
<td>0.38</td>
<td>0.26</td>
<td>0.98</td>
<td>0.62</td>
<td>0.83</td>
</tr>
</tbody>
</table>

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.

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

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.

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

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.
**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 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.

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 well 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.

GPT-who is a statistical-based approach that outperforms other statistical-based approaches but is unsurprisingly outperformed by fine-tuned methods in 2 out of 4 benchmarks. In the case of statistical-based methods, it is typically very hard to come close to fine-tuned performance as such methods rely only on derived properties of the text and do not utilize the full raw text in training as is the case in the latter (Jawahar et al., 2020). Despite this, GPT-who can exceed fine-tuned LM performance by 10% for 2 benchmarks.

**4.5 Running Time**

We measured the time taken for the one-time training or fine-tuning and inference for 6 testbeds from the InTheWild Dataset (the largest of all our benchmarks). We compare the average running times of DetectGPT, BERT, GPT-who and a stylometric detector in Table 6 and find that GPT-who is the fastest as it eliminates the need for any LM fine-tuning and makes a single inference call per text sample. Other methods require either LM fine-tuning or multiple inference calls (for example, DetectGPT). This computational load is greater than a single forward inference pass through one LM (GPT2) followed by logistic regression which is what GPT-who requires.

<table>
<thead>
<tr>
<th>Method</th>
<th>One-Time Training</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DetectGPT</td>
<td>&gt;10 hours</td>
<td>60 sec</td>
</tr>
<tr>
<td>BERT</td>
<td>~1.5 hours</td>
<td>2 sec</td>
</tr>
<tr>
<td>Stylometry</td>
<td>~1.5 hours</td>
<td>2 sec</td>
</tr>
<tr>
<td>GPT-who</td>
<td>20 min</td>
<td>0.8 sec</td>
</tr>
</tbody>
</table>

Table 6: Average Running time over 6 testbeds from the InTheWild dataset.

**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. However, this finding does not imply that humans are not producing uniform text. It is important to note that our work cannot provide support for or refute the UID hypothesis which comes from psycholinguistic studies such as those in Section 2.1. Our work shows that, given our operationalization of UID based on prior works, machine text is relatively more uniform than human-written text. While this might seem contradictory to UID theory, it does not still disprove that humans are uniform in their language production.

We conjecture that this unexpected finding is because we use GPT-2’s probability distribution to calculate surprisal, which is potentially a poor approximation of the “human” probability distribution. A closer-to-human probability distribution might (or might not) show humans to be more uniform than machines, though this determination is not within the scope of this work. It is crucial to note that uniformity is relative, and while machines are more uniform under this operationalization, it would still be true that human text is uniform as per the human’s probability distribution (that we do not have access to and can only approximate using some LM distribution, for example, GPT2-XL in our case). UID theory does not make any predictions on where machine-generated text might lie in the uniform to non-uniform spectrum but only indicates that humans are arranging utterances evenly as per their own language distribution.

Irrespective of its alignment with what theory suggests, we find that the UID-based features are very useful in distinguishing authors, which is the focus of this work. This is an important consideration and helps disentangle the utility of UID-inspired features from the cognitive plausibility of those feature calculations or UID approximations.

Thus, this operationalization of UID does not imply that humans are “less human” or machines are “more human” as it is an approximation of a theory that states that humans are uniform as per their language distribution. It does not have any further implications for machine-generated text and is unable to predict what happens in the case of machines.

We 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 psycholinguistically-inspired domain-agnostic 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.
- 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.

6 Conclusion

We propose GPT-who, a statistical-based machine-generated text detector that utilizes features inspired by the Uniform Information Density (UID) principle rooted in the observation that humans tend to evenly distribute information or surprisal in language. We leverage UID-based features, including variance of surprisals and minimally/maximally uniform sub-sequences extracted from the surprisal sequence generated by an off-the-shelf LM. We demonstrate that these features are highly effective in discerning machine-generated text from human-generated text as they capture nuances in how models and humans distribute surprisal in their texts. Our findings have implications for enhanced text authenticity assessment.
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.

Acknowledgments

This work was in part supported by National Science Foundation (NSF) awards #1820609, #1950491, and #2131144.

References


Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,


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.