RESEARCH BRIEF • OCTOBER 2025

Artificial Writing and Automated Detection

Based on BFI Working Paper No. 2025-116, "Artificial Writing and Automated Detection," by Brian Jabarian and Alex Imas, University of Chicago

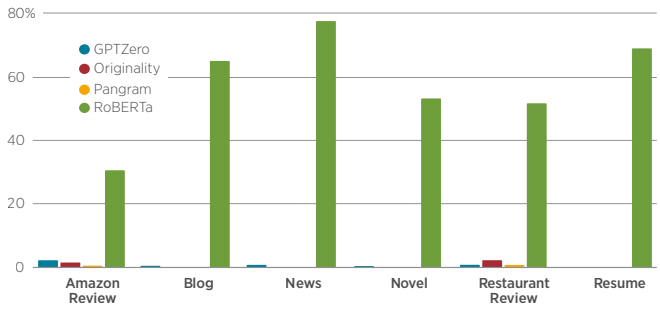
Commercial AI detection tools significantly outperform open-source alternatives, with Pangram achieving near-zero error rates while open source options misclassify up to 78% of human text as AI-generated. A new policy framework allows institutions to systematically compare detectors based on their tolerance for false accusations versus missed AI content.

Generative Artificial Intelligence tools have been adopted faster than any other technology on record, giving rise to writing that is either assisted or entirely completed by Large Language Models (LLMs). The ubiquity of AI-generated writing across domains such as school assignments and consumer reviews presents a new challenge to stakeholders aiming to detect whether content was written by humans. While automated detection tools hold promise, their accuracy claims are difficult to verify since they rely on proprietary data and methods.

In this paper, the authors audit the set of leading AI detection tools and offer a framework to evaluate how they should be incorporated into potential policies. They evaluate four detectors—three commercial tools (Pangram, OriginalityAI, GPTZero) and one open-source baseline (RoBERTa)—on their ability to minimize two critical errors: false positives (wrongly flagging human text as AI) and false negatives (missing actual AI content).

Understanding detector performance requires grasping how these tools work. A detector evaluates text and assigns it a score, such that higher scores imply a greater likelihood that the text is AI-generated. A detector's performance

Figure 1 · Detector False Positive Rates by Genre



depends critically on the score threshold at which it classifies content as AI-generated. A higher threshold implies that the detector requires a higher score to classify a passage as AI-generated; this will naturally decrease the false positive rate while at the same time increasing the false negative rate.

Given this threshold trade-off, the authors evaluate performance in two ways: First using detector-optimized thresholds calculated to maximize the difference between true positive rate and false positive rate, and second by manipulating thresholds exogenously to demonstrate how policy designers can adjust detector settings based on their tolerance for different types of errors.

For their evaluation, the authors use a 1,992-passage text corpus that spans six everyday genres (news, blogs, consumer reviews, novels, restaurant reviews, and résumés) as input for their evaluation. 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). They also examine the effectiveness of AI "humanizers" (StealthGPT) in potentially bypassing detectors.

They find the following:

- Commercial detectors significantly outperform open-source alternatives across all metrics and AI models. Among commercial options, Pangram achieves essentially zero false positive rates and false negative rates on medium-length to long passages, both when using detector-optimized thresholds and exogenously-set thresholds. The false positive rate and false negative rate increase slightly on short passages, but remain well below reasonable policy thresholds.
- The performance gap between commercial and open-source tools is substantial. OriginalityAI and GPTZero constitute a secondary tier among commercial detectors with partial strengths, making the choice between the two dependent on the user's priority: minimizing false positive rate favors GPTZero, while maximizing ability to distinguish AI from human text favors OriginalityAI. In contrast, the open-source RoBERTa base is deemed unsuitable for high-stakes applications, misclassifying most human text with false positive rates of approximately 30-78% across scenarios.
- Pangram's false negative rate is robust to the use of current "humanizers" and remains low even when AI-generated passages are modified using tools such as StealthGPT. The other detectors are less robust to humanizers, with GPTZero largely losing its capacity to detect AI-generated text, showing false negative rate scores around 50% and above across most genres and LLM models.
- 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. Cost-per-true-positive analysis sharpens the price gap, making Pangram the most cost-efficient detector.
- The policy caps framework, which sets exogenous thresholds to test detector robustness, reveals that Pangram is the only detector that meets stringent policy requirements without compromising the ability to detect AI text. When policy caps are set at very low levels (0.5% false positive rate), Pangram continues detecting AI content effectively while other detectors see sharp degradation in their detection capabilities.

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.

READ THE WORKING PAPER

NO. 2025-116 · AUGUST 2025

Artificial Writing and Automated Detection

bfi.uchicago.edu/working-papers/artificial-writing-and-automated-detection

ABOUT OUR SCHOLARS

Brian Jabarian
Howard and Nancy Marks Fellow and Roman Family Center for Decision Research Principal Researcher, Chicago Booth

Alex Imas
Roger L. and Rachel M. Goetz Professor of Behavioral Science, Economics and Applied AI and Vasilou Faculty Scholar, Chicago Booth

THE UNIVERSITY OF CHICAGO
BECKER FRIEDMAN INSTITUTE FOR ECONOMICS

Written by Abby Hiller • Designed by Maia Rabenold