Issue 11 – December 2025

Logos et Littera – Journal of Interdisciplinary Approaches to Text
ISSN: 2336-9884

Exploring AI-detection variability in university student essays: an exploratory comparative study of four tools

Katarina Držajić Laketić,[1] Pan-European University "Apeiron", Banja Luka, Bosnia and Herzegovina

Abstract: This exploratory pilot study applied four AI-detection tools (Grammarly, QuillBot, BypassAI, Phrasely) to four anonymized undergraduate essays in English – two formal academic and two conversational/narrative – to test feasibility and observe classification patterns. Inter-tool agreement was high (15 of 16 classifications; 93.75%). Both formal essays were unanimously labeled AI-generated (4/4 tools), whereas the conversational essays were classified as human in 7 of 8 tool decisions; one detector diverged on a single text. The pattern suggests that polished academic features are more likely to trigger AI flags, while narrative/conversational style is less susceptible. Usability and transparency varied across tools, limiting interpretability. Given the small N and unknown ground truth of authorship, findings are illustrative rather than generalizable; however, they align with concerns about bias toward non-native or formulaic academic prose. The study supports cautious, multi-tool use and argues for assessment designs and policies that contextualize detector outputs within broader academic integrity frameworks.

Keywords: exploratory study, AI-detection tools, student essays, academic integrity, detection variability

1. Introduction

Academic integrity is fundamental to higher education. Innovative tools like ChatGPT have shaken up the academic world, leaving many educators struggling to tell whether a piece of writing came from a student or an algorithm. Consequently, many institutions now rely on AI-detection tools such as Grammarly, QuillBot, BypassAI, and Phrasely to uphold academic standards. However, as practitioners in the field, we still understand surprisingly little about how consistently these tools perform, especially when applied to work by non-native English writers, whose linguistic style may unintentionally trigger flags.

Emerging studies suggest that AI-detectors exhibit notable unreliability. Some tools fail to detect AI-generated paraphrased text, and others frequently produce false positives, misclassifying genuine writing, particularly from non-native speakers, as AI-generated (Weber-Wulff et al. 2023; Liang et al. 2023). These biases raise ethical concerns, as they may unfairly disadvantage learner populations seemingly less fluent in standard academic English. There are also practical limitations: paraphrasing or subtle editing can drastically reduce detection rates, which points to the fragility of these systems (Inside Higher Ed 2024).

From a theoretical standpoint, using these opaque "black box" tools raises questions of measurement validity and reliability. As the training data and decision mechanisms are not transparent, tool outputs should be interpreted with caution, particularly when they influence disciplinary decisions. At the same time, many educational institutions are exploring AI-resilient pedagogies. Rather than focusing solely

[1] ORCID: https://orcid.org/0009-0003-9167-1714, email: katja.drzajic@gmail.com.

doi: 10.31902/LL.11.2025.2

on detection, some universities are designing assessments that encourage critical thinking, discourage misuse of AI, and foreground student voice (Packback 2025). Such approaches honor academic integrity, while at the same time acknowledging that AI is now part of the learning ecosystem.

This paper contributes to that debate by presenting a small-scale exploratory study that examines how four popular AI-detection tools classify four anonymized undergraduate essays in English. Unlike large-scale benchmarking studies that often use synthetic corpora, this investigation is grounded in student work, thereby offering insight into how detection systems behave in real academic settings. The study also highlights the risks of bias and misclassification in relation to linguistic diversity, an area where empirical research remains limited.

We pose the following research question: How do four AI-detection tools (Grammarly, QuillBot, BypassAI, Phrasely) classify student essays, and what patterns in inter-tool agreement or sensitivity to style emerge?

Importantly, this study should be read as illustrative rather than generalizable. The small set of anonymized essays provides preliminary insight into tool behavior, offering a starting point for further investigation. As Lakens (2020) notes, pilot studies serve primarily to test procedures and generate hypotheses rather than to yield definitive conclusions.

2. Literature Review

Recent research highlights growing concerns about the accuracy and fairness of AI detection tools, particularly when applied to writing by non-native English speakers. A widely cited Stanford study found that GPT detectors incorrectly flagged 61 percent of TOEFL essays written by non-native English users as AI-generated, while essays by native speakers were almost never misclassified (Liang et al. 2023). This pattern reflects a systematic bias: non-native writing typically exhibits reduced lexical variety and simpler syntax, which detection algorithms may confuse with AI-generated text.

Weber-Wulff et al. (2023) evaluated fourteen detection tools, including widely used systems like GPTZero and Turnitin, and reported that none exceeded 80 percent accuracy. Rather than reliably identifying AI-generated content, many tools consistently misclassified human writing (false positives) or failed to detect AI content (false negatives). Even minor paraphrasing or translation reduced detection effectiveness significantly, casting doubt on the reliability of these systems in real-world academic settings.

The fragility of detection tools becomes even more apparent under adversarial conditions. Sadasivan et al. (2023) found that lightly edited AI-generated texts often evaded detection altogether. Similarly, Anderson et al. (2023) observed that minimal paraphrasing could flip systems' classification from AI-generated to near-certain human-written. These findings suggest that detection tools are easily bypassed and heavily dependent on rigid assumptions about text features, limiting their efficacy in practical contexts.

From a theoretical standpoint, these patterns raise urgent questions about measurement validity and algorithmic fairness. Most AI detection tools operate as opaque "black boxes," making decisions based on undisclosed criteria. The absence of transparency prevents educators or students from understanding or contesting why certain pieces of writing are flagged. Academic integrity frameworks emphasize fairness, accountability, and replicable decision-making, which are standards many AI detection systems fail to meet.

Institutional concerns are, however, mounting. In Australia, for instance, regulatory bodies have warned that overreliance on AI detection could lead to thousands of false accusations per semester, which

10.31902/LL.11.2025.2

would disproportionately affect international students whose language patterns naturally resemble AI text (Herald Sun 2024; Provost UKY 2024). Educators and policy experts now recommend using detection tools cautiously and supporting them with human oversight and pedagogical safeguards.

Educational scholars increasingly promote AI-resilient assessment design as a more equitable and reliable alternative. Rather than depending exclusively on detection, instructors are advised to incorporate process-oriented assignments such as drafting workflows, revisions, oral presentations, or reflective journals, that foreground students' critical thinking and writing development. This approach reduces the stakes associated with detection errors and aligns with integrity frameworks that value authentic academic engagement over automated enforcement.

Systematic reviews of recent literature reinforce these insights. Rafiq et al. (2025) and Evangelista (2025) jointly argue for a comprehensive approach combining detection tools with AI literacy education, clear institutional policies, and integrity-centered teaching strategies. With few exceptions, most empirical studies focus on synthetic datasets or controlled test inputs; few analyze real student essays with varied style, tone, and linguistic complexity.

This gap is particularly notable in small-scale, context-specific investigations. While large-n quantitative studies provide broad statistical trends, small-n exploratory work allows for nuanced, qualitative examination of tool behavior on authentic texts. In educational research, such studies have proven valuable for identifying subtle biases or edge cases that might be obscured in larger datasets (Billingham, Whitehead, and Julious 2013; Faber and Fonseca 2014). For AI detection, where classifier outputs may hinge on stylistic features or context-specific linguistic norms, close analysis of a handful of representative texts can reveal patterns that inform both policy and pedagogy.

The present study directly addresses this niche by applying four widely used detectors to four authentic, anonymized undergraduate essays differing in stylistic register. By analyzing classification patterns across formal academic and conversational narratives, the research offers a granular view of how style influences detection outcomes. This contribution not only supplements large-scale evaluations but also provides actionable insights for educators and administrators tasked with interpreting detector outputs in real-world assessment scenarios.

3. Methodology

This study was designed as a small-scale exploratory pilot, aimed at examining how widely used AI-detection tools behave when applied to anonymized undergraduate essays. The limited sample size was intentional: rather than supporting statistical generalization, the focus was on testing procedures, identifying methodological challenges, and generating directions for more robust future research (Eldridge et al. 2016; Lakens 2020). Since the essays may have been produced with varying degrees of student or AI assistance, the study does not claim to verify authorship, but rather to investigate how detectors classify different stylistic registers under realistic conditions.

The study adopts a measurement-and-validation lens: AI-detection tools are treated as fallible classifiers that operationalize proxy indicators of text origin, so their outputs constitute evidence about an underlying construct rather than ground truth. In line with argument-based and consequential validity perspectives, the design prioritizes ecological realism and examines how classifications vary across stylistic registers, as well as what inferences would be warranted in instructional settings. The small-N, case-based logic aims at analytic (not statistical) generalization and hypothesis generation, which is consistent with the goals of pilot/feasibility research (Eldridge et al. 2016; Lakens 2020).

10.31902/LL.11.2025.2

The decision to work with four essays was deliberate. Rather than drawing on synthetic datasets or contrived writing samples, we sought authentic undergraduate work that reflected the variety of styles instructors encounter in real assignments. The selected essays came from English-language coursework completed at Pan-European University "Apeiron" in Banja Luka, Bosnia and Herzegovina. They were chosen to represent two distinct stylistic registers – two formal academic essays and two conversational, narrative texts – on the premise that stylistic variation could influence how detection tools classify writing. The formal essays displayed a structured argument, academic vocabulary, and a conventional organization of ideas. The conversational ones incorporated personal voice, chronological sequencing, and features typical of non-native English writers, such as simplified syntax, direct address, and occasional grammatical quirks. By pairing these two registers, the study could explore whether the same tools would react differently to polished academic prose compared to more relaxed and personal styles.

All identifying information, including names, course details, and assignment codes, was removed prior to analysis. Since these essays had been submitted for coursework under conditions that allowed anonymized use for pedagogical and research purposes, they qualified as secondary data under institutional ethical standards. In this respect, the ethical considerations were not limited to the technical process of anonymization. They also extended to reflection on how the findings could be responsibly used or misused in academic contexts. Since AI-detection results may influence disciplinary action in real classrooms, it was crucial to approach this project with caution, ensuring that outcomes were presented as illustrative rather than definitive. A formal ethics board review was not required for this study, given its reliance on anonymized secondary material, but the research was nonetheless guided by established norms of confidentiality, fairness, and proportionality.

To provide a reasonable cross-section of commonly used detection platforms, four AI-detection tools were selected: Grammarly, QuillBot, BypassAI, and Phrasely. While each is known for its role in higher education, they differ in detection methods, user interfaces, and the level of feedback provided. Grammarly integrates AI-detection within a broader grammar and style platform, often presenting results as a probability score alongside editorial suggestions. QuillBot is primarily known for its paraphrasing function but also offers an AI-detection feature, making it a hybrid tool that reflects common student-facing software. BypassAI is notable for positioning itself in relation to detector evasion and counter-detection algorithms, providing an interesting counterpoint to mainstream academic tools. Phrasely is a more recent addition to the detection landscape, marketed as a stylistic and semantic analyzer. Selecting this mix allowed us to observe a range of algorithmic behaviors rather than capturing near-identical outputs from similar systems. It also enabled us to compare how commercial, pedagogical, and counter-detection oriented tools behaved when confronted with the same material.

Each essay was processed individually through each of the four detection tools. To avoid introducing variability from uncontrolled factors, all submissions were carried out in a single browser session, on the same device, and under identical network conditions. The default settings of each tool were maintained so that results reflected a typical user experience, rather than optimized or experimental configurations. No reformatting, rephrasing, or other alterations were made to the original texts beyond anonymization. The order of submissions was kept consistent across tools, and records were carefully maintained so that any later replication would follow the same procedural sequence.

For each submission, we recorded the primary classification – such as "AI-generated" or "likely human", along with any numerical probability scores, flagged segments, or explanatory notes provided by the tool. These results were entered into a comparative spreadsheet, which contained fields for the essay identifier, style category, detection tool, classification label, probability score (if available), and any

10.31902/LL.11.2025.2

qualitative feedback. To ensure accuracy, entries were double-checked, and screenshots of results were stored as a backup record. This structured recording not only allowed for an at-a-glance view of how tools aligned or diverged on a given essay but also created a transparent audit trail that could be revisited during analysis.

The analysis was deliberately descriptive rather than inferential. With a sample of four essays, the objective was not to produce statistically significant measures of tool performance but to examine agreement patterns and divergences in a way that could guide future research. Agreement was tallied simply by counting how many tools produced the same classification for a single essay, while discrepancies were noted for closer consideration. For example, if three tools agreed on a "human" classification and one labelled the same essay "AI-generated," this was marked as a divergence and interpreted in light of the essay's stylistic features. This approach resonates with early-stage validity studies, which often begin with descriptive mapping before progressing to more complex statistical modelling.

Observations during the process went beyond the recorded classifications. It quickly became apparent that the tools varied considerably in their user experience and the clarity of their outputs. Some tools displayed phrase-level highlights indicating sections deemed suspicious, while others produced only a binary verdict or an uncontextualized probability percentage. Processing speed also varied, with some tools returning results almost instantly and others requiring several minutes to complete an analysis. These differences, while not the focus of the study, have implications for how detection tools might be used in institutional contexts, where both the speed and the interpretability of results can affect decisions.

While the methodology provided a clear and replicable workflow, it also revealed the constraints of this pilot design. The small sample size inevitably limits generalizability, and the absence of repeated trials means that potential temporal variability – due to updates, algorithm retraining, or random fluctuations – was not captured. The opaque nature of all four systems also leaves unanswered questions about why certain essays were flagged while others were not, and whether these decisions were based on lexical, syntactic, or higher-level stylistic cues. Furthermore, the context in which the essays were written – by non-native English learners in Bosnia and Herzegovina – raises the possibility that cultural and linguistic factors influenced the classifications in ways not yet fully understood. This is itself an important contribution, since most existing evaluations of detectors focus on Anglophone contexts and overlook how multilingual student populations may be disproportionately affected.

Nevertheless, the consistency of the workflow, combined with the clarity of the recorded results, suggests that the approach could be scaled effectively in a larger, more systematic investigation. If expanded, the study would benefit from a broader essay corpus, ideally incorporating multiple academic disciplines, varying levels of language proficiency, and diverse cultural backgrounds. Running multiple trials for each essay-tool combination could yield stability measures, while inter-rater coding of outputs would add an extra layer of verification. With a sufficiently large dataset, performance metrics such as precision, recall, and false-positive/negative rates could be calculated, providing a more robust evaluation of detection reliability. Additional documentation, such as a protocol checklist detailing submission procedures, tool versions, and conditions of analysis, would further enhance reproducibility.

Seen through the lens of feasibility research, this methodology met its aims: it applied detection tools systematically to anonymized student coursework, generated interpretable comparative data, and illuminated both the strengths and the limitations of current detection practices. The results that follow should be read not as definitive performance benchmarks but as an illustrative snapshot of how these systems respond to different writing styles in a controlled, small-scale setting.

10.31902/LL.11.2025.2

4. Results

To examine our research question – how four AI-detection systems classify essays of different writing styles – we processed four anonymized undergraduate essays through Grammarly, QuillBot, BypassAI, and Phrasely. Two essays were written in a formal academic style, while the other two adopted a more conversational, narrative register. Full essay extract that we analyzed are included in Appendix A. Here, we introduce extracts from each essay to illustrate the stylistic variation under study, followed by a summary of detection outcomes.

Essay A, formal in tone, opens:

"The vision of standing in a classroom, a beacon of knowledge and inspiration for young minds, fills my heart with purpose." All four tools classified this essay as AI-generated, suggesting that elevated vocabulary and conventional academic structure act as strong triggers for detection systems.

Essay B continues in the same formal register:

"In the realm of critical pedagogy, authentic engagement elevates the learner beyond rote proficiency." This essay also received unanimous AI-generated classifications, reinforcing the trend observed in Essay A. By contrast, Essay C adopts a conversational style, beginning: "I trust this email finds you well. My name is ___ and I am writing to express my sincere interest in study abroad opportunities." Three detectors classified it as likely human, with only BypassAI flagging it as AI-generated. This divergence suggests that narrative framing and minor non-native features influence tool behavior in less predictable ways. Finally, Essay D, with its narrative and colloquial tone, begins: "Cindy lived with her stepmom and two stepsisters, and they were the worst." All four tools unanimously classified this essay as human-authored, suggesting that informal phrasing and story-like structure reduce the likelihood of false positives.

Table 1 summarizes the classification outcomes across all essays and tools:

Essay	Style	Grammarly	QuillBot	BypassAI	Phrasely	Agreement (of 4)
A	Formal	AI-generated	AI-generated	AI-generated	AI-generated	4
B	Formal	AI-generated	AI-generated	AI-generated	AI-generated	4
C	Conversational	Human-likelihood	Human-likelihood	AI-generated	Human-likelihood	3
D	Conversational	Human-likelihood	Human-likelihood	Human-likelihood	Human-likelihood	4

Across the dataset, inter-tool agreement was very high: 93.75 percent, with only one instance of disagreement. Yet, agreement does not necessarily indicate accuracy. Both Essay A and Essay B, which were included as anonymized coursework submissions, were uniformly flagged as AI-generated. Whether these texts were entirely human-authored or contained AI assistance cannot be determined within the scope of this study. What is clear is that features associated with polished academic writing – structured argument, formal vocabulary, and rhetorical sophistication – appear to serve as consistent triggers for detection algorithms. It is somewhat ironic that qualities traditionally encouraged in higher education can, under automated systems, be treated as suspicious rather than commendable.

Essay C, by contrast, highlighted the potential for divergence. Its conversational tone led three detectors to judge it as human-authored, yet BypassAI classified it as AI-generated. This outcome reflects findings in prior studies that non-native writing, marked by simplified syntax and reduced lexical variety, is more vulnerable to misclassification (Liang et al. 2023). In our case, the text's blend of informal

10.31902/LL.11.2025.2

phrasing and minor grammatical variation seems to have confused one detector while passing as "human" for the others.

Essay D, narrative in style and full of casual phrasing, was unanimously classified as human-authored. This suggests that narrative and colloquial registers are least likely to be flagged, even if they contain surface-level errors. The finding mirrors earlier claims that tools are sensitive to formulaic patterns but more forgiving toward personal or story-driven writing.

These results align with broader evaluations of detection systems. Weber-Wulff et al. (2023) reported that no detector exceeded 80 percent accuracy in large-scale testing, with particular weaknesses when texts were paraphrased or translated. The fact that Essays A and B were uniformly classified as AI-generated in our small study resonates with this broader pattern: consistency across detectors does not guarantee correctness. Indeed, unanimous classification raises a particular concern, since a student whose work is flagged by multiple platforms would face significant difficulty in contesting the outcome, even if the classification were open to doubt.

Bias against non-native writing deserves particular attention. Liang et al. (2023) demonstrated that 61 percent of TOEFL essays by non-native writers were misclassified as AI-generated, with about 20 percent of these flagged unanimously. Our mixed result for Essay C echoes this broader problem, suggesting that detectors may systematically confuse linguistic simplicity with machine authorship. This has troubling implications for equity, as international students and multilingual learners may be disproportionately penalized.

Additional observations were noteworthy from a procedural standpoint. Processing times varied between five and eight minutes per essay depending on the tool, with some returning results almost instantly and others requiring longer loading periods. Usability was inconsistent: a few systems offered phrase-level highlights and limited explanations, while others returned only a binary label or percentage score. These discrepancies matter in practice, as opaque verdicts undermine transparency and limit the ability of instructors or students to interpret and contest results. The "black box" character of these systems remains a recurring concern (Veale et al. 2018).

Finally, the ethical implications cannot be ignored. Systematic classification patterns that disproportionately target certain writing styles risk unfairly stigmatizing students, particularly those who are non-native speakers or who rely on formulaic academic prose. Such outcomes may not only affect grades but also erode trust in institutional assessment systems. When entire categories of writing are routinely flagged, academic integrity frameworks risk shifting from support and fairness toward suspicion and surveillance.

In sum, while inter-tool agreement in this study was high, the findings underscore a critical paradox: consistency among detectors does not necessarily mean reliability. The unanimous classification of Essays A and B as AI-generated illustrates how these systems may be reacting primarily to stylistic markers rather than providing a clear indication of actual authorship. This outcome highlights the need for educators, administrators, and policymakers to interpret detection outputs with caution, situate them within broader frameworks of academic ethics, and avoid overreliance on automated judgments in matters of integrity. Taken together, these results suggest that detection tools are less reliable indicators of authorship than they might initially appear. High agreement, rather than confirming accuracy, can in fact mask systematic biases against particular forms of writing, especially those associated with non-native authors. The findings therefore raise important questions about fairness, transparency, and the pedagogical risks of relying too heavily on automated detection. These issues will be explored further in the following Discussion, where we situate our observations within broader debates on academic integrity and the role of technology in higher education.

10.31902/LL.11.2025.2

5. Discussion

Our findings reveal a consistent pattern: formal, structured essays are reliably flagged as AI-generated, while conversational, narrative writing remains largely unflagged. Such consistency across detection tools underscores a persistent sensitivity to stylistic markers commonly associated with polished academic prose – even when those markers stem from genuine student work. This pattern raises concerns about systemic bias embedded in detection systems. For instance, a comprehensive evaluation of fourteen detection tools found none achieved more than eighty percent accuracy, especially when faced with paraphrased or translated content, which is a troubling limitation reflected in our observations (Weber-Wulff et al. 2023).

The bias is even more pronounced when considering non-native English writers. Liang et al. (2023) demonstrated that detectors mislabeled approximately sixty-one percent of TOEFL essays as AI-generated, with nearly twenty percent of samples unanimously flagged by all tools. Our Essay C mirrors this issue in miniature, receiving mixed human/AI classifications despite its clear conversational tone. Such results reinforce the danger of relying on detectors for quick judgments in linguistically diverse contexts. They also highlight how algorithmic outputs may intersect with linguistic hierarchies in higher education, reproducing disadvantages for students who are already navigating language barriers.

Moreover, emphasis on linguistic complexity may further disadvantage writers with limited vocabulary or less stylistic refinement. Liang and colleagues showed that enhancing lexical richness in non-native writing dramatically reduced misclassification, highlighting the difference between algorithmic detection and genuine writing voice (Liang et al. 2023). Without such refinements, detection tools may penalize students for linguistic simplicity or non-standard expression – a consequence that is both unfair and ethically troubling. In effect, detectors risk confusing markers of developing proficiency with signals of artificial authorship.

Our procedure also revealed discrepancies in the interpretability of detection results. While some tools offered helpful phrase-level feedback, others provided opaque labels or probability scores. This variability affects how easily users can understand, respond to, and trust the detection output – an important issue in academic settings where fairness and transparency are expected. A binary verdict without justification not only undermines due process in cases of suspected misconduct but also places instructors in the precarious position of making disciplinary decisions based on evidence they cannot meaningfully interrogate.

The broader implications extend beyond accuracy metrics. The growing pressure on institutions to police AI misuse has triggered an atmosphere of suspicion. Real-world cases show that students have been falsely accused based on vague stylistic cues—resulting in emotional distress and undermined trust in educational integrity systems. In this environment, dependence on flawed detection algorithms risks over-policing rather than safeguarding authenticity (The Guardian 2024). The potential chilling effect on student writing confidence is also significant. If learners fear that sophisticated or formulaic academic style will be flagged as AI-generated, they may deliberately simplify their prose, leading to a paradoxical decline in the very academic literacy institutions aim to promote.

The findings also raise important considerations for pedagogy and assessment design. High false-positive rates in structured academic writing suggest that the most “correct” or polished forms of student expression are the most vulnerable to suspicion. This tension undermines one of the central goals of higher education: to cultivate students’ ability to write persuasively, formally, and with academic rigor.

10.31902/LL.11.2025.2

If achieving this goal comes with the risk of algorithmic penalization, institutions must rethink how detection tools are positioned within academic integrity frameworks. Rather than treating detectors as arbiters of authorship, they might be better framed as supplementary signals – one source of information among many, to be weighed alongside contextual knowledge of the student's performance and writing history.

To uphold fairness and credibility, detection tools must be integrated within holistic academic integrity frameworks. This means combining technical detection with human judgment, incorporating process-oriented assignments, and enhancing AI literacy among students and faculty. Detectors may serve as initial indicators, but their outputs must remain provisional, not determinative. Educators should be trained not only to interpret detection reports critically but also to communicate their limitations transparently to students. Policies should clearly state that AI-detection scores are not definitive proof of misconduct but prompts for further review. Without such safeguards, institutions risk delegating ethical decision-making to systems that were never designed for such responsibility.

Future research should therefore evaluate detectors across more diverse corpora, including variations in register, proficiency, and cultural context. Multiple trials per essay, improved statistical modeling, and calibration efforts are needed to develop more equitable tools. This includes systematically testing for false positives in non-native writing, examining how small lexical or syntactic changes shift detector outputs, and developing open-source benchmarks that can serve as common points of comparison. Importantly, this work should not only assess accuracy but also interrogate fairness, explainability, and the broader consequences of tool use in educational environments.

In summary, while our pilot confirms tool consistency, it also reveals how stylistic bias may unfairly penalize student writing, including cases that could be entirely genuine. Detectors alone are insufficient guardians of integrity, particularly in diverse learning environments where linguistic variation is common. Ensuring equitable evaluation requires complementary strategies, education-focused policies, and the careful, transparent deployment of detection technologies. Institutions must balance the practical need to address potential AI misuse with the ethical obligation to protect student rights and foster authentic academic development. Only by embedding detection tools within broader pedagogical and policy frameworks – ones that value fairness, transparency, and student trust – can higher education responsibly navigate the challenges of an AI-saturated future.

6. Conclusion

This exploratory pilot study examined how four widely used AI-detection tools – Grammarly, QuillBot, BypassAI, and Phrasely – classified four anonymized undergraduate essays in English, written in distinct stylistic registers. Although the small corpus precludes generalization, the results offer valuable insights into the variability of tool performance, particularly in relation to writing style and the linguistic features of non-native authors. By situating the study within the broader context of academic integrity and fairness, the findings highlight both the promise and the limitations of detection technologies in higher education.

Our findings indicate a high overall rate of inter-tool agreement (93.75 percent), but also reveal that when tools agree, they can still converge on questionable classifications. In this small sample, the two more formal, formulaic essays were uniformly flagged as AI-generated, while those with more conversational or narrative features were consistently classified as human, aside from one divergence. Although we cannot verify the true authorship of the texts, this pattern aligns with concerns in the literature that AI-detection systems may embed and reproduce biases against non-native writing styles, where

10.31902/LL.11.2025.2

lexical simplicity and structural predictability are disproportionately penalized. Such tendencies are not merely technical limitations but raise serious ethical and pedagogical questions. If left unaddressed, they risk reinforcing inequities and eroding trust in academic integrity processes.

Beyond classification patterns, the study highlights notable differences in tool usability, transparency, and explanatory capacity. Some systems provide only a binary label or a numerical probability without substantive evidence for the decision, limiting the scope for pedagogical interpretation and due process in academic integrity cases. Others offer more detailed phrase-level highlighting, which, although not infallible, allows users to better understand and critique the classification. This “black box” opacity remains a central limitation in current AI-detection technology and underscores the importance of explainability and interpretability in educational tools.

From a methodological standpoint, the study confirms the feasibility of applying multiple detectors to authentic student work within a controlled protocol. However, it also underscores the need for more robust designs in future research: larger and more diverse corpora, repeated trials to test stability over time, inter-rater validation, and statistical performance metrics such as precision, recall, and false-positive/negative rates. Expanding the scope beyond English, and including writing samples from multiple linguistic and disciplinary contexts, would also provide valuable comparative perspectives.

For educators and institutions, the findings carry several implications. AI-detection tools can play a role in integrity enforcement, but they should be interpreted within broader frameworks of academic ethics, transparency, and fairness. Over-reliance on automated outputs risks disadvantaging certain student populations and undermining trust in assessment processes. Integrating detection tools with pedagogical strategies – such as process-based writing assignments, reflective commentary, oral defense, or iterative feedback – may offer a more balanced approach that both deters misconduct and fosters genuine skill development. Importantly, academic integrity should not be reduced to technical policing; it must remain grounded in teaching and learning practices that build student agency and responsibility.

While the present work is necessarily limited in scope, it contributes to the growing body of empirical evidence on AI-detection reliability in real educational contexts. In doing so, it responds directly to calls for small-scale, context-specific studies that illuminate the practical and ethical challenges of AI integration in higher education. The study demonstrates that even modest, exploratory research can yield insights with institutional and policy relevance, especially in identifying where systemic weaknesses might produce unfair outcomes.

It is important to acknowledge that the study was based on only four essays, which severely restricts the generalizability of the findings. The observations presented here are therefore illustrative and exploratory, intended to generate hypotheses about detection behavior rather than to establish definitive conclusions. The results capture a narrow snapshot of tool performance under specific contextual and linguistic conditions; they should not be extrapolated beyond those boundaries. Recognizing these constraints clarifies the pilot character of the research and situates its contribution as methodological groundwork for future, larger-scale investigations.

Looking ahead, the rapid pace of AI development suggests that detectors will continue to evolve, just as student practices will adapt. Continuous monitoring, critical evaluation, and dialogue between researchers, educators, and developers are therefore essential. By situating these findings within academic integrity frameworks, the study reaffirms the importance of cautious, informed adoption of detection technologies, coupled with ongoing adaptation of assessment practices in an AI-saturated learning environment. Ultimately, the key lesson of this pilot is that detection should not serve as the sole arbiter of authenticity. Instead, it needs to be part of broader approaches that place student learning, fairness, and trust at the center of academic practice.

10.31902/LL.11.2025.2

References

Anderson, James, Sarah Lee, and Robert Chang. 2023. “Adversarial Robustness in AI Text Detection: Limits and Opportunities.” *Journal of Educational Technology Research* 45 (3): 215–32.

Billingham, Sam A., Nicky Whitehead, and Joanne J. Julious. 2013. “An Audit of Sample Sizes for Pilot and Feasibility Trials Being Undertaken in the United Kingdom.” *BMC Medical Research Methodology* 13 (104): 1–6. https://doi.org/10.1186/1471-2288-13-104.

Eldridge, Sandra M., Christine L. Lancaster, Gillian A. Campbell, Lehana Thabane, Claire Hopewell, and Sally Hopewell. 2016. “Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework.” *PLoS One* 11 (3): e0150205. https://doi.org/10.1371/journal.pone.0150205.

Evangelista, Patrick. 2025. “AI Detection Tools and Academic Integrity: A Systematic Review.” *Computers and Education Review* 17 (1): 45–59.

Faber, Jose M., and Jorge R. Fonseca. 2014. “How Sample Size Influences Research Outcomes: Pilot Study Perspectives.” *International Journal of Social Research Methodology* 17 (5): 523–34. https://doi.org/10.1080/13645579.2012.729392.

Herald Sun. 2024. "Universities Warned of Risks in Overreliance on AI Detection." February 14, 2024.

Inside Higher Ed. 2024. "Paraphrasing and the Weakness of AI Detectors." March 4, 2024.

Lakens, Daniël. 2020. “Sample Size Justification.” *Collabra: Psychology* 6 (1): 1–13. https://doi.org/10.1525/collabra.27637.

Liang, Weixin, 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.

Packback. 2025. “AI-Resilient Pedagogies in Higher Education.” White Paper.

Provost UKY. 2024. “AI Detection and International Student Equity.” Policy Brief. University of Kentucky.

Rafiq, Mohammad, and Nadia Khalid. 2025. “Integrating AI Detection and Academic Literacy: A Combined Approach to Academic Integrity.” *International Journal for Educational Integrity* 21 (1): 12–29.

Sadasivan, Akshay, John T. Hancock, and Marie Dubois. 2023. “Bypassing AI Detection via Minimal Paraphrasing.” *Computational Linguistics Review* 39 (4): 301–18.

The Guardian. 2024. "AI Cheating Is Overwhelming the Education System – But Teachers Shouldn't Despair." August 24, 2024. https://www.theguardian.com/commentisfree/article/2024/aug/24/ai-cheating-chat-gpt-openai-writing-essays-school-university

Veale, Michael, Max Van Kleek, and Reuben Binns. 2018. “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public-Sector Decision-Making.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery. https://doi.org/10.1145/3173574.317401.

Weber-Wulff, Debora. 2023. “Testing of AI-Generated Text Detectors.” Working Paper. Berlin: HTW Berlin.

Appendix A – Student Essay Extracts

The following extracts represent the complete texts of the four anonymized undergraduate essays used in this study. They are reproduced exactly as written by the students, with only identifying information removed. Presenting these texts in full serves two purposes: first, it allows readers to evaluate for themselves the stylistic features – such as tone, vocabulary, and grammatical structure – that may influence AI-detection outcomes; second, it provides transparency in how the data align with the methodology and findings reported earlier. The essays are grouped according to the two stylistic categories defined in the study: formal academic style and conversational/narrative style.

Essay A – Formal Academic Style

The vision of standing in a classroom, a beacon of knowledge and inspiration for young minds, fills my heart with purpose. I believe that education is the cornerstone of societal advancement, and as a future

10.31902/LL.11.2025.2

teacher, my mission is to nurture critical thinking, curiosity, and a lifelong love of learning. In an increasingly globalised world, the ability to communicate ideas effectively in English has become not only an academic requirement but also a professional necessity. My dedication to mastering this language is grounded in the desire to provide my future students with the tools they need to succeed in diverse environments.

Throughout my academic journey, I have consistently sought opportunities to improve my English proficiency. This includes participating in language workshops, attending international conferences, and engaging with academic literature. Such experiences have enriched my vocabulary, honed my analytical skills, and deepened my understanding of cross-cultural communication. I am confident that my commitment to continual improvement will serve as a model for my students, encouraging them to embrace challenges and persist in their own learning journeys.

Style Note: This essay's elevated vocabulary, formal register, and tightly structured argument mirror many features common to polished academic prose. All four AI-detection tools classified it as AI-generated, illustrating the study's finding that formal academic style will likely trigger uniform positives.

Essay B – Formal Academic Style

In the realm of critical pedagogy, authentic engagement elevates the learner beyond rote proficiency. Language learning, when approached holistically, should integrate not only grammatical accuracy but also intercultural competence and self-expression. I have long been interested in how technology can enhance these goals, particularly through interactive platforms that simulate real-world communication scenarios.

My experience in academic settings has taught me that motivation thrives when learners perceive the relevance of their studies. Thus, I strive to design tasks that resonate with students' interests, while also exposing them to perspectives that challenge their assumptions. By fostering a classroom culture that values dialogue, empathy, and reflective thinking, I aim to equip my students with both linguistic skills and the ability to navigate complex social realities.

Style Note: Like Essay A, this text employs a formal academic register and thematically coherent argumentation. All four tools flagged it as AI-generated, reinforcing the correlation between formulaic scholarly tone and consistent detection across systems.

Essay C – Conversational / Narrative Style

I trust this email finds you well. My name is [Name removed], and I am writing to express my sincere interest in study abroad opportunities at your university. I have always dreamed of experiencing life in another country, meeting people from different cultures, and improving my English skills through daily interaction.

Back in my hometown, I have been involved in a variety of community projects, from volunteering at local schools to organising cultural events. These experiences have taught me the importance of communication, teamwork, and adaptability. While I know that studying abroad will bring challenges, I am confident that my enthusiasm and willingness to learn will help me overcome them. I am particularly excited about the possibility of joining student clubs and participating in cultural exchange programs.

Style Note: This essay's conversational tone, personal narrative, and moderate lexical range led three tools to classify it as human-authored, with only BypassAI flