WORKING PAPER · NO. 2025-116

Artificial Writing and Automated Detection

Brian Jabarian and Alex Imas

AUGUST 2025

THE UNIVERSITY OF CHICAGO
BECKER FRIEDMAN INSTITUTE
FOR ECONOMICS

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Artificial Writing and Automated Detection

Brian Jabarian Alex Imas

This version: August 26, 2025

Abstract

Artificial Intelligence (AI) tools are increasingly being used for written deliverables in a wide variety of domains. In some cases, the recipient of the deliverables wants to ensure that the content was written by a human rather than an AI tool, e.g., ensuring assignments were completed by students, product reviews written by actual customers, etc. This creates a demand for AI detection tools that minimize two key statistics: the False Negative Rate (FNR), which corresponds to the proportion of AI-generated text that is falsely classified as human, and the False Positive Rate (FPR), which corresponds to the proportion of human-written text that is falsely classified as AI-generated. We evaluate four commercial and open-source AI-text detectors—Pangram, OriginalityAI, GPTZero and RoBERTa—on these dimensions using a large corpus of human and AI-generated text that spans across topics, length, and AI models. First, we find that detectors vary in their capacity to minimize FNR and FPR, with the commercial detectors outperforming open-source. Second, most commercial AI detectors perform remarkably well, with Pangram in particular achieving a near zero FPR and FNR within our set of stimuli; these results are stable across AI models. Third, while Pangram's performance largely holds up on very short passages (< 50 words) and is robust to "humanizer" tools (e.g., StealthGPT), the performance of other detectors becomes case-dependent. Finally, we consider the implementation of detectors as policy, noting that a policy designer faces a trade-off between maximizing the probability of detecting true AI-generated text while minimizing the risk of false accusations. Given this tradeoff, we propose an evaluation metric that uses *policy caps*—a scale-free, detector-independent measure that corresponds to the designer's tolerance for false positives or negatives—to compare detectors. Using this metric, we show that Pangram is the only detector that meets a stringent policy cap (FPR ≤ 0.005) without compromising the ability to accurately detect AI text.

Contact: Brian Jabarian; Booth School of Business, University of Chicago, CAAI, CESifo, J-PAL, IGL, JILAEE, brian.jabarian@chicagobooth.edu. Alex Imas; Booth School of Business, University of Chicago, CAAI, CESifo, HCEO, NBER; alex.imas@chicagobooth.edu.
Acknowledgments: We thank Kevin Bryan for helpful feedback. Ziyue Feng and Andrew James provided excellent research assistance. All remaining errors are our own.
Funding: Brian Jabarian gratefully acknowledges funding support from the University of Chicago Booth Center for Applied Artificial Intelligence, the Becker-Friedman Institute Program in Behavioral Economics Research and the Google Cloud Research Program.
Disclosure. All authors declare that they have no financial or personal conflicts of interest related to this study.

1 Introduction

Generative Artificial Intelligence tools have been adopted faster than any other technology on record (Bick, Blandin, and Deming, 2024). This has given rise to writing that is either assisted or entirely completed by Large Language Models (LLMs). From assignments completed for school to company documents and consumer reviews, AI-generated writing is becoming ubiquitous (Hanley and Durumeric, 2024). Stakeholders across a wide variety of domains are increasingly facing this new reality and are actively trying to figure out how to react to it. But a key pre-condition for designing the optimal policy for dealing with AI-generated writing is to detect it in the first place. While studies have shown that AI experts can successfully detect LLM-written text, this is not a scalable solution given the deluge writing that may need to be screened (Russell, Karpinska, and Iyyer, 2025). Automated detection tools hold promise in allowing the recipients of potentially AI-generated writing to design policy that is in-line with their objectives. Yet published claims about detector accuracy are hard to verify: they are often based on private data, focus on a single threshold, and ignore firm (cost of signal accuracy) (Tang, Chuang, and Hu, 2024), market (short length texts flooding economic markets) (Tardelli et al., 2022), or policy constraints (trade-offs between different social risks in detection) (Crothers, Japkowicz, and Viktor, 2023). Independent studies that examine AI detection are out of date and focus on a set of tools that rely on antiquated techniques (Weber-Wulff et al., 2023).

The current paper aims to audit the set of leading AI detection tools and offers a framework to evaluate how they should be incorporated into potential policies. A detector evaluates text and assigns it a score, such that higher scores imply a greater likelihood that the text is AI-generated. We consider the objective function of a policy designer who aims to minimize both the False Negative Rate (FNR)—failing to detect actual AI writing—and the False Positive Rate (FPR)—falsely flagging human writing as AI of a detector. A detector's performance on these metrics depends critically on the threshold (τ) used to classify a text as AI-generated or not based on the assigned score. A higher threshold implies that the detector requires a higher score to classify a passage as AI-generated; this will naturally decrease the FPR while at the same time (weakly) increasing the FNR. A higher threshold can thus be interpreted as greater conservatism.

We evaluate performance in several ways. First, we report results on FPR and FNR using detector-specific optimal thresholds. These thresholds are calculated to maximize the distance between the True Positive Rate (TPR) and the FPR.[1] Next, we perform a

[1]This corresponds to maximizing Youden's J statistic.

sensitivity analysis for each detector by manipulating the threshold exogenously and reporting the resultant FPR and FNR. We then generate additional statistics to evaluate detectors: a) *area under the ROC curve* (AUROC), which corresponds to the probability that a randomly chosen AI-generated passage is ranked above a randomly chosen human-generated passage by the detector and b) the $\Delta-$Mean, which corresponds to the cardinal distance between the detector's average score assigned to actual AI-generated passages versus human-generated passages.

Our analysis evaluates three leading commercial detectors, Pangram, OriginalityAI, and GPTZero, together with an open-source baseline (RoBERTa-base), on their FPR, FNR, and the two additional statistics. As input, we use a 1,992-passage text corpus that spans six everyday genres (news, blogs, consumer reviews, novels, restaurant reviews, and résumés). Verified human-generated text is matched with AI-generated text using four frontier LLMs (GPT-4.1, Claude Opus 4, Claude Sonnet 4, Gemini 2.0 Flash). We also examine the effectiveness of AI "humanizers" (StealthGPT) in potentially bypassing detectors.

The results are clear-cut. Pangram stands out as the only AI detector maintaining policy-grade levels on our main metrics when evaluated on all four generative AI models. On medium-length to long passages, Pangram achieves essentially *zero* FPRs and FNRs within our sample, both when using detector-optimized thresholds and exogenously-set thresholds. Both the FPR and FNR increase slightly on short passages, but remain well below reasonable policy thresholds. The AUROC is one or arbitrarily close to one, and the $\Delta-$mean is nearly 1 (0.9-1) in almost all genres and LLM models used, except the $\Delta-$mean in restaurant review between 0.8049 and 0.8878. Pangram also performs surprisingly well on text 'stubs' (< 50 words). Importantly, Pangram's FNR is robust to the use of current "humanizers"; the FNR remains low even when AI-generated passages are modified using tools such as StealthGPT. OriginalityAI and GPTZero constitute a secondary tier with partial strengths, making the choice between the two dependent on the user's priority: minimizing FPR favors GPTZero, while maximizing AUROC favors OriginalityAI. But neither is suitable for very short text ('stubs') and both are susceptible to "humanizers." RoBERTa base is deemed unsuitable for high-stakes applications.

We also consider the cost of implementing a detector-based policy. After converting vendor fees into cost per correctly flagged AI passage, Pangram is two times cheaper than OriginalityAI and is almost three times cheaper than GPTZero both overall and on shorter passages.

Finally, since FPR and FNR are negatively correlated, the policy designer faces a trade-off between maximizing the probability of detecting true AI-generated text while minimizing the risk of false accusations. The key choice variable that governs this tradeoff

is the threshold for classifying text as AI-generated or not. The threshold is thus a critical choice variable for the policy designer that depends on her objective function. In settings where missing AI-generated text is costlier than false positives, the threshold should be set at a lower level; in settings where false accusations are costlier than failing to detect AI-generated text, the threshold should be set higher.

Given this choice variable, we propose an evaluation metric that uses *policy caps*—a scale-free, detector-independent measure that corresponds to the detector's tolerance for false positives—to compare detector performance. Specifically, to evaluate detectors, we first set a policy cap x% of maximum allowable false positives (e.g., 0.5%, 1%, etc) and then back-end adjust the detector's internal thresholds such that the FPR $\leq x$% on the relevant sample. Using this metric, we find that Pangram is the only detector that meets stringent policy requirements without compromising the ability to detect AI text, while OriginalityAI comes second, failing slightly short on for higher policy caps. The open-source tool RoBERTa base loses functionality when stringent policy rates are imposed.

An important question emerges when considering the implementation of AI-detection tools: At what stage of the writing process should LLM involvement be flagged? For example, it may be the case that an LLM entirely generated a text, rephrased a human-generated draft written by a non-native speaker, or corrected grammar through an AI-based app like Grammarly. It seems natural for institutions to adapt their policy responses to these different usages of AI. A much anticipated development in AI detection techniques lies in the capacity to detect at which stages of the writing process the LLM was used. Before this development, however, our methodology already offers a portable toolkit for detection as a function of the policy designer's objective function. By constructing their *policy caps* accordingly, designers can account for different stages of possible AI usage. For example, setting a conservative FPR cap would be advisable for education settings, as this would facilitate LLM-assisted writing (e.g., Grammarly)—since it would unlikely be flagged—at a cost of a potentially higher FNR. As more technical progress by AI detectors is made, it will be easier to move away from binary classification, which will allow more targeted mechanisms to emerge.

Finally, it is important to note that we do not expect the results presented here to remain constant over time. Detector performance will likely be a continuous technical arms race between detectors, LLMs, the sophistication of "humanizers," and how users strategically adapt their AI usage for the writing process. Because of this, routine and transparent audits, similar to stress tests in banking, should become part of platform compliance, with results published so thresholds can be reset before large-scale abuse emerges.

The remainder of the paper is organized as follows. Section 2 describes the 1,992-

passage benchmark, audit protocol, and error metrics. Section 3 presents the main performance results on accuracy, threshold robustness, and length effects. Section 4 probes two stress tests: very short passages (<50 word 'stubs') and "artificial humanizers." Section 5 explores the implications of policy, comparing the costs of detection accuracy, and introducing a framework to guide detector choice under different FNR versus FPR objectives. Section 6 concludes.

2 Methodology

2.1 Benchmark Dataset: Human and AI-Generated texts

We rely on human-authored text and passages matched in length and content generated by LLMs. Our dataset comprises a corpus of 1,992 human passages total; for each source model, we generate an AI corpus of size 1,992 to match. Human content was selected pre-2020 to ensure that it was not AI-generated.

Human-Written Texts. Our collection of human-written samples originates from six public datasets. They contain distinct writing styles, lengths, and text structures: formal news, informal blogs, structured professional documents, short restaurant or retail reviews, and long novels. The entire universe of this collection, from which we draw our subset, includes journalistic pieces from CC-News; first-person narratives from posts in the Bar-Ilan Blog Authorship Corpus; structured résumés from the Kaggle Résumé dataset; reviews from entries in the Yelp Review Full set; product assessments from Amazon Reviews spread across 30 categories; and excerpts from pre-2000 English novels on Project Gutenberg.

AI-Generated Texts. We generate AI synthetic equivalents using GPT-4.1, Claude Opus 4, Claude Sonnet 4, and Gemini 2.0 Flash.[2] The temperature is adjusted to 0.7, and the output duration is capped at twice the input length, with a maximum of 4,096 tokens. The prompt used, exactly reproduced in Appendix E, directs the model to reflect the substance and structure of arguments without duplicating the original language and specific linguistic features, specific to human writing.[3]

Descriptive Statistics. Our dataset comprises 1,992 human passages per source model, consisting of 200 Amazon reviews, 200 blogs, 300 news articles, 1,000 novel excerpts,

[2]The complete analysis in this paper is also replicated with GPT-3.5. This model is not included in the main document due to its outdated status, but is outlined in the Appendix since this could still provide insight into the output of smaller language models or open-source models.

[3]This mimetic replication is what AI humanizers are designed for; we test this in Section 4.4.

100 restaurant reviews, and 192 resume passages. The AI corpus exactly reflects these human counts ($N_{\text{AI}} = N_H$). The benchmark was crafted to match human and AI texts on three key axes: length, lexical variety, and readability.[4] This approach facilitates the detection of genuine stylistic discrepancies by minimizing the imbalance of the data set in AI detection evaluations. One key dimension of variation is the length of the passages, which we demarcate as long (novel excerpts and resume), medium (news and blogs), and short (Amazon and restaurant reviews).

To give a few examples of such characteristics, word count comparisons already reveal genre similarities: the average word lengths for human-generated long passages, such as novel and resume, are 966.77 and 789.51 respectively, while the average lengths for the AI-generated novel and resume are 1037.75 and 833.65. For medium passages, human-generated news articles average 459.88 words; AI-generated passages are 458.40 words. Finally, for short passages, the average length for human-generated Amazon reviews is 79.57 words while AI-generated text is 76.68. In subsequent detection analyzes, both AI and human-generated text segments are considered individual observational entities.

2.2 Experimental Audit Design: Testing AI Detectors

AI-detectors. We consider three industry-leading commercial AI detectors, Pangram, GPTZero, and OriginalityAI, as well as an open-source RoBERTa classifier fine-tuned with public GPT-2 output.[5]

Performance measures. Our performance measures are based on two key metrics: the False Positive Rate and the False Negative Rate. The False Positive Rate (FPR) corresponds to a Type I error, representing the probability that a passage genuinely authored by a *human* is incorrectly identified as being generated by AI. It is formally defined as follows. Let $y \in \{0, 1\}$ denote the authorship, where $y = 1$ signifies an AI-generated text, and $s \in [0, 1]$ represents the detector score, with higher s corresponding to a higher likelihood of being AI-generated. An AI detector will classify the text as AI-generated if the score $s \geq \tau_{\text{raw}}$, where τ_{raw} corresponds to the choice of threshold that the detector uses to classify the text as AI-generated or not. Given this setup, let:

$$\text{FPR}(\tau_{\text{raw}}) = \Pr[s \geq \tau_{\text{raw}} \mid y = 0].$$

[4] See Table A.1 in Appendix A for detailed corpus size and linguistic attributes per genre model.

[5] Appendix D contains comprehensive versions and endpoint details for each detector. The replication package includes all associated code and logs. Refer to the repository README for detailed replication procedures.

The False Negative Rate (FNR) corresponds to a Type II error, representing the probability that an AI-generated text is not classified as being AI-generated by the detector. Let the FNR be defined as follows:

$$\text{FNR}(\tau_{\text{raw}}) = \Pr[s < \tau_{\text{raw}} \mid y = 1].$$

It is readily apparent that the two key metrics are a function of the threshold τ_{raw}. Setting a lower threshold will classify more text as AI-generated, but this will (weakly) increase the FPR; at the same time, setting a lower threshold will (weakly) decrease the FNR. Given this trade-off between FPR and FNR, the choice of the threshold is a key variable to evaluate performance.

We will evaluate the detector performance in two ways. The first evaluates FPR and FNR based on a detector-optimized threshold. This threshold is calculated by maximizing the difference between the True Positive Rate (TPR) and the False Positive Rate (FPR) within the given set of texts, where the TPR corresponds to the rate at which the detector correctly classifies a passage as AI-generated. Note that this exercise is akin to choosing a threshold that maximizes Youden's J statistic. Importantly, the threshold can also be a key choice variable for the policy designer depending on their objective function. A higher tolerance for FPR over FNR implies setting a lower threshold, while a higher tolerance for FNR over FPR implies setting a higher threshold. Our second set of performance metrics vary the thresholds exogenously, demonstrating how a policy designer who wants to guarantee a certain FPR or FNR statistic may engage with the detector when deciding how to implement it in practice.

We also define two additional statistics to evaluate detector performance. First, the Area Under the ROC Curve (AUROC) represents the probability that a randomly chosen passage generated by AI surpasses a randomly selected human-authored passage in ranking in terms of its detector-assigned score. It is formally defined as follows:

$$\text{AUROC} = \Pr[s_i > s_j \mid y_i = 1, y_j = 0].$$

Second, $\Delta-$Mean represents the average detector-assigned score to AI-generated text minus the average score assigned to human-generated text. A large $\Delta-$Mean implies that the detector achieves clear separation between AI and human-generated text, while a small $\Delta-$Mean implies potentially significant overlap. It is defined formally as follows:

$$\Delta\text{–Mean} \;=\; \mathbb{E}[s \mid y = 1] \;-\; \mathbb{E}[s \mid y = 0].$$

We now proceed to use these metrics to evaluate the AI detectors in our consideration set.

3 Key Results

3.1 Overall Performance

Detector-Optimized Thresholds. We begin by computing the detector-optimized thresholds. For each detector, we calculate the thresholds that yield the highest difference between TPR - FPR and then obtain the FNR by calculating 1 - TPR within our corpus. Calibrating the thresholds on our corpus gives each detector the chance to perform at its best. The TPR - FPR maximizing thresholds are reported in Table 1.

Table 1: Thresholds Maximizing TPR-FPR (Youden's J) by Genre.

Genre	GPTZero	Originality	Pangram	RoBERTa
amazon review	0.8750	0.3508	0.0077	0.9972
blog	1.0000	0.0111	0.1273	0.9693
news	0.9783	0.4294	0.2362	0.9989
novel	1.0000	0.0032	1.0000	0.9997
restaurant review	0.9491	0.3257	0.0021	0.9904
resume	1.0000	0.7202	0.9994	0.9995

We report performance results using these detector-optimized thresholds in Tables 2 and 3.

False Positive Rate. Table 2 presents FPR results across all text genres. Pangram emerges as the most cautious detector. Pangram achieves a *zero* FPR on longer passages and essentially a zero FPR on medium passages. The FPR increases a bit on shorter passages, but never rises above 0.01. Both GPTZero and OriginalityAI keep the FPR at 0.01 or below on medium to long passages and below 0.03 on shorter passages. RoBERTa base diverges significantly, misclassifying most of the human text, with FPRs of approximately 0.30-0.69 in all scenarios.

Table 2: Detector False Positive Rates by Genre.

Genre	GPTZero	Originality	Pangram	RoBERTa
amazon review	0.0238	0.0170	0.0050	0.3063
blog	0.0050	0.0000	0.0000	0.6488
news	0.0100	0.0026	0.0008	0.7775
novel	0.0045	0.0025	0.0000	0.5315
restaurant review	0.0100	0.0218	0.0075	0.5175
resume	0.0000	0.0013	0.0000	0.6914

False Negative Rate. Next, we consider detector performance in terms of FNR across LLM models and text genres. Table 3 presents the results. Here we see a larger separation between Pangram and the other detectors. On medium to long texts, Pangram achieves near *zero* FNR.[6] The FNR stays below 0.01 even for shorter text (except for Gemini 2.0 Flash-generated restaurant reviews, where it was 0.02). OriginalityAI and GPTZero compete for second place. OriginalityAI's FNR was at 0.02 or below for medium to long passages, but increased to 0.05 or below for shorter passages. GPTZero's FNR was 0.05 or below for medium to longer passages, and increased to 0.07 for shorter passages. RoBERTa base was again an exception, missing up to 51% of the AI-generated passages.

[6]The FNR was greater than zero only for GPT-4.1 and Gemini 2.0 Flash-generated blogs, where the FNR was still less than or equal to 0.01.

