

Who Keeps Up with Whom?

Assessing GenAI Text Detection Across Version Updates

Ahnaf Chowdhury Niloy¹, Golnoush Haddadian¹, and Min Kyu Kim¹

¹College of Education and Human Development, Georgia State University, Atlanta, USA.

Correspondence: aniloy1@gsu.edu

Abstract

Artificial Intelligence use raises academic integrity concerns, yet institutions rely on detection tools of questionable reliability. Rapid version updates to chatbots, detectors, and humanization tools remain unaccounted for in detection research. Across 960 detection scores from two waves, the findings reveal detection outcomes are fundamentally reshaped by these parallel updates.

1 Introduction

Writing-based assessments occupy a central role in higher education as instruments for assessing student learning and cognitive engagement. Sweller's (1988, 2010) Cognitive Load Theory posits that schema construction and long-term retention of learning material require effortful mental processing. When students delegate writing tasks to Generative Artificial Intelligence (GenAI) tools, this cognitive engagement is bypassed. Fan et al. (2024) introduced the term "metacognitive laziness" to describe the cognitive disengagement that emerges when students rely on generative AI to complete writing tasks, and Gerlich (2025) found that AI-driven cognitive offloading is associated with lower critical thinking performance. Empirical evidence indicates that this is already occurring in practice. Students use AI partially or completely to complete their academic work, primarily because it saves time and goes undetected (Basu, 2023; Gill et al., 2024; Niloy et al., 2024a; Niloy et al., 2024b).

In response, AI-text detection tools such as Turnitin and GPTZero have been widely adopted as institutional mechanisms for enforcing academic

integrity. These tools self-report accuracy rates of over 90%, yet empirical evidence consistently challenges these claims (Gao et al., 2023; Mlambo, 2024; Perkins et al., 2024; Walters, 2023). The detection challenge is further compounded by AI-powered humanization tools designed specifically to rewrite AI-generated text to evade detection, reducing detection accuracy by up to 80 percent under high-level humanization conditions (Tarim and Onan, 2025; Wang et al., 2024). Hasham (2026) argues that fear of detection remains the primary resistance against AI misuse, yet Niloy and Huda (2025) note that human evaluation of AI content is not feasible at scale due to systemic issues. If detection tools cannot reliably fulfill this role, academic integrity enforcement lacks a credible mechanism.

In order to ensure students do not have such flexibility; defensive measures must be taken. These measures require tools that help educators protect academic integrity. Given that empirical evidence consistently challenges the accuracy claims of AI detection tools (Gao et al., 2023; Mlambo, 2024; Perkins et al., 2024; Walters, 2023), it is essential to highlight the specific angles that remain unaddressed before revisiting whether AI text detection is accurate or not from a different perspective.

Current detection research has gaps that limit its practical applicability. First, studies conducted so far assess either a single chatbot, a single detector, or both (Bellini et al., 2024; Elkhayat et al., 2023; Weber-Wulff et al., 2023). This approach ignores the possibility of multi-tool use cases, as users are not limited to a single chatbot, nor do instructors rely on a uniform detection tool (Walters, 2023). Second, AI tools produce texts in distinct writing

styles (Fraser et al., 2025; Godwin-Jones, 2024; Berber Sardinha, 2024; Kujur, 2025), with style differences emerging from simple prompt changes. However, current studies are yet to explore whether different writing styles are detected differently. Third, AI tools have diversified, with specialized tools emerging for certain tasks. For instance, one such type is AI humanization tools that modify AI output to make it more human-like to avoid detection. They apply techniques such as lexical substitution, structural paraphrasing, syntactic variation, and noise injection that move text away from AI-typical signatures while preserving meaning (Wang et al., 2025). The effect of these tools on detection is yet to be examined empirically and systematically. Fourth, AI tools are evolving rapidly through frequent updates. Yet, detection studies tend to focus on a single timepoint, treating what is fundamentally a changing environment as if it were fixed. This creates a problem of temporal instability, meaning that detection outcomes at one timepoint may not hold after a version update. Every tool in the ecosystem, whether the chatbot, the humanizer, or the detector, undergoes continuous updates, and each update changes the conditions under which the others operate.

For the purposes of this study, reliability refers to whether AI-text detection scores can be trusted as accurate indicators of AI involvement. With this, we raise a fundamental question that is yet unaddressed in current detection research: Is AI-text detectability reliable when generative AI systems, detection tools, and humanization tools undergo version updates?

2 Method

2.1 Research Design

This study employed a quasi-experimental design to evaluate AI-text detection across two measurement waves. The wave variable represents two distinct states of the same ecosystem: Wave 1 (V1, September 2025) and Wave 2 (V2, May 2026), separated by major version updates to all three components: chatbots, detection tools, and the humanization tool. These updates occurred on commercial schedules outside the researcher's control, functioning as a naturally occurring manipulation between waves. Because the topic, chatbots, writing styles, detectors, and procedure were held constant across both waves, differences in detection outcomes between the tools' initial

version (V1) and the latter version (V2) can be attributed to the version updates rather than to changes in procedure. **Table 1** (see Appendix) presents the specific versions of all tools used across both waves.

2.2 Tools and Experimental Conditions

Six generative AI chatbots (ChatGPT, Copilot, Gemini, Claude, Grok, and Meta AI) produced the sample essays. The two most widely deployed AI-text detection tools in higher education (Turnitin and GPTZero) scored them. One AI-powered humanization tool, HIX AI, processed each essay at maximum rewriting intensity, using its *Aggressive* setting in V1 and its *Latest* setting in V2. All essays addressed a single topic, a high school organic chemistry question on producing ester and carboxylic acid from a two-carbon alkene, to control for content variation. Four writing styles were used: *default*, *academic*, *creative*, and *informative*.

Academic writing is characterized by high lexical density, nominalization, and epistemic hedging (Biber and Gray, 2010; Hyland, 2005, 2008). Creative writing is marked by figurative language and syntactic variability (Biber, 1988). It also draws on a wider range of vocabulary than other styles, a property commonly measured using the Type-Token Ratio (Biber, 1988; Kao and Jurafsky, 2012). Informative writing favors common vocabulary, short transparent sentences, and explicit cohesive devices (Fisher and Frey, 2013). Default style captures unmodified AI output as a baseline. These represent the dominant range of composition tasks in higher education and are linguistically distinct in ways that early evidence suggests affect detection outcomes (Aull, 2019; Fraser et al., 2025; Goulart, 2024).

2.3 Data Collection Procedure

In V1, each chatbot generated five essays in each of four writing styles across five independent conversation sessions, resulting in 120 original essays. Each of these essays was then processed through HIX AI to produce 120 humanized counterparts, yielding 240 essays for V1. The same procedure was repeated in V2, producing a further 240 essays and a total of 480 across both waves. Each essay was then evaluated using both Turnitin and GPTZero, yielding 960 detection scores. Both detectors report scores on a scale ranging from 0 to 100, with higher values indicating greater

likelihood of AI generation. **Table 2** (see Appendix) presents the prompting sequence used in each independent session with the AI chatbot.

2.4 Data Analysis

Data preparation and descriptive statistics were conducted in Python 3.14, and inferential analyses in R 4.5.2. The principal method was a Linear Mixed Effects Model with detection score as the dependent variable, chatbot, writing style, detector, humanization, wave, and their key interactions as fixed effects, and essay as the random intercept. The Wilcoxon signed-rank test and Mann-Whitney U test were used for paired and independent comparisons respectively, after Shapiro-Wilk tests confirmed non-normality.

3 Analysis and Findings

3.1 Descriptive Overview

The final dataset comprised 960 detection scores from 480 essays, consisting of 120 original and 120 humanized essays per wave, each evaluated by both Turnitin and GPTZero. Overall detection scores ranged from 0 to 100 ($M = 49.48$, $SD = 36.77$). GPTZero produced higher mean detection scores than Turnitin (53.30 vs 45.67), though with substantially greater variability. Detection scores were higher in V2 ($M = 57.75$) than in V1 ($M = 41.22$), though this overall increase is misleading. Some chatbots became more detectable between waves while others became less detectable, and the two detectors evolved in opposite directions. Non-humanized essays scored considerably higher than humanized essays (61.94 vs 37.03), confirming that humanization suppressed detection across the sample.

3.2 Detection Patterns Across Waves

In V1, a clear chatbot detectability hierarchy existed with nine of the fifteen pairwise comparisons reaching significance after Bonferroni correction. Differences ranged from 9.14 points (Gemini vs. Grok, $r = .23$) to 46.72 points (Meta AI vs. ChatGPT, $r = .56$). GPTZero was substantially more sensitive than Turnitin for several chatbots, with particularly large differences for Meta AI and Claude. Humanization was highly effective in V1. GPTZero scores collapsed from 87.97 to 10.02 following humanization ($W = 1.00$, $p < .001$, $r = .86$), and Turnitin showed a significant but more moderate reduction from 40.04 to 26.84

($W = 442.50$, $p < .001$, $r = .39$). **Table 4** (see Appendix) presents all significant V1 pairwise comparisons.

By V2, this landscape had changed fundamentally. Only one of the fifteen pairwise comparisons remained significant, Claude vs. Gemini ($\Delta EMM = 10.97$, $r = .24$), with all other chatbot pairs statistically indistinguishable. ChatGPT rose from $EMM = 24.66$ to 58.61, while Meta AI declined from 71.39 to 58.00, and the detector asymmetry that defined V1 had also converged substantially. Most strikingly, humanization effects were non-significant for both Turnitin ($W = 3137.50$, $p = .393$, $r = .08$) and GPTZero ($W = 2991.50$, $p = .094$, $r = .15$) in V2. Humanization, nearly decisive in V1, had no significant effect in V2 for either detector. **Table 5** (see Appendix) presents the V2 pairwise comparisons.

3.3 Style Effects on Detectability

Writing style produced no significant main effect on detection scores ($\chi^2(3) = 0.19$, $p = .979$). This finding of statistical insignificance is itself meaningful. This challenges prior literature suggesting that academic writing is systematically more detectable than creative or informative styles (Aull, 2019; Fraser et al., 2025; Goulart, 2024). A plausible explanation is that as AI chatbots become more stylistically capable, the genre-specific signatures that detectors previously relied upon may have lost their discriminating power. The informative style, however, revealed a directional reversal that illustrates the broader instability across version updates. Informative writing shifted from the least detectable style in V1 ($EMM = 36.47$) to the most detectable in V2 ($EMM = 60.42$). This reversal does not indicate that detectors improved specifically for informative writing. Rather, it shows that the relative ranking of styles is itself unstable across version updates. When the overall detection landscape shifts between versions, the position of any given style within that landscape can change entirely, independent of anything specific to the style itself.

3.4 Evidence of Temporal Instability

To test the significance of fixed effects and their interactions within the LMM, Type III Wald chi-squared tests were applied. **Table 3** (see Appendix) presents the full omnibus test results for all fixed effects and interactions. These tests confirmed that

instability in detection outcomes was statistically significant at every level of the detection landscape.

The main effect of Wave was significant ($\chi^2(1) = 27.66, p < .001$), but the more informative evidence lies in the interaction terms. The Chatbot $\times$ Wave interaction ($\chi^2(5) = 98.33, p < .001$) confirmed that chatbot-specific detectability patterns changed significantly between version updates, with Meta AI showing the largest shift ($\beta = -89.83, SE = 11.49$). The Detector $\times$ Wave interaction ($\chi^2(1) = 60.14, p < .001$) confirmed that the two detectors did not evolve equivalently, with GPTZero's advantage over Turnitin substantially reduced under V2 detector versions ($\beta = -65.14, SE = 8.40$). The Chatbot $\times$ Detector $\times$ Wave interaction ($\chi^2(5) = 122.65, p < .001$) confirmed that the relationship between chatbot identity and detector response was structurally different under V2 tool versions than under V1 tool versions. Most critically, the Detector $\times$ Humanization $\times$ Wave interaction ($\chi^2(1) = 135.73, p < .001, \beta = 61.89, SE = 5.31$) captured the simultaneous collapse of humanization effectiveness across both detectors following the version updates between waves.

While the LMM interactions confirmed significant change across version updates, they could not identify which of the three components evolving simultaneously drove the observed shifts. Two supplementary diagnostics addressed this attribution question. The Spearman rank correlation between V1 and V2 chatbot detection means assessed rank stability, and the per-wave variance of chatbot detection means assessed output homogenization. The Spearman rank correlation was $\rho = -0.26$ ($p = .623$), indicating near-zero rank stability. The variance of chatbot detection means collapsed by 95%, from 296.94 in V1 to 14.51 in V2. By contrast, the two detectors evolved in opposite directions, with Turnitin's sensitivity to non-humanized text increasing and GPTZero's declining. This asymmetric detector evolution, combined with the near-inverted chatbot ranking, is more consistent with substantial chatbot evolution than with detector-driven change, though this interpretation remains inferential given that all three components updated simultaneously between waves.

4 Discussion

This study's central finding is that AI-text detection tools performed substantially below their vendor

accuracy claims, and that this limited performance was further destabilized by version updates to all three ecosystem components. GPTZero advertises a 99% accuracy rate for detecting advanced language models (GPTZero, 2025), yet detection scores on fully AI-generated, non-humanized essays produced an overall mean of only 61.94 across all chatbots and detectors. This represents a gap of more than 37 points on entirely AI-generated, unmodified content, suggesting the shortfall is not marginal but structural. If detection tools cannot reliably flag content that is fully AI-generated, their ability to accurately classify human-AI collaborative writing is even more questionable. A student who wrote an essay independently but used AI to refine their language may face a real risk of being flagged incorrectly, not because they cheated but because the detection tool cannot distinguish between full AI generation and legitimate AI assistance.

Beyond this low baseline, the detection patterns observed in V1 did not persist into V2. The chatbot detectability hierarchy inverted, detector sensitivity converged, and humanization no longer functioned as an evasion strategy. These shifts occurred through routine commercial version updates, not exceptional technological events. This raises a concern more pressing than the limitations of single-timepoint research designs. Students are being evaluated, educators are forming judgments, and institutions are making policy decisions based on detection scores whose meaning shifts between version updates without these stakeholders' awareness. A score produced under one set of tool versions may mean something entirely different under the next.

This finding extends Walters' (2023) methodological critique that single-timepoint assessments produce findings that may not generalize as tools evolve. The present study confirms this concern empirically: instability is a property of the competitive balance among multiple simultaneously evolving components, not simply a matter of one tool improving or declining. When any of these components update, it changes the conditions under which the others operate. The humanization collapse in V2 illustrates this clearly. Prior studies documented humanization as highly effective at suppressing detection (Wang et al., 2024; Tarım and Onan, 2025), and the present findings do not contradict this for V1. However, the reason humanization became non-significant in V2

is not that updated detectors became better at detecting humanized text. In V2, updated chatbot versions produced text that was already substantially less detectable before any humanization was applied. This left humanization with fewer detection elements to suppress. Humanization's effectiveness was therefore not a stable property of the tool itself but of its interaction with chatbot outputs and detector sensitivity at the specific timepoint. The chatbot convergence pattern adds a further layer to this interpretation. In V1, the six chatbots produced meaningfully different detection profiles spanning a range of approximately 47 points. By V2, this range had compressed to approximately 11 points and the V1 ranking had largely inverted. This pattern is more consistent with chatbot evolution than with detector-driven changes, though the interpretation remains inferential. This suggests V2 chatbot outputs became more uniformly human-like, reducing detectability across platforms.

Dugan et al. (2024) characterized AI text detection as an arms race in which detector performance degrades as language models evolve. The present findings support this characterization, showing that such degradation can occur across multiple platforms simultaneously within a single update cycle. The absence of a significant style effect further reinforces the systemic nature of detection failure documented in this study. Detection scores did not differ meaningfully across academic, creative, informative, and default writing styles, suggesting the unreliability observed is not confined to any particular writing style. Prior literature suggested academic writing produces higher detection rates than creative or informative styles (Aull, 2019; Fraser et al., 2025; Goulart, 2024), but the present findings do not replicate this pattern. A plausible explanation is that modern AI systems have become stylistically capable enough that the genre-specific signatures detectors previously relied upon are no longer effective (Zaitsu et al., 2025). Even the directional reversal of the informative style, from least detectable in V1 to most detectable in V2, is best understood not as a style effect but as further evidence that detection patterns shift unpredictably across version updates regardless of genre. These findings collectively show that detection outcomes can change fundamentally between version updates, not because students may behave differently, but because the tools used to evaluate

them have changed. This has real consequences for students, educators, and institutions who rely on detection scores to make judgments about academic integrity. A student who used AI minimally for language refinement in V1 and received a low detection score may receive a substantially different score for the same behavior in V2, not because their conduct changed but because the ecosystem did. An educator who calibrated their judgment to V1 detector behavior may be applying outdated standards to V2 scores without knowing it. These concerns bear directly on the fairness of the process and the validity of interpreting detection scores as reliable evidence of academic misconduct, particularly when tool versions change without the knowledge of those acting on them.