Table 3: Detector Performance by Genre and Model: FNR

Model	Genre	GPTZero	Originality	Pangram	RoBERTa
GPT-4.1	amazon review	0.0050	0.0076	0.0000	0.5100
GPT-4.1	blog	0.0000	0.0000	0.0050	0.0700
GPT-4.1	news	0.0033	0.0000	0.0000	0.1433
GPT-4.1	novel	0.0000	0.0000	0.0000	0.1830
GPT-4.1	restaurant review	0.0200	0.0000	0.0000	0.2800
GPT-4.1	resume	0.0000	0.0000	0.0000	0.0156
Claude Opus 4	amazon review	0.0500	0.0148	0.0050	0.2550
Claude Opus 4	blog	0.0300	0.0208	0.0000	0.1650
Claude Opus 4	news	0.0167	0.0103	0.0033	0.0567
Claude Opus 4	novel	0.0320	0.0020	0.0000	0.0320
Claude Opus 4	restaurant review	0.0500	0.0435	0.0000	0.3300
Claude Opus 4	resume	0.0000	0.0000	0.0000	0.0052
Claude Sonnet 4	amazon review	0.0250	0.0000	0.0050	0.2650
Claude Sonnet 4	blog	0.0200	0.0069	0.0000	0.0700
Claude Sonnet 4	news	0.0000	0.0034	0.0000	0.0700
Claude Sonnet 4	novel	0.0060	0.0060	0.0000	0.0400
Claude Sonnet 4	restaurant review	0.0200	0.0417	0.0100	0.1700
Claude Sonnet 4	resume	0.0000	0.0000	0.0000	0.0052
Gemini 2.0 Flash	amazon review	0.0500	0.0156	0.0050	0.3150
Gemini 2.0 Flash	blog	0.0450	0.0000	0.0100	0.2900
Gemini 2.0 Flash	news	0.0100	0.0034	0.0033	0.1433
Gemini 2.0 Flash	novel	0.0000	0.0000	0.0000	0.2230
Gemini 2.0 Flash	restaurant review	0.0700	0.0448	0.0200	0.1700
Gemini 2.0 Flash	resume	0.0000	0.0000	0.0000	0.1250

3.2 Threshold Sensitivity

Results in the previous subsection used each detector's optimized thresholds. We now explore the robustness of each detector's FPR and FNR to exogenous changes in this threshold.

False Positive Rate. Pangram dominates the other detectors across all thresholds. As shown in Table 4 below, the FPR is essentially zero across all thresholds 0.5 and greater (which includes the detector-optimized threshold). The FPR increases to at most .001 (less than a tenth of one percent) when the very loose threshold of 0.1 is imposed. Coming in second, OriginalityAI's FPR ranges between 0.001 at the tightest threshold and 0.003 at the loosest threshold. GPTZero's FPR does not change across the threshold range used,

staying at around 0.007. RoBERTa-base performs the worst, labeling more than 90% of human passages as "AI" even at the highest threshold.

Table 4: False Positive Rates at Different Raw Thresholds

Detector	0.1	0.3	0.5	0.7	0.9
gptzero	0.0071	0.0071	0.0071	0.0071	0.0071
originality	0.0027	0.0011	0.0011	0.0011	0.0011
pangram	0.0010	0.0005	0.0000	0.0000	0.0000
roberta-base-detector	0.9774	0.9608	0.9503	0.9327	0.8991

False Negative Rate. As shown in Table 5, Pangram's FNR ranges between 0.0045 and 0.038, depending on the threshold and LLM model used; it only misses between 0.0161 and 0.0382 of AI-generated text even at the strictest threshold at 0.9. The FNR for GPTZero ranges between 0.002 and 0.030 depending on the threshold and LLM model. OriginalityAI performs worse than both detectors: depending on the LLM model, the FNR can be as high as .300 even on the loosest threshold of 0.1, increasing to .424 at the strictest threshold of 0.9. Finally, RoBERTa's relatively low FNR scores are misleading, largely driven by the fact that it incorrectly classifies so many passages as AI-generated.

Table 5: False Negative Rates at Different Raw Thresholds

		Thresholds				
Model	**Detector**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
GPT-4.1	gptzero	0.0020	0.0020	0.0025	0.0025	0.0025
	originality	0.1102	0.1431	0.1606	0.1820	0.2166
	pangram	0.0045	0.0075	0.0131	0.0136	0.0161
	roberta-base-detector	0.0060	0.0120	0.0176	0.0291	0.0452
Claude Opus 4	gptzero	0.0291	0.0291	0.0291	0.0291	0.0296
	originality	0.2989	0.3514	0.3733	0.3957	0.4242
	pangram	0.0120	0.0136	0.0216	0.0246	0.0286
	roberta-base-detector	0.0015	0.0025	0.0060	0.0080	0.0206
Claude Sonnet 4	gptzero	0.0085	0.0085	0.0085	0.0085	0.0085
	originality	0.1333	0.1687	0.1877	0.2051	0.2367
	pangram	0.0065	0.0085	0.0151	0.0171	0.0201
	roberta-base-detector	0.0025	0.0035	0.0050	0.0085	0.0146
Gemini 2.0 Flash	gptzero	0.0146	0.0146	0.0146	0.0146	0.0146
	originality	0.0154	0.0242	0.0270	0.0347	0.0424
	pangram	0.0141	0.0181	0.0272	0.0292	0.0382
	roberta-base-detector	0.0075	0.0126	0.0171	0.0266	0.0462

3.3 AUROC and Δ-Mean

Overall discrimination (AUROC). Table 6 reports the results on the general discrimination between AI and human-generated text across the genre / length of the text and the LLM model. Pangram achieves nearly flawless classification across all four source models. For medium to long passages, the AUROC scores are 1.0000 for the vast majority of categories.[7] Even for shorter passages, the AUROC never dips below 0.9979. OriginalityAI scores are high but lower than Pangram's across the board. GPTZero's scores are lower still, approaching 0.9600 for shorter passages. Meanwhile, the RoBERTa base performs around or below random (corresponding to 0.5) on most categories.

Table 6: Detector Performance by Genre and Model: AUROC

Model	Genre	GPTZero	Originality	Pangram	RoBERTa
GPT-4.1	amazon review	0.9849	0.9996	0.9998	0.6271
GPT-4.1	blog	0.9975	1.0000	0.9998	0.5701
GPT-4.1	news	0.9933	1.0000	1.0000	0.4708
GPT-4.1	novel	0.9980	1.0000	1.0000	0.6165
GPT-4.1	restaurant review	0.9850	1.0000	0.9998	0.5870
GPT-4.1	resume	1.0000	1.0000	1.0000	0.4371
Claude Opus 4	amazon review	0.9625	0.9965	0.9987	0.7845
Claude Opus 4	blog	0.9825	0.9996	1.0000	0.5672
Claude Opus 4	news	0.9867	0.9994	0.9999	0.4063
Claude Opus 4	novel	0.9815	0.9999	1.0000	0.6445
Claude Opus 4	restaurant review	0.9700	0.9841	0.9996	0.6224
Claude Opus 4	resume	1.0000	0.9999	1.0000	0.3965
Claude Sonnet 4	amazon review	0.9775	0.9997	0.9996	0.7626
Claude Sonnet 4	blog	0.9875	0.9993	1.0000	0.5763
Claude Sonnet 4	news	0.9950	1.0000	1.0000	0.3957
Claude Sonnet 4	novel	0.9950	1.0000	1.0000	0.6244
Claude Sonnet 4	restaurant review	0.9850	0.9954	0.9999	0.6114
Claude Sonnet 4	resume	1.0000	1.0000	1.0000	0.4060
Gemini 2.0 Flash	amazon review	0.9625	0.9973	0.9993	0.7023
Gemini 2.0 Flash	blog	0.9750	1.0000	0.9998	0.5672
Gemini 2.0 Flash	news	0.9900	1.0000	1.0000	0.4546
Gemini 2.0 Flash	novel	0.9975	1.0000	1.0000	0.6113
Gemini 2.0 Flash	restaurant review	0.9600	0.9921	0.9979	0.5570
Gemini 2.0 Flash	resume	1.0000	1.0000	1.0000	0.5256

[7] Only two LLM x genre score are below 1.0000, with the lowest (GPT-4.1 news) at 0.9998.

Score–separation (Δ-Mean). Pangram and GPTZero perform similarly well in score separation between AI and human-generated text, with mean differences (Δ-Mean) of approximately 0.805 to 1.0 (see Table 7). OriginalityAI achieves smaller score separation that is more LLM model dependent: it performs well with Gemini Flash (0.873 - 0.999), but substantially less so with Claude Opus 4 (0.417 - 0.976) and GPT-4.1 (0.709 - 0.999). The RoBERTa base shows an almost negligible separation between AI and human-generated text (0.002-0.199), demonstrating that its scores barely differentiate between the two types of passages.

Table 7: Detector Performance by Genre and Model: Δ-Mean

Model	Genre	GPTZero	Originality	Pangram	RoBERTa
GPT-4.1	amazon review	0.9675	0.9811	0.9642	0.1153
GPT-4.1	blog	0.9950	0.9722	0.9850	0.1090
GPT-4.1	news	0.9864	0.9982	0.9964	0.0185
GPT-4.1	novel	0.9960	0.7086	1.0000	0.0076
GPT-4.1	restaurant review	0.9700	0.9801	0.8859	0.0800
GPT-4.1	resume	1.0000	0.9994	1.0000	0.0191
Claude Opus 4	amazon review	0.9250	0.8984	0.9279	0.1990
Claude Opus 4	blog	0.9650	0.6869	0.9599	0.1078
Claude Opus 4	news	0.9733	0.9491	0.9872	0.0418
Claude Opus 4	novel	0.9630	0.4175	1.0000	0.0132
Claude Opus 4	restaurant review	0.9380	0.7750	0.8529	0.0923
Claude Opus 4	resume	1.0000	0.9759	1.0000	0.0193
Claude Sonnet 4	amazon review	0.9550	0.9744	0.9396	0.1936
Claude Sonnet 4	blog	0.9750	0.7509	0.9897	0.1244
Claude Sonnet 4	news	0.9900	0.9873	0.9933	0.0350
Claude Sonnet 4	novel	0.9900	0.7028	1.0000	0.0132
Claude Sonnet 4	restaurant review	0.9700	0.9048	0.8878	0.1105
Claude Sonnet 4	resume	1.0000	0.9984	1.0000	0.0193
Gemini 2.0 Flash	amazon review	0.9250	0.9506	0.9128	0.1739
Gemini 2.0 Flash	blog	0.9500	0.9023	0.9477	0.1025
Gemini 2.0 Flash	news	0.9800	0.9933	0.9898	0.0071
Gemini 2.0 Flash	novel	0.9950	0.9731	1.0000	0.0029
Gemini 2.0 Flash	restaurant review	0.9200	0.8728	0.8049	0.0783
Gemini 2.0 Flash	resume	1.0000	0.9993	1.0000	0.0140

4 Robustness

4.1 Performance on 'Stubs'

The analysis in the previous section has shown that the performance of AI detectors decreases in shorter passages. However, numerous settings such as job boards, gig-work sites, and social media where AI detection would be useful often feature even shorter passages (< 50 words), which we refer to as 'stubs.' We evaluated each detector using 8% of our corpus comprising passages under 50 words ($N_H = N_{AI} = 160$ per model). These stubs of text span the same categories as before.

Among the four detectors, only Pangram, GPTZero, and RoBERTa assign a score to each stub. *OriginalityAI* provides a probability for certain stubs, but indicates a 'length filter' error for others. The selective rejection of certain stubs precludes the use of OriginalityAI for this analysis.

False-positive rate. Table B.1 Pangram shows notable conservatism, with mostly zero FPRs across all categories, reaching at most 0.025 in restaurant reviews. GPTZero comes in at a close second, with mostly very low FPR statistics. In comparison, the open-source *RoBERTa-base* detector incorrectly identifies between 0.2948 and 0.4871 of authentic human text.

False-negative rates. Table B.2 presents the FNR statistics on stubs. Pangram mostly produces FNR statistics close to zero, with a notable exception of 'news' passages generated by Claude Opus 4 and Gemini 2.0 Flash. GPTZero has a substantially higher miss rate across categories. RoBERTa performs the worst among the three, with a miss rate ranging from 0 to 0.4444.

Overall discrimination. As shown in Table B.3, the AUROC for Pangram on stubs is high, ranging from 0.9630 to 1. It ranges from 0.7778 to 1 for GPTZero, and from 0.5679 to 0.7900 for RoBERTa.

4.2 Robustness to Artificial Humanizers

Recent "humanizer" AI systems such as *StealthGPT* promise to artificially generate "humanized" versions of original AI-generated texts. This means that the AI-humanized versions contain linguistic features specifically observed in human writing. Such AI humanizers aim evading AI detection. To examine the robustness of AI detectors to such

"humanizers," we fed every AI passage through the StealthGPT default rewrite endpoint and re-scored the outputs of the four AI detectors.

Since the goal of humanizers is to avoid detection, the relevant statistic is the FNR[8] Pangram's performance is largely robust to the humanizer. For longer passages, Pangram detects nearly 100% of AI-generated text. The FNR increases a bit as the passages get shorter, but still remains low. The other detectors are less robust to humanizers. The FNR for Originality.AI increases to around 0.05 for longer text, but can reach up to 0.21 for shorter text, depending on the genre and LLM model. GPTZero largely loses its capacity to detect AI-generated text, with FNR scores around 0.50 and above across most genres and LLM models. RoBERTa does similarly poorly with high FNR scores throughout.

5 Assessing Tradeoffs

5.1 Detection Costs

Raw Costs (API Fees). Commercial APIs charge a fee per passage, with prices varying by genre and vendor, as summarized in Table B.5 and Table B.7 for shorter passages. Raw fees alone already tilt the economics in favor of Pangram. Averaging across genres, the per-call gap is stark, \$0.0227 for Pangram versus \$0.0415 for OriginalityAI and \$0.0569 for GPTZero. Similarly, the average cost for shorter passages for Pangram is \$0.0016, two times cheaper than OriginalityAI at \$0.0029 and almost three times cheaper than GPTZero at \$0.0040.

Cost-per-true-positive (CPTP). Raw costs alone do not inform about how many AI passages a detector actually catches, so we translate them into a payoff-adjusted metric—*cost per true positive*. Let c_g denote the API price per pass in genre g and FNR_g the genre-specific miss rate from Table 3. The expected *dollar cost per correctly detected AI passage* is

$$\text{CPTP}_g \;=\; \frac{c_g}{1-\text{FNR}_g},$$

so a lower CPTP means cheaper detection for the same true positive.[9] Converting fees to *cost-per-true-positive* (CPTP) sharpens the price gap.[10] Once each detector miss is charged to

[8] Table B.4 displays the FNR for stealth-GPT-generated texts for different detectors and LLM models.

[9] See details in Table B.5. In genres where a detector never misses (FNR $= 0$), CPTP equals the raw fee. If FNR $= 1$ (100% misses) the denominator is zero and the CPTP is economically undefined: One pays but can never detect.

[10] See details in Table B.6.

the numerator, Pangram remains the low-cost leader in all genres and on average: \$0.0228 per correctly flagged AI passage versus \$0.0416 for OriginalityAI and \$0.0575 for GPTZero, making Pangram the most cost-efficient detector for both full-length passages and stubs. Likewise, average cost per correctly identified AI short passage is the lowest for Pangram at \$0.0017, compared to Originality at \$0.0029, and GPTZero at \$0.0046.

5.2 Policy-Caps and Trade-offs

Previously, we calculated optimal thresholds for each detector by maximizing the difference between TPR and FPR. We used these detector-optimized thresholds to evaluate performance in Tables 2 and 3. We also assessed performance as a function of exogenously-set thresholds across detectors (see Tables 4 and 5). The former method produces different thresholds depending on the detector used, while the latter method fixes the caps across detectors. The former method prevents an apples-to-apples comparison of detectors because of the relationship between the FPR, FNR, and the threshold used; a threshold that minimizes FPR may be increasing the FNR above a tolerable level, or vice versa. Instead, a policy designer may want to choose a certain allowable FPR or FNR and then optimize the detector to minimize the other metric given that constraint. Fixing this constraint, which we refer to as a policy cap, also allows for an apples-to-apples comparison between detectors.

Why Use Policy-Caps? Every API allows the user to optimize a score threshold $\tau_{\text{raw}} \in [0, 1]$ in a way that is user-friendly and can be easily changed. We demonstrate this process in Section 3.1. However, this process is not ideal for a policy designer trying to choose whether to use an AI detector, and if so, which one. This involves making a choice that optimizes the designer's objective function that takes as its inputs the realized FPR and FNRs in the relevant context. The detector-optimized threshold is difficult use here for two reasons. First, each detector's scoring scale is idiosyncratic: the same numeric cut-off has a different operational meaning between AI detectors. Second, the detector-optimized threshold is estimated by maximizing over the FPR and FNR simultaneously. This makes it difficult to choose a detector based on an objective function that, for example, places more weight on one statistic (e.g., the FPR) than the other.

Instead, the designer can set a *policy cap* = x on a rate of her choice and then back-end adjust the detector's internal threshold so that the rate is less than x on a set of content entering the platform. In this section, we show that such policy caps can deliver more interpretable cross-detector comparisons.

A policy cap has three benefits. First, it is *scale-free*: it sets an upper bound on one error

rate (e.g., α for FPR) that is common across detectors, regardless of their internal score scales. Second, it allows policy designers to select a cap that matches their objective function, e.g., risk tolerance for Type I (or Type II) errors, without having to access the detector's proprietary algorithm. Third, this construction allows for a transparent separation between a *policy constraint* (the fixed cap) and the performance ranking, which allows for clearer differentiation between detectors.

Illustrative Example: FPR Policy Cap We now illustrate this exercise by imposing different FPR policy caps and reporting the resulting FNRs across models.

Did the detectors meet the caps? Table C.1 presents results on whether detectors comply to the caps imposed. We see that all detectors largely comply with the imposed policy caps. For example, Pangram realizes exactly 0.01 FPR when the policy cap is 0.01 and an FPR of 0.05 when the cap is 0.05

Comparing detectors. Table C.2 shows the resulting FNR rates at each FPR policy cap. Pangram has the best performance across the FPR caps imposed. For example, at the most conservative FPR cap of 0.0001, Pangram still achieves an FNR rate of around 0.01 on most LLM models. This FNR rate drops when the policy cap is loosened to 0.01. On the other hand, neither GPTzero nor Originality.AI do very well under the most stringent FPR policy cap, but begin to perform much better once that cap is loosened to 0.01. RoBERTa continues to miss the majority of AI passages even at the loosest cap.

Take-away. Reasonable FPR policy caps (0.005 to 0.01) leave the FNR rates of Pangram largely unchanged. Stringent policy caps degrade the performance of OriginalityAI and GPTZero. RoBERTa displays poor performance under any cap.

6 Conclusion

We evaluate the current batch of dominant AI-text detectors using a 1,992-passage multi-genre corpus in conjunction with four advanced LLMs. Three key findings are obtained.

First, a clear ranking in detectors emerges. Pangram maintains near-perfect accuracy across long and medium length texts. It achieves very low error rates even on shorter passages and 'stubs.' These results are robust to exogenous changes in the threshold and the use of 'humanizers' such as StealthGPT. The other two commercial detectors, Originality.AI and GPTZero, perform well on long and medium length texts, but, depending on the detector, struggle on short and 'humanized' passages. The open-source RoBERTa baseline both misclassifies human-written texts and fails to identify up to 51% of AI-generated

passages. Second, recalculating vendor fees as *cost-per-true-positive* (CPTP) shows that Pangram is substantially cheaper to implement than both OriginalityAI GPTZero. Third, the use of *policy caps* facilitates cross-detector comparisons and allows the policy designer to select a detector based on a straightforward optimization of their objective function.

While we have thus far focused on auditing existing AI text detectors, it is important to note that the implications of AI detection for writing and text-based work more generally are not obvious. LLMs are incredibly valuable tools that can facilitate idea generation and help tighten writing. At the same time, the use of LLMs to off-load a task where the receiver explicitly desires human input creates a host of agency problems. The use of AI text detectors in practice must thus strike a delicate balance to avoid discouraging the former while mitigating the issues posed by the latter. As highlighted in Section 1, setting more or less stringent policy caps is one way to navigate this trade-off. We believe that a more careful treatment of these issues requires a formal game theoretic treatment that we leave for future work.

References

Bick, Alexander, Adam Blandin, and David J Deming (2024). "The rapid adoption of generative AI." Working paper. National Bureau of Economic Research. [2]

Crothers, Evan N., Nathalie Japkowicz, and Herna L. Viktor (2023). "Machine-generated text: A comprehensive survey of threat models and detection methods." *IEEE Access* 11: 70979, -–71013. [2]

Hanley, Hans WA, and Zakir Durumeric (2024). "Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites." In *Proceedings of the international AAAI conference on web and social media*. vol. 18, 542–56. [2]

Russell, Jenna, Marzena Karpinska, and Mohit Iyyer (2025). "People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text." *arXiv preprint arXiv:2501.15654*, [2]

Tang, Ruixiang, Yu-Neng Chuang, and Xia Hu (2024). "The Science of Detecting LLM-Generated Text." *Communications of the ACM* 67 (4): 50–59. [2]

Tardelli, Serena, Marco Avvenuti, Maurizio Tesconi, and Stefano Cresci (2022). "Detecting inorganic financial campaigns on Twitter." *Information Systems* 103: 101769. [2]

Weber-Wulff, Debora, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington (2023). "Testing of detection tools for AI-generated text." *International Journal for Educational Integrity* 19 (1): 1–39. [2]

Appendix

A Tables for Descriptive Statistics

In the main analysis, we removed duplicates of texts across genres. Previously, there were 200 texts for each of the restaurant review and resume categories. After removing duplicates, there are now unique 100 restaurant reviews and 192 resumes. Sample sizes of other categories remain the same.

Table A.1: Descriptive Statistics

Model	Genre	N_H	N_{AI}	Mean words		TTR		FK grade	
				H	AI	H	AI	H	AI
GPT 4.1	amazon review	200	200	79.57	76.68	0.7863	0.8581	7.40	13.36
GPT 4.1	blog	200	200	198.98	197.11	0.7144	0.7896	6.24	9.95
GPT 4.1	news	300	300	459.88	458.40	0.6194	0.6629	10.95	14.81
GPT 4.1	novel	1000	1000	966.77	1037.75	0.5127	0.5191	44.79	9.56
GPT 4.1	restaurant review	100	100	133.08	129.39	0.7683	0.8262	6.07	11.06
GPT 4.1	resume	192	192	789.51	833.65	0.5329	0.5781	19.64	16.88
Claude Opus 4	amazon review	200	200	79.57	79.01	0.7863	0.8929	7.40	12.25
Claude Opus 4	blog	200	200	198.98	202.54	0.7144	0.8153	6.24	8.21
Claude Opus 4	news	300	300	459.88	450.59	0.6194	0.7014	10.95	14.71
Claude Opus 4	novel	1000	1000	966.77	982.35	0.5127	0.5632	44.79	8.04
Claude Opus 4	restaurant review	100	100	133.08	133.51	0.7683	0.8574	6.07	9.93
Claude Opus 4	resume	192	192	789.51	784.51	0.5329	0.6094	19.64	15.84
Claude Sonnet 4	amazon review	200	200	79.57	82.23	0.7863	0.8931	7.40	14.11
Claude Sonnet 4	blog	200	200	198.98	200.26	0.7144	0.8145	6.24	9.62
Claude Sonnet 4	news	300	300	459.88	435.94	0.6194	0.6929	10.95	16.54
Claude Sonnet 4	novel	1000	1000	966.77	952.40	0.5127	0.5587	44.79	10.32
Claude Sonnet 4	restaurant review	100	100	133.08	135.25	0.7683	0.8540	6.07	11.38
Claude Sonnet 4	resume	192	192	789.51	744.32	0.5329	0.5908	19.64	18.88
Gemini 2.0 Flash	amazon review	200	200	79.57	77.53	0.7863	0.8501	7.40	9.62
Gemini 2.0 Flash	blog	200	200	198.98	197.49	0.7144	0.7822	6.24	7.21
Gemini 2.0 Flash	news	300	300	459.88	430.54	0.6194	0.6478	10.95	12.91
Gemini 2.0 Flash	novel	1000	1000	966.77	970.35	0.5127	0.5246	44.79	8.39
Gemini 2.0 Flash	restaurant review	100	100	133.08	134.02	0.7683	0.8181	6.07	7.92
Gemini 2.0 Flash	resume	192	192	789.51	787.84	0.5329	0.5424	19.64	13.73

Notes: H = Human; TTR = Type–Token Ratio; FK = Flesch–Kincaid. Counts refer to unique passages after de-duplication and pre-2020 filtering. The AI sample size exactly matches the human count by construction ($N_{\text{AI}} = N_{\text{H}}$=1,992).

Note that FK grade for "Novels" is high because there are some old novels that do not have full stops. Consequently, the number of sentences is 1 for some samples, hence the calculation for either FK grade is inflated.

B Tables for Robustness Tests

B.1 Additional Tables for Stubs

Table B.1: False Positive Rates by Genre (Passages < 50 Words)

Genre	GPTZero	Pangram	RoBERTa
amazon review	0.0299	0.0187	0.2948
blog	0.0172	0.0000	0.4871
news	0.0000	0.0000	0.3611
restaurant review	0.0333	0.0250	0.3667

Table B.2: Detector Performance by Genre and Model: FNR—Passages < 50 Words

Model	Genre	GPTZero	Pangram	RoBERTa
GPT-4.1	amazon review	0.0149	0.0000	0.4179
GPT-4.1	blog	0.0000	0.0172	0.0862
GPT-4.1	news	0.1111	0.0000	0.0000
GPT-4.1	restaurant review	0.0667	0.0000	0.2667
Claude Opus 4	amazon review	0.1194	0.0149	0.3731
Claude Opus 4	blog	0.1034	0.0000	0.1552
Claude Opus 4	news	0.4444	0.2222	0.4444
Claude Opus 4	restaurant review	0.1667	0.0000	0.1667
Claude Sonnet 4	amazon review	0.0597	0.0000	0.1343
Claude Sonnet 4	blog	0.0517	0.0000	0.2241
Claude Sonnet 4	news	0.0000	0.0000	0.4444
Claude Sonnet 4	restaurant review	0.0667	0.0333	0.1000
Gemini 2.0 Flash	amazon review	0.1343	0.0149	0.4179
Gemini 2.0 Flash	blog	0.1379	0.0345	0.1207
Gemini 2.0 Flash	news	0.3333	0.1111	0.0000
Gemini 2.0 Flash	restaurant review	0.2000	0.0667	0.3667

Table B.3: Detector Performance by Genre and Model: AUROC—Passages < 50 Words

Model	Genre	GPTZero	Pangram	RoBERTa
GPT-4.1	amazon review	0.9776	0.9984	0.6195
GPT-4.1	blog	0.9914	0.9982	0.7298
GPT-4.1	news	0.9444	1.0000	0.5679
GPT-4.1	restaurant review	0.9500	0.9978	0.6789
Claude Opus 4	amazon review	0.9254	0.9944	0.7603
Claude Opus 4	blog	0.9397	1.0000	0.6822
Claude Opus 4	news	0.7778	0.9630	0.7160
Claude Opus 4	restaurant review	0.9000	0.9956	0.6956
Claude Sonnet 4	amazon review	0.9552	0.9969	0.7636
Claude Sonnet 4	blog	0.9655	1.0000	0.7574
Claude Sonnet 4	news	1.0000	1.0000	0.7531
Claude Sonnet 4	restaurant review	0.9500	0.9989	0.7900
Gemini 2.0 Flash	amazon review	0.9179	0.9951	0.7117
Gemini 2.0 Flash	blog	0.9224	0.9979	0.6742
Gemini 2.0 Flash	news	0.8333	0.9877	0.6296
Gemini 2.0 Flash	restaurant review	0.8833	0.9856	0.6922

B.2 Additional Tables for Humanizers

Table B.4: Detector Performance by Genre and Model for Stealth Generated Text: FNR

Model	Genre	GPTZero	Originality	Pangram	RoBERTa
GPT-4.1	amazon review	0.6784	0.0888	0.0250	0.5650
	blog	0.4450	0.0227	0.0000	0.1550
	news	0.5067	0.0743	0.0200	0.9400
	novel	0.4474	0.0240	0.0000	0.4310
	restaurant review	0.6600	0.1905	0.0100	0.4800
	resume	0.6354	0.0573	0.0000	0.1510
Claude Opus 4	amazon review	0.6700	0.0414	0.0400	0.5500
	blog	0.5850	0.0414	0.0050	0.8200
	news	0.5667	0.0438	0.0367	0.5667
	novel	0.7730	0.0292	0.0000	0.4540
	restaurant review	0.6400	0.0353	0.0200	0.2300
	resume	0.6562	0.0469	0.0000	0.2812
Claude Sonnet 4	amazon review	0.6350	0.0702	0.0250	0.5800
	blog	0.5276	0.0349	0.0101	0.8241
	news	0.4448	0.0642	0.0268	0.9498
	novel	0.6329	0.0414	0.0000	0.4750
	restaurant review	0.6700	0.0482	0.0100	0.2500
	resume	0.6927	0.0990	0.0000	0.2552
Gemini 2.0 Flash	amazon review	0.6400	0.0625	0.0500	0.4300
	blog	0.5250	0.0299	0.0200	0.8100
	news	0.3533	0.0441	0.0133	0.7100
	novel	0.4369	0.0366	0.0000	0.4410
	restaurant review	0.5600	0.2073	0.0100	0.4800
	resume	0.2552	0.0573	0.0000	0.2448

B.3 Tables for the Economic Costs of Detection

Table B.5: Per-passage API fee (USD) by genre and detector

Genre	Pangram	OriginalityAI	GPTZero
Amazon review	$0.00383	$0.00698	$0.00959
Blog post	$0.00986	$0.01794	$0.02464
News article	$0.02292	$0.04171	$0.05730
Novel excerpt	$0.05189	$0.09444	$0.12972
Restaurant review	$0.00647	$0.01177	$0.01617
Résumé fragment	$0.04168	$0.07586	$0.10421
Average (6 genres)	**$0.02277**	**$0.04145**	**$0.05694**

Table B.6: Cost per *true* AI detection (USD) by genre

Genre	Pangram	OriginalityAI	GPTZero
Amazon review	$0.00385	$0.00704	$0.00991
Blog post	$0.00989	$0.01806	$0.02524
News article	$0.02296	$0.04189	$0.05773
Novel excerpt	$0.05189	$0.09462	$0.13096
Restaurant review	$0.00652	$0.01217	$0.01685
Résumé fragment	$0.04168	$0.07586	$0.10421
Average (6 genres)	**$0.02280**	**$0.04161**	**$0.05748**

Table B.7: Per-passage API fee (USD) for short passage (<50 words) by genre and detector

Genre	Pangram	OriginalityAI	GPTZero
Amazon review	$0.00163	$0.00296	$0.00407
Blog post	$0.00155	$0.00282	$0.00388
News article	$0.00178	$0.00324	$0.00444
Restaurant review	$0.00146	$0.00266	$0.00366
Average (4 genres)	**$0.00161**	**$0.00292**	**$0.00401**

Table B.8: Cost per *true* AI detection (USD) for short passage (<50 words) by genre

Genre	Pangram	OriginalityAI	GPTZero
Amazon review	$0.00164	$0.00296	$0.00443
Blog post	$0.00157	$0.00282	$0.00419
News article	$0.00194	$0.00324	$0.00571
Restaurant review	$0.00150	$0.00266	$0.00418
Average (4 genres)	**$0.00166**	**$0.00292**	**$0.00463**

C Tables for Policy Caps

Table C.1: False-positive Rates Across Models at Candidate Policy Caps.

Detector	0.0001	0.005	0.01	0.05	0.10
gptzero	0.0000	0.0000	0.0071	0.0071	0.0071
originality	0.0005	0.0055	0.0104	0.0476	0.0876
pangram	0.0005	0.0050	0.0100	0.0502	0.1003
roberta-base	0.0005	0.0050	0.0100	0.0497	0.1024

Table C.2: False-negative rates at candidate FPR caps

		Policy caps (FPR target)				
Model	**Detector**	0.0001	0.005	0.010	0.050	0.100
FNR						
GPT-4.1	gptzero	1.0000	1.0000	0.0020	0.0020	0.0020
	originality	0.2873	0.0718	0.0362	0.0000	0.0000
	pangram	0.0095	0.0010	0.0005	0.0000	0.0000
	roberta-base	1.0000	0.9985	0.9955	0.9809	0.9418
Claude Opus 4	gptzero	1.0000	1.0000	0.0291	0.0291	0.0291
	originality	0.5129	0.2304	0.1368	0.0077	0.0000
	pangram	0.0176	0.0035	0.0015	0.0005	0.0005
	roberta-base	1.0000	0.9965	0.9945	0.9679	0.9277
Claude Sonnet 4	gptzero	1.0000	1.0000	0.0085	0.0085	0.0085
	originality	0.3215	0.0963	0.0539	0.0027	0.0005
	pangram	0.0110	0.0010	0.0005	0.0000	0.0000
	roberta-base	1.0000	0.9980	0.9975	0.9794	0.9433
Gemini 2.0 Flash	gptzero	1.0000	1.0000	0.0146	0.0146	0.0146
	originality	0.0749	0.0083	0.0039	0.0011	0.0000
	pangram	0.0211	0.0035	0.0030	0.0010	0.0010
	roberta-base	1.0000	0.9975	0.9965	0.9829	0.9383

D Detector APIs and Built-in Probability Fields

- **Pangram** (API v2025-05): probability field `prob_ai`.
- **GPTZero** (API v2): `average_generated_prob`.
- **OriginalityAI** ("Turbo" endpoint): `probability_ai` returned on a 0–100 scale; we divide by 100.
- **ZeroGPT** (API v1): `ai_percentage` / 100.
- **RoBERTa-base, open source**: the 2020 OpenAI detector fine-tuned on public GPT-2 outputs; we keep the soft-max probability of the AI class.

E AI Models and Prompt for AI-Generated Texts

E.1 AI Models Versions

- "GPT-4.1", "id": "gpt-4.1-2025-04-14", "provider": "openai"

- "Claude Opus 4", "id": "claude-opus-4-20250514", "provider": "anthropic"
- "Claude Sonnet 4", "id": "claude-sonnet-4-20250514", "provider": "anthropic"
- "Gemini 2.0 Flash", "id": "gemini-2.0-flash", "provider": "google",

E.2 Prompt Engineering

SYSTEM + USER PROMPT SUPPLIED TO `gpt-3.5-turbo`

```
You are a writing assistant.

Write an original passage on the topic: {topic}.
Target length: {word_count} words.

Be clear and human-like; avoid copying or referencing specific
texts.

Do not include the topic or these instructions in your output.
Return only the generated passage.
```

E.3 AI Error Rate: Example

Figure E.1: Example of OriginalityAI API return

genre	source	topic	ai_generated	stealthgpt_rephrased	human_verdict	ai_verdicts
novel	Haggard_Mines_1885.txt	Discovery of da Silvestra's remains, relics, and map; survival struggle in the mountains; hunting antelope for food; life-or-death decisions with rifles	{'gpt-4.1-2025-04-14': 'The sun was already arcing low behind the ragged peaks when we stumbled upon what was left of da Silvestra. We had been following the ridge for hours, boots scraping over scree and shale, lungs burning in the thin air. It was Anton who saw the glint—something unnatural among the gray stones—a buckle, perhaps, or the edge of a battered brass compass. We all stopped, the ...	{'gpt-4.1-2025-04-14': 'When we finally found what remained of da Silvestra, the sun was already a low arc behind the jagged mountaintops. For hours we had been making our way along the ridge, boots scraping over scree and shale, our lungs burning from the thin air. Anton was the first to notice something off in the coloring of the stones—weirdly grayish, instead of just gray—the sign of a cam...	{'pengram': {'ai_likelihood': 0.0001280837588840061, 'prediction': 'Unlikely AI'}, 'originality': {'error': '422 Client Error: Unprocessable Entity for url: https://api.originality.ai/api/v3/scan'}, 'gptzero': {'error': '429 Client Error: Too Many Requests for url: https://api.gptzero.me/v2/predict/text'}, 'roberta-base-detector': 0.9998247027397156}	{'gpt-4.1-2025-04-14': {'pengram': {'ai_likelihood': 1.0, 'prediction': 'Highly Likely AI'}, 'originality': {'error': '422 Client Error: Unprocessable Entity for url: https://api.originality.ai/api/v3/scan'}, 'gptzero': {'error': '429 Client Error: Too Many Requests for url: https://api.gptzero.me/v2/predict/text'}, 'roberta-base-detector': 0.9998149275779724}}

F Human and AI-Generated Writing Examples

F.1 Original Human texts: Normal and Short

F.1.1 Normal Length Human Text

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The proposed training, which would have been provided by volunteers at no cost to the state, would occur during orientation for legislators at the beginning of each session. The bill was not prompted by the Dakota Access Pipeline protests, but several who testified in support of the training said better cultural understanding could have prevented some of the conflicts that arose during the protests.The Senate Government and Veterans Affairs Committee amended the bill to a legislative management study to consider the need for cultural competency training for legislators, other elected and appointed officials and state employees. Sen. Shawn Vedaa, R-Velva, a member of the committee, said Tuesday, Feb. 14, the bill was amended to a study because several committee members felt requiring the training "was overstepping legislation."Sen. Dick Dever, R-Bismarck, spoke in favor of the bill as a way to repair relationships that have been strained during the pipeline protests."I think there have been damages done to the relationships between our general population and the population south of here through recent events," Dever said.The Senate voted to amend the bill to a study, but ultimately the bill failed in a 20-26 vote on Tuesday.Sen. Richard Marcellais, D-Belcourt, a member of the Turtle Mountain Band of Chippewa and one of the bill's sponsors, said he is disappointed with the vote."It's not going to improve the communications or relations between the state and tribes," he said.Marcellais sponsored similar legislation in 2009 that also failed in the Senate with a similar vote.

F.1.2 Short Length Human Text

"18,NOVEMBER,2002"; "BLOG"; https://huggingface.co/datasets/barilan/blog_authorship_corpus

‘Thought- OK...so, I’m all for midgets and whatnot. I’m not all, "down with midgets," but really, don’t you think it’d be rude to say, "Up with midgets?" Ya di amore ‘

F.2 AI-Generated Texts: Normal and Short

F.2.1 GPT-4.1 Versions

GPT-4.1 Version: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

In the wake of heightened tensions between the state of North Dakota and its tribal nations following the Dakota Access Pipeline protests, lawmakers considered a bill intended to address persistent misunderstandings and improve relationships. The original measure called for mandatory cultural competency training for all state lawmakers and officials, aiming to foster greater awareness of tribal histories, treaties, and contemporary issues. Proponents argued that such training was essential for building trust and ensuring informed decision-making, especially after the discord and miscommunication that arose during the protests. Despite strong support from Native American leaders and some legislators, the proposal faced skepticism from others who questioned the necessity and logistics of mandatory training. Critics expressed concerns about government overreach and the effectiveness of compulsory sessions, suggesting that respect and understanding could not be legislated. After contentious debate, the bill was amended, softening its requirements by replacing the mandate with a recommendation to study the feasibility and potential benefits of cultural competency programs. Even with the amendment, the proposal struggled to gain traction. Opponents maintained their reservations, and some viewed the study as an unnecessary step that would not lead to tangible change. When

brought to a Senate vote, the bill failed to pass, effectively halting the initiative. The rejection disappointed advocates who saw an opportunity to bridge divides, while its defeat underscored ongoing challenges in addressing the complex relationship between North Dakota's government and its tribal communities.

GPT-4.1 Version: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

Choosing respectful language is important when referring to individuals of short stature. Many prefer the term "little people" or "person with dwarfism" to promote dignity and understanding.

F.2.2 Claude Opus 4 Versions

Claude Opus 4 Version: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The North Dakota Senate chamber fell silent as lawmakers cast their votes on a measure that had stirred considerable debate. The proposed legislation, which would have mandated cultural competency training for all state legislators, ultimately failed by a margin of 20-26. Even an amended version, scaled back to merely establish a study committee on the issue, could not garner sufficient support. The bill emerged from ongoing tensions that have marked North Dakota's political landscape since the Dakota Access Pipeline protests brought national attention to the state. Advocates argued that mandatory training would help bridge the divide between state government and tribal nations, fostering better understanding of Native American history, sovereignty, and contemporary challenges. They pointed to miscommunications and cultural misunderstandings during the pipeline controversy as evidence of the need for such education. Opponents, however, viewed the requirement as government overreach. Some senators expressed concerns about the precedent of mandating specific training for elected officials, while others questioned who would develop and deliver such programs. The fiscal impact also drew criticism, with estimates suggesting significant costs for implementing comprehensive training across

the legislature. The defeat represents a setback for those seeking to improve state-tribal relations through institutional change. Tribal leaders expressed disappointment, noting that voluntary efforts had proven insufficient in the past. Meanwhile, the bill's sponsors vowed to continue pursuing alternative approaches to foster cultural understanding. As North Dakota moves forward, the challenge of building meaningful dialogue between state government and tribal nations remains unresolved, leaving both sides to seek new pathways toward mutual understanding and cooperation

Claude Opus 4 Version: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The protest signs reading "Down with discrimination" confused matters when shortened to "Down with little people" - an unfortunate abbreviation that seemed to contradict the supportive message intended for those with dwarfism.

F.2.3 Claude Sonnet 4 Versions

Claude Sonnet 4 Version: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The North Dakota Senate chamber witnessed heated debate Tuesday as lawmakers grappled with legislation requiring cultural competency training for state representatives. The original bill, which would have mandated diversity and inclusion education for all legislators, faced significant pushback from conservative members who argued it represented government overreach. Senator Maria Rodriguez, the bill's primary sponsor, emphasized the importance of preparing lawmakers to serve North Dakota's increasingly diverse population. "Our constituents deserve representatives who understand the communities they serve," she stated during floor debate. The training would have covered topics including implicit bias, cultural awareness, and effective communication across different demographic groups. Opposition senators raised concerns about mandatory participation and questioned the necessity of formal training. Senator James Mitchell argued that elected

officials already possess the judgment needed to represent their districts effectively. "We don't need bureaucratic seminars to teach us how to do our jobs," he declared. Recognizing the bill's uncertain fate, supporters attempted a compromise amendment that would establish a study committee to examine cultural competency needs rather than implementing immediate training requirements. This softer approach aimed to gather data and build consensus for future legislative sessions. Despite the amendment's more moderate stance, the modified legislation still failed to gain sufficient support. The final tally of 20 votes in favor and 26 against reflected the deep philosophical divide within the chamber regarding government-mandated diversity initiatives. The defeat leaves North Dakota without formal cultural competency requirements for its lawmakers, continuing ongoing statewide discussions about inclusion and representation in government

Claude Sonnet 4 Version: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

I understand youŕe looking for wordplay, but Id́ prefer not to create jokes that could be hurtful toward people with dwarfism. Id́ be happy to help you craft other types of puns or wordplay instead - perhaps something with "up" and "down" phrases

F.2.4 Gemini Flash 2.0 Versions

Gemini Flash 2.0 Version: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The North Dakota Senate chamber buzzed with restrained energy as the vote tally flashed across the screen. A bill mandating cultural competency training for state legislators had just failed, replaced instead with a measure calling for a comprehensive study on the topic. The proposed training, intended to foster understanding and empathy towards diverse communities within the state, had sparked weeks of heated debate. Several senators voiced concerns that the training, however well-intentioned, represented an overreach of legislative authority. They questioned

the curriculum, the potential for biased instruction, and the overall effectiveness of mandated sensitivity programs. Some argued that genuine understanding couldn't be forced through legislation. The shadow of the Dakota Access Pipeline protests loomed large over the discussion. The contentious standoff, years prior, had left deep scars on the relationships between the state government and Indigenous communities. Some senators worried that mandating cultural competency training, in the wake of such division, would be perceived as a performative gesture, further straining already fragile connections. The approved study, proponents argued, would allow for a more nuanced and informed approach. It would involve gathering data on existing cultural awareness initiatives, identifying specific needs within the legislature, and exploring best practices from other states. The hope was that a thorough understanding of the issue would pave the way for a more effective, and less divisive, solution. Whether this study would ultimately lead to meaningful change, or simply delay action, remained to be seen

Gemini Flash 2.0 Version: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

My brain short-circuits. "Great...vertically-challenged...human?" I stammer, cringing. "Up with midgets!" suddenly seems less a toast, more a lawsuit waiting to happen.

F.3 AI-Humanized Versions of AI-Generated texts

F.3.1 AI-Humanized Versions of GPT-4.1

AI-Humanized Versions of GPT-4.1: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

After protests against the Dakota Access Pipeline led to tensions between North Dakota and its tribal nations, lawmakers took up a bill aimed at improving relations and addressing misunderstandings. The original measure mandated cultural competency training for all state lawmakers and officials, targeting an intimate and

thorough unveiling of the significant histories, treaties, and current sovereign nation issues that the tribes would want any state manager to understand. Proponents framed the necessity for such training as essential to trust building, and a precondition for informed decisions in light of the significant tribal issues that might come up. Even with the firm backing of Native American leaders and certain legislators, the proposal met with skepticism from a number of quarters. Some questioned not so much the motivation for the effort—everyone agreed that it would not be a bad thing for all law enforcement officers to have a better understanding of and respect for Native cultures—but they did take issue with the necessity and logistics of mandating the kind of comprehensive, in-depth training that seemed to be called for. They worried it might just be another instance of government overreach and didn't seem to have much faith in the idea that what was being proposed could actually work. Even with the amendment, the proposal found it difficult to get traction. The folks who opposed the measure maintained their reservations, and some viewed the proposal as an unnecessary step not likely to lead to tangible change. When it came up for a Senate vote, the bill failed to pass, effectively halting the initiative. Advocates of the amendment were disappointed, because they had viewed its passage as an opportunity to bridge some divides

AI-Humanized Versions of GPT-4.1: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

When choosing language, itś important to be respectful—especially when referring to individuals of short stature. The Council of Mayo Clinic in 2017 recommended always using the phrases or terms "little people" and "person with dwarfism." They argue these ways of being descriptive are more dignified than past or potentially ambiguous terminology.

F.3.2 AI-Humanized Versions of Claude Sonnet 4

AI-Humanized Versions of Claude Sonnet 4: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The State Senate held a contentious debate Tuesday over a bill requiring cultural competency training for state leaders. Lawmakers had to wrestle with the issue. And they weren't alone. Testifying before the Senate Education Committee in opposition were members of a prominent conservative organization Whitney Mangus, a part-time state employee who leads the training and development division, said only two other Basin cities had had the experience of being part of the training. Train was to be used in that context, not to train up, but to show whose lives were to be accepted and whose not. Senator Maria Rodriguez, the primary sponsor of the bill, highlighted the significance of equipping lawmakers to represent North Dakota's ever-diversifying populace. "The constituents that I represent deserve to have representatives who understand the communities that they serve," she said during the floor debate. The training, as proposed, would have covered such subjects as implicit bias, cultural proficiency, and effective communication across the many different demographic groups that make up our state. Senators in opposition expressed misgivings about making participation compulsory and doubted whether the program's formal training was really needed. Senator James Mitchell said he thought the training was unnecessary because he believed that the kind of people who get elected to public office are just inherently good judges of the public's interest. "We don't need bureaucratic seminars to teach us how to do our jobs," he declared. Acknowledging the unclear destiny of the bill, Allies attempted a compromise amendment establishing a study committee to assess the cultural competency range of this bill. This softer approach works to gather data and build a coalition for the next session of the legislature. Even with this more moderate approach, the amended legislation could not garner enough support to pass. When all was said and done, the vote tally came to 20 in favor, 26 against. That reflects, quite clearly, the deep philosophical divide within the chamber over whether the federal government ought to be mandating diversity. North Dakota lawmakers aren't required to be culturally competent,

thanks to a recent court loss. The state remains without any formal requirements for inclusion and representation in government, even as the aforementioned factors are part of ongoing conversations in the state.

AI-Humanized Versions of Claude Sonnet 4: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

I get that you're after some witty wordplay, but I'd rather not come up with gags that might offend folks with dwarfism. If you're set on using that type of humor, then I'm not the person to help you. On the other hand, if you're interested in creating puns or wordplay that are, like, the opposite of offensive, then hit me up. I'd be more than happy to help you with that.

F.3.3 AI-Humanized Versions of Claude Opus 4

AI-Humanized Versions of Claude Opus 4: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The North Dakota Senate chamber was silent as lawmakers voted on a measure that had stirred considerable debate. The proposed legislation mandated cultural competency training for all state legislators and ultimately failed by a margin of 20-26 and an amended version, scaling back to merely establishing a study committee on the issue, also could not garner sufficient support. They could not even get a committee going to look into the problem. Tensions have marked North Dakota's political landscape since the Dakota Access Pipeline protests brought the state national attention, and the bill is a product of that discord. Advocates argue it would help bridge the divide between state government and tribal nations, fostering a better understanding of Native American history, sovereignty, and contemporary challenges. They point to the pipeline controversy and cultural misunderstandings that arose during it as evidence for the need for such education. Critics, however, perceived the mandate as governmental overreach. Concerns were expressed about the precedent of requiring specific training for people in public office. Who, that was also a

question, would develop and deliver the training? And then could the taxpayers foot the bill? Estimates suggested that implementing comprehensive training all over the place—that is, across the whole legislature—could cost a pretty penny. The defeat marks a step back for individuals who are trying to improve the relations between the state and tribes through institutional change. Disappointment was expressed by tribal leaders, who noted that in the past, voluntary efforts to improve relations had not worked. Meanwhile, two sponsors of the bill stated that they would pursue other avenues in the future to try to improve the understanding of tribal culture. Building meaningful dialogue between state government and tribal nations in North Dakota today feels like a challenge yet to be met, going on a decade since handfuls of North Dakota American Indians first raised their concerns to state leadership over go-it-alone decisions that affect tribal nations. Mutual understanding and cooperation appear to be the objectives both sides are seeking and seem to be at the moment. But how to get there is an unanswered question.

AI-Humanized Versions of Claude Opus 4: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The signs reading "Down with discrimination" confused things when they were shortened to "Down with little people" - an unfortunate abbreviation that seemed to contradict the supportive message intended for those with dwarfism.

F.3.4 AI-Humanized Versions of Gemini Flash 2.0

AI-Humanized Versions of Gemini Flash 2.0: Normal Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

The Senate chamber in North Dakota buzzed with contained energy when the vote tally appeared onscreen. A bill requiring cultural competency training for state legislators had just gone down to defeat, replaced instead with a mandate to do a comprehensive study of the topic. The proposed training was intended to promote understanding and empathy toward the multitude of different

communities that make up North Dakota and seemed like a no-brainer. Yet the very idea of training had sparked several weeks' worth of animated discussion. Some of that discussion could even be classified as argument. Many senators were concerned that the training, as well-intentioned as it might be, amounted to an overreach of legislative authority. They questioned: The curriculum: What are the trainers teaching? The potential for biased instruction: Who is doing the teaching? The overall effectiveness of the programs: Do these kinds of sessions actually accomplish anything? Some went so far as to suggest that if Congress truly wanted to get to the bottom of the divisive issues facing our nation, it ought to do just the opposite of what it was mandating: hold training sessions in which all participants would have the freedom to speak candidly about their feelings on the hot-button issues of the day. The Dakota Access Pipeline protests cast a long shadow over the conversation. The years-long standoff had left a number of appearances and accusations on both sides. Many senators were worried about making bad relationships worse, especially when some of their own Democratic colleagues appeared to be hostile to anything in favor of law enforcement, as had happened in the past. Supporters of the approved study argued that it would permit a more detailed and clearer way to proceed. They contended that it would lead to the collection of data on different kinds of existing cultural awareness initiatives, the identification of the specific kinds of cultural needs that existed within the legislature, and the examination of which practices from other states worked best. They hoped that by thoroughly understanding the problem, they could arrive at a more effective, less divisive solution.

AI-Humanized Versions of Gemini Flash 2.0: Short Length

YEAR:2014; GENRE: "NEWS"; SOURCE: https://huggingface.co/datasets/cc_news

I short-circuit my brain. "Vertically challenged humans?" I stammer and cringe inside. "Up with midgets!" suddenly seems less a toast than a possible lawsuit waiting to happen.

G Scientific Transparency: Replication Note

The complete code, data, and step–by–step README are archived on GitHub https://github.com/brianjabarian/DetectionAI. The repository contains:

- `/all_code/` – Jupyter notebooks that reproduce every table and figure, including `200 - DetectionAI - Statistics.ipynb`, which builds the headline performance tables in one click.
- `/all_json/` – placeholders for the raw human passages and the paired AI-generated rewrites.
- `results/` – CSVs generated by the notebooks; the paper's tables are direct copies of these files.

System requirements. All scripts run under Python 3.10+ with the packages listed in `requirements.txt`. A minimal replication can be executed with:

```
git clone https://github.com/brianjabarian/DetectionAI.git
cd DetectionAI
pip install -r requirements.txt  % installs pandas, sklearn, torch, etc.
```

Quick start. After installation, open `notebooks/Code_statistics_part.ipynb` and run all cells; this notebook loads the JSON dataset and writes the summary tables to `results/`. Running the remaining notebooks reproduces the short-text stress tests and all robustness checks. Researchers may reproduce every number in the paper by executing the notebooks in order; no manual tweaking or proprietary data is required.

H Full Replication with GPT-3.5

H.1 Descriptive Statistics

Synthetic counterparts are produced with `gpt-3.5-turbo` at the same temperature 0.7 as for the four other main AI models and a length cap set to twice the source length, up to 2,048 tokens (snapshot, May 15, 2025). We used the same prompt as for the other models available in the Appendix E. Table H.1 provides the same summary of the human and AI corpus as for the other AI models.

Table H.1: Composition of the DetectionAI Benchmark

Genre	N_H	N_{AI}	Mean words H	Mean words AI	TTR-H	TTR-AI	FK grade-H	FK grade-AI
News (CC-News)	300	300	473.88	387.38	0.5578	0.5510	10.79	13.53
Blogs	200	200	198.98	192.35	0.6299	0.7125	6.24	10.65
Résumés	200	200	800.06	750.05	0.4758	0.4717	19.67	15.27
Yelp reviews	200	200	133.08	139.73	0.7396	0.7917	6.07	11.39
Amazon reviews	200	200	79.57	80.25	0.7597	0.8113	7.40	13.33
Novels (pre-2000)	1000	1000	1018.09	621.26	0.4159	0.4306	46.51	9.87

Notes: H = Human; TTR = Type–Token Ratio; FK = Flesch–Kincaid. Counts refer to unique passages after de-duplication and pre-2020 filtering. The AI sample size exactly matches the human count by construction (N_{AI} = N_H=2,100).

H.2 Baseline and Genre Heterogeneity

Overall Baseline Table H.2 shows that two commercial ensembles, *Originality AI* and *Pangram*, achieve perfect separation of AI and human passages (AUROC = 1.00). Their difference in mean classification accuracy is below five basis points and statistically indistinguishable from zero under our genre-clustered bootstrap (95% CI: –0.014 ; 0.023). *GPTZero* trails by less than one percentage point (AUROC = 0.993), while the open-source *RoBERTa* baseline performs barely above chance (AUROC $<$ 0.273), misclassifying more than a third of passages even at its Youden-optimal policy targeted cap.

Table H.2: Overall Detector Performance

Detector	AUROC	Δ-Mean
Originality AI	1.000	0.996
Pangram	1.000	0.991
GPTZero	0.993	0.987
RoBERTa baseline	0.275	−0.335

Heterogeneity across genres. The accuracy of the decomposition by textual domain in Table H.3 reveals that the perfect performance of *Originality AI* and *Pangram* extends to all genres, including nineteenth- and twentieth-century fiction. *GPTZero* displays mild variation, its entropy heuristic copes least well with short, opinion-laden consumer reviews (AUROC ≈ 0.98) and best with literary prose (AUROC ≈ 0.999). In contrast, the detector *RoBERTa* collapses when stylistic conventions diverge from its training data, recording AUROC = 0.13 in novels and never exceeding 0.60 in any genre. The wide gap between

commercial ensembles and open-source baselines thus persists across all content types, cautioning institutions that lean on cost-free detectors in settings rich in stylistic variation.

Table H.3: Detector Performance by Genre

Genre	GPTZero	Originality	Pangram	RoBERTa
Amazon review	0.980	1.000	1.000	0.558
Blog	0.985	1.000	1.000	0.407
News	0.997	1.000	1.000	0.197
Novel	0.999	1.000	1.000	0.128
Restaurant review	0.980	1.000	1.000	0.520
Résumé	0.998	1.000	1.000	0.474

H.3 Performance on Short Texts

Originality AI is omitted for GPT-3.5 because the vendor's API refused to score texts less than 50 words, raising a length-filter exception; we treat these outputs as missing rather than zeros.

Table H.4: Detector Performance on Passages < 50 Words

Detector	AUROC	Δ-Mean
Pangram	0.999	0.904
GPTZero	0.969	0.938
RoBERTa baseline	0.593	0.115

H.4 AI-Generated Synthetic Writing for GPT-3.5

H.4.1 Normal Length AI-Generated Version

The AI-generated version of the previous human original version is below:

GPT-3.5-Generated Version

```
"In a recent legislative session, a well-intentioned effort to
bridge the gap between state lawmakers and tribal nations fell
short of its initial goal. The proposed bill sought to establish
a comprehensive program of free cultural competency training for
legislators. Proponents of the bill argued that such an initiative
was crucial for fostering a deeper understanding of the unique
histories, cultures, and governance structures of the indigenous
```

peoples within the state.The training was envisioned as a tool to equip lawmakers with the necessary insights to craft policies that would honor tribal sovereignty and promote collaboration on issues affecting tribal lands and communities. Advocates of the bill believed that enhanced cultural awareness among legislators could lead to more informed decision-making and mitigate long-standing tensions between state and tribal governments.Despite its noble aspirations, the bill faced considerable resistance. Critics cited concerns over the cost implications and the effectiveness of the proposed training program. After vigorous debates and numerous committee reviews, the bill's original mandate was significantly diluted. In its place, lawmakers agreed to a compromise: an amendment to conduct a comprehensive study on the potential impact and feasibility of such training.The revised measure tasked a joint committee with exploring the benefits and challenges of implementing cultural competency training. The committee was to examine best practices from other states, consult with tribal leaders, and consider the logistical aspects of providing training to a diverse legislative body.While the original bill's failure was a disappointment to those who had championed its cause, the amended resolution offered a glimmer of hope. The study, if conducted thoroughly and with genuine input from all stakeholders, could lay the groundwork for future initiatives. It had the potential to pave the way for a more culturally responsive legislative process, one that recognized the importance of understanding and respecting the rich tapestry of indigenous cultures within the state."

H.4.2 Normal Length AI-Humanized Version of AI-Generated Version

The AI-humanized of the AI-generated version of the previous human original version is below:

StealthGPT-Generated Version

'This recent legislative session saw a good-faith effort aimed at bridging the gap between state lawmakers and tribal nations. An initial push was made to establish a comprehensive program of free

cultural competency training for legislators, but the measure failed to pass a key committee. Proponents of the bill argued for its necessity, saying it was crucial to achieving a deeper understanding of the unique histories, cultures, and governance structures of the indigenous peoples within the borders of the state.The training was envisioned as a tool to give lawmakers the insights necessary to craft policies that would honor tribal sovereignty and, as a result, promote collaboration on issues affecting tribal lands and communities. Advocates of the bill believed that enhanced cultural awareness among legislators could lead to a lot less pipe decision-making (I think the word you're looking for here is ". . . could presage a future of more informed decision-making") and possibly avoid long-standing tensions between state and tribal governments. Lawmakers were pictured as better equipped to fulfill their end of the honor deal if they had a bit more humility and a lot more insight.Even with its high-minded goals, the bill ran into a good deal of opposition. Critics raised alarms about the possible price tag and the basic effectiveness of the proposed training. After lots of back-and-forth and a good number of committee reviews, the bill's overall mandate was cut back in a meaningful way. In its place, lawmakers signed off on a compromise: an amendment to do a study that would look in depth at the possible impacts of the training and its feasibility.The amended measure assigned a joint committee to look into the pros and cons of mandating cultural competency training for legislators. The committee was directed to study what works and what doesn't in other states, to confer with knowledgeable tribal leaders, and to think through the many challenges in training a body as diverse as our legislature.Though it was disappointing for those who had fought valiantly for the original bill, the resolution—amended at the last minute—offered a glimmer of hope. If done thoroughly and with genuine input from all the stakeholders, the promised study could serve as a kind of prelude for future initiatives. Its potential seemed to be paving the way for a much more responsive kind of process, one that would have the same kind

```
of responsiveness to the indigenous cultures within the state that
any legislation affecting that group ought to have.‘
```

H.4.3 Short AI-Generated Version

The AI-generated version of the previous human original version is below:

GPT-3.5-Generated Version

```
’The term "midgets" is widely considered pejorative and outdated,
with "dwarf" or "short stature" being preferred language when
referring to individuals with such medical conditions.’
```