5 Implications

5.1 Detection Research

The most direct methodological implication is that single-chatbot, single-timepoint detection studies cannot characterize detection reliability as a property of the current technological ecosystem. The between-chatbot range of 47 points in V1 means that a study using only ChatGPT would have concluded that detection tools are largely ineffective, while a study using only Meta AI would have reached the opposite conclusion. Neither would have been wrong given its design, but neither would have produced generalizable conclusions. The field requires longitudinal, multi-chatbot, multi-detector evaluation frameworks that treat these components as evolving entities, and the present study offers one such template.

5.2 Detection Technology

Current detection tools attempt to classify submitted text as AI-generated or human-written, yet both the texts being evaluated and the tools doing the evaluation change continuously between version updates. Detection scores for fully AI-generated content shifted substantially between V1 and V2, meaning that the same numerical score carries a different evidentiary weight depending on which tool version produced it. A score of 60 under V1 tool versions reflects a different detection context than a score of 60 under V2 tool versions. This calls for a fundamentally different approach to detection.

A more reliable direction would be to verify the origin of text at the point of generation rather than classifying it after submission. AI developers could embed invisible attribution markers that distinguish between fully AI-generated content and AI-assisted human writing. Google's *SynthID* technology is experimentally doing this for Gemini outputs, showing that the idea may be technically achievable. Under this approach, when a student submits work, the detection platform would ask a simple question: did you use AI? If they say no but their submitted text carries an AI-generated marker, the system flags the mismatch. If they say yes, the student submits both their original draft and the AI-refined version. The educator sees exactly what the student wrote and what AI changed. This approach also helps students who used AI only for language refinement, including non-native English speakers, to show their genuine contribution openly rather than being unfairly flagged by an opaque system (Liang et al., 2023; Perkins et al., 2024). While broader adoption would require industry-wide watermarking standards, this approach points toward a detection paradigm built on transparency rather than suspicion.

5.3 Institutional Practice

The principle that assessment drives instruction is directly relevant here (Black and Wiliam, 1998). When assessment instruments are unreliable, the learning behaviors they are designed to encourage become distorted. If students believe detection is unpredictable, they either lose confidence in the fairness of the system or exploit the inconsistency for academic dishonesty. If educators cannot trust detection scores, they either over-rely on them or ignore them.

No detection score should trigger disciplinary action without human review of the broader context, including the assignment type, the student's prior work, and the version of the detector at the time of evaluation. This concern is especially acute in educational contexts where essay-based assessments carry substantial institutional weight (Chong and McArthur, 2023; Sivasubramaniam, 2024). In these environments, an inaccurate detection judgment does not simply affect a grade. It can jeopardize a student's academic record, their standing within a competitive system, and their long-term educational trajectory. The instability documented in this study means that students in these contexts face not only the well-documented

risks of false positives but also the risk that detection conditions have shifted since the policy under which they are evaluated was adopted.

Institutions and educators need training materials that distinguish between AI use that replaces learning and AI use that supports it. Framing all AI use as dishonest creates incentives for concealment rather than transparency. A more productive approach defines acceptable AI collaboration in terms of the student's cognitive contribution rather than tool use alone.

6 Conclusion

Detection studies have consistently treated AI-text detection as a static phenomenon, evaluating tools at a single timepoint using fixed versions of chatbots and detectors. The present study addressed this gap by examining whether detection reliability holds when chatbots, detectors, and humanization tools undergo simultaneous version updates. The findings show that detection reliability shifted fundamentally within a single ordinary update cycle, driven not by exceptional disruption but by the routine pace of version updates. This is precisely what makes the findings consequential for students, educators, and institutions whose decisions rest on the assumption that detection scores are stable. As the first study to examine detection reliability across simultaneous version updates of all three technological components, this work establishes a replicable template that future research can build upon. The single-topic design controlled for content variation; future studies should apply this framework across multiple disciplines and assignment types to test whether the findings generalize beyond a controlled setting. Expanding the design to include additional detection tools and humanization platforms would further strengthen generalizability. Also, to determine the rate of false positives, scripts produced by humans (with and without the assistance from AI) can also be another measurement aspect for future works. Longitudinal designs with more than two waves would allow researchers to track whether the instability documented here accelerates, stabilizes, or reverses as the technology matures. Finally, qualitative and corpus-linguistic analysis of the generated essays could identify the specific linguistic features that shift between versions, providing detector developers with evidence-based targets for classifier improvement.

Acknowledgments

The authors declare that this work received no external funding and that there are no acknowledgements to report.

References

- Ahmed M. Elkhatat, Khaled Elsaid, and Saeed Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1):17.
- Ahnaf C. Niloy, Md. A. Bari, Jakia Sultana, Rup Chowdhury, Fareha M. Raisa, Afsana Islam, Saadman Mahmud, Iffat Jahan, Moumita Sarkar, Salma Akter, Nurunnahar Nishat, Muslima Afroz, Amit Sen, Tasnem Islam, Mehedi H. Tareq, and Md. A. Hossen. 2024a. Why do students use ChatGPT? Answering through a triangulation approach. *Computers and Education: Artificial Intelligence*, 6(6):100208. <https://doi.org/10.1016/j.caeai.2024.100208>
- Ahnaf C. Niloy, Reaz Hafiz, Bhuiyan M. Hossain, Farhana Gulmeher, Naznin Sultana, Khondokar F. Islam, Fabiha Bushra, Samiya Islam, Syed I. Hoque, Md. Rahman, and Sayma Kabir. 2024b. AI chatbots: A disguised enemy for academic integrity? *International Journal of Educational Research Open*, 7:100396. <https://doi.org/10.1016/j.ijedro.2024.100396>
- Ahnaf C. Niloy and Tazreen Huda. 2025. Is AI only to Blame? Assessing Teachers' Perceived Challenges in AI Detectability. *Asian Journal of Assessment in Teaching and Learning*, 15(2):21–42.
- Arahan Kujur. 2025. A comparative analysis of AI-generated and human-written text: Linguistic patterns, detection accuracy, and implications for modern communication. SSRN. <https://ssrn.com/abstract=5833302>
- Bishal Basu. 2023. ChatGPT and its impact on education sector. *Daily Sun*. <https://www.daily-sun.com/20/673514>
- Cathrine A. Gao, Fredrich M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6:75. <https://doi.org/10.1038/s41746-023-00819-6>
- Debora Weber-Wulff, 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):26. <https://doi.org/10.1007/s40979-023-00146-z>
- Doris Y. K. Chong and Jan McArthur. 2023. Assessment for learning in a Confucian-influenced culture: Beyond the summative/formative binary. *Teaching in Higher Education*, 28(6):1395–1411.
- Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK.
- Douglas Biber and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1):2–20. <https://doi.org/10.1016/j.jeap.2010.01.001>
- Douglas Fisher and Nancy Frey. 2013. *Better Learning Through Structured Teaching: A Framework for the Gradual Release of Responsibility*. ASCD, Alexandria.
- İsmail Tarım and Aytuğ Onan. 2025. Can You Detect the Difference? *arXiv*. <https://doi.org/10.48550/arXiv.2507.10475>
- John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.
- John Sweller. 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2):123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In David Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz (editors), *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17. Association for Computational Linguistics. <https://aclanthology.org/W12-2502>
- Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2025. Detecting AI-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, 82:2233–2278. <https://doi.org/10.48550/arXiv.2406.15583>
- Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Bloomsbury Academic, United Kingdom.
- Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21. <https://doi.org/10.1016/j.esp.2007.06.001>
- Larissa Goulart. 2024. *Variation in University Student Writing: A Communicative Text Type Approach*.

- John Benjamins Publishing Company.
<https://doi.org/10.1075/scl.117>
- Laura Aull. 2019. Linguistic Markers of Stance and Genre in Upper-Level Student Writing. *Written Communication*, 36(2):267–295.
<https://doi.org/10.1177/0741088318819472>
- Liam Dugan, Alyssa Hwang, Filip Trhlik, Andrew Zhu, Josh M. Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Michael Gerlich. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6.
<https://doi.org/10.3390/soc15010006>
- Mike Perkins, Jasper Roe, Binh H. Vu, Darius Postma, Don Hickerson, James McGaughran, and Huy Q. Khuat. 2024. Simple techniques to bypass GenAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21.
<https://doi.org/10.1186/s41239-024-00487-w>
- Paul Black and Dylan Wiliam. 1998. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.
<https://doi.org/10.1080/0969595980050102>
- Philani B. Mlambo. 2024. Investigating ChatGPT: A Threat or Benefit to Higher Education? *Futurity Education*, 4(1):274–289.
<https://doi.org/10.57125/FED.2024.03.25.14>
- Robert Godwin-Jones. 2024. Generative AI, Pragmatics, and Authenticity in Second Language Learning. *arXiv*, abs/2410.14395.
- Shiva D. Sivasubramaniam. 2024. Academic Integrity in South Asia: Focus on India, Pakistan, and Sri Lanka. In S. E. Eaton (editor), *Second Handbook of Academic Integrity, Springer International Handbooks of Education*. Springer, Cham.
https://doi.org/10.1007/978-3-031-54144-5_88
- Sukhpal S. Gill, Minxian Xu, Panos Patros, Huaming Wu, Rupinder Kaur, Kamalpreet Kaur, Stephanie Fuller, Manmeet Singh, Priyansh Arora, Ajith K. Parlikad, Vlado Stankovski, Ajith Abraham, Soumya K. Ghosh, Hanan Lutfiyya, Salil S. Kanhere, Rami Bahsoon, Omer Rana, Schahram Dustdar, Rizos Sakellariou, Steve Uhlig, and Rajkumar Buyya. 2024. Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems*, 4:19–23.
<https://doi.org/10.1016/j.iotcps.2023.06.002>
- Tony Berber Sardinha. 2024. AI-Generated vs Human-Authored Texts: A Multidimensional Comparison. *Applied Corpus Linguistics*, 4:100083.
<https://doi.org/10.1016/j.acorp.2023.100083>
- Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2025. Humanizing the machine: Proxy attacks to mislead LLM detectors. *arXiv*, abs/2410.19230.
<https://doi.org/10.48550/arXiv.2410.19230>
- Valentina Bellini, Fredrico Semeraro, Jonathan Montomoli, Marco Cascella, and Elena Bignami. 2024. Between human and AI: assessing the reliability of AI text detection tools. *Current Medical Research and Opinion*, 40(3):353–358.
<https://doi.org/10.1080/03007995.2024.2310086>
- Wataru Zaitzu, Mingzhe Jin, Shunichi Ishihara, Satoru Tsuge, and Mitsuyuki Inaba. 2025. Stylometry can reveal artificial intelligence authorship, but humans struggle: A comparison of human and seven large language models in Japanese. *PLOS One*, 20(10):e0335369.
<https://doi.org/10.1371/journal.pone.0335369>
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779.
<https://doi.org/10.1016/j.patter.2023.100779>
- William H. Walters. 2023. The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, 7(1):20220158.
<https://doi.org/10.1515/opis-2022-0158>
- Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2024. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*. Advance online publication. <https://doi.org/10.1111/bjet.13544>
- Zemina Hasham. 2026. Are universities ready for AI-native academic integrity? Turnitin.
<https://www.turnitin.com/blog/are-universities-ready-for-ai-native-academic-integrity>

A Appendices: Analysis Tables.

Table 1.

AI Tools, Detectors, and Humanizers Used for the Experiment

Tools	Category	Version 1 (V1)	Version 2 (V2)
ChatGPT	Chatbot	GPT-4o	GPT-5
Claude	Chatbot	Opus 4.1	Sonnet 4.6
Gemini	Chatbot	Flash 2.5	Pro 3.1
Grok	Chatbot	V4	V4.3
Copilot	Chatbot	GPT 4.1 Mini	GPT 5.2
Meta AI	Chatbot	Llama 4	Muse Spark
HIX AI Bypass*	AI Humanizer	Aggressive	HIX 2.1 (Latest)
Turnitin	AI Detector	July 2024 Patch	February 2026 Patch
GPTZero	AI Detector	Model 3.9b	Model 4.5b

Note: HIX AI Bypass typically do not publicly publish their version updates on AI Humanizer models or specific dates of update. However, they have highlighted a toolkit upgradation taking place in 2026. Since V1 data collection was conducted in September 2025 and V2 data collection in April-May 2026, this study safely assumes two different versions being examined for HIX AI Bypass.

Table 2.

Prompt Sequence with Action Explanation

Sl.	Prompt Sequence	Action
1	"Imagine you are a student and need to write an answer to a question on organic chemistry. You need to write the answer in a paragraph structure and add chemical equations as an example to your response. Now write the answer to the following question: How can you get Ester and Carboxylic Acid from a 2 Carbon Alkene?"	The chatbot generates a response essay in its default style. The author copies the response and saves it as a Word file, labeling it X1.
2	The author pastes the X1 essay back into the same chat session and prompts: " <i>Rewrite the following essay in Academic writing style: [pasted X1 essay]</i> "	The chatbot rewrites the pasted essay in academic style. The author saves the output and labels it XA1.
3	The author pastes the X1 essay back into the same chat session and prompts: " <i>Rewrite the following essay in Creative writing style: [pasted X1 essay]</i> "	The chatbot rewrites the X1 essay in creative style. The author saves the output and labels it XC1.
4	The author pastes the X1 essay back into the same chat session and prompts: " <i>Rewrite the following essay in Informative writing style: [pasted X1 essay]</i> "	The chatbot rewrites the X1 essay in informative style. The author saves the output and labels it XI1.

Note: The complete prompt sequence (Sl. 1 to 4) was repeated in a new conversation session for each of the five default essay generations per chatbot, producing X1 through X5 with their corresponding XA, XC, and XI variants. The numbering convention applies to all six chatbots used in the study.

Table 3.*Type III Omnibus Tests of Fixed Effects from the Linear Mixed Effects Model*

Effect	χ^2	<i>df</i>	<i>p</i>
Main Effects			
Chatbot	191.39	5	< .001
Detector	93.84	1	< .001
Wave	27.66	1	< .001
Humanization	3.44	1	.064
Style	0.19	3	.979
Two-Way Interactions			
Chatbot × Detector	116.62	5	< .001
Detector × Humanization	104.51	1	< .001
Chatbot × Wave	98.33	5	< .001
Detector × Wave	60.14	1	< .001
Chatbot × Humanization	49.20	5	< .001
Chatbot × Style	43.65	15	< .001
Wave × Humanization	5.59	1	.018
Style × Humanization	3.65	3	.302
Wave × Style	1.23	3	.746
Style × Detector	0.41	3	.938
Three-Way Interactions			
Detector × Humanization × Wave	135.73	1	< .001
Chatbot × Detector × Wave	122.65	5	< .001
Chatbot × Style × Wave	38.18	15	< .001
Chatbot × Detector × Humanization	32.74	5	< .001
Chatbot × Style × Detector	25.01	15	.050
Chatbot × Humanization × Wave	13.39	5	.020
Style × Detector × Wave	4.39	3	.222
Style × Humanization × Wave	1.48	3	.688

Note: χ^2 values are Wald chi-square statistics derived from Type III tests.

Table 4.*Significant Bonferroni-Corrected Chatbot Pairwise Comparisons — Wave 1*

Comparison	<i>EMM</i>₁	<i>EMM</i>₂	<i>ΔEMM</i>	<i>p</i>adj	<i>Effect Size</i>
MetaAI vs. ChatGPT	71.39	24.66	+46.72	< .001	.56
MetaAI vs. Grok	71.39	26.96	+44.42	< .001	.48
MetaAI vs. Claude	71.39	38.88	+32.51	< .001	.46
MetaAI vs. Gemini	71.39	36.10	+35.29	< .001	.39
Copilot vs. ChatGPT	49.31	24.66	+24.65	< .001	.40
Copilot vs. Grok	49.31	26.96	+22.35	.001	.31
MetaAI vs. Copilot	71.39	49.31	+22.08	.023	.24
Gemini vs. ChatGPT	36.10	24.66	+11.44	.001	.31
Gemini vs. Grok	36.10	26.96	+9.14	.045	.23

Note: EMM = estimated marginal mean derived from the LMM reference grid, averaging over all levels of Style, Detector, and Humanization. Only pairwise comparisons reaching statistical significance (*p*adj < .05) after Bonferroni correction for 15 comparisons are reported. *ΔEMM* = difference between estimated marginal means (first minus second chatbot). Positive values indicate the first-named chatbot produced higher detection scores. *r* = effect size. All remaining comparisons were non-significant.

Table 5.*Significant Bonferroni-Corrected Chatbot Pairwise Comparisons — Wave 2*

Comparison	<i>EMM</i>₁	<i>EMM</i>₂	<i>ΔEMM</i>	<i>p</i>adj	<i>Effect Size</i>
Claude vs. Gemini	63.74	52.76	+10.97	.034	.24

Note: EMM = estimated marginal mean derived from the LMM reference grid. Only one pairwise comparison reached statistical significance (*p*adj < .05) after Bonferroni correction for 15 comparisons, reflecting the substantial convergence of chatbot detection scores in Wave 2 relative to Wave 1. All remaining comparisons were non-significant and are available from the corresponding author upon reasonable request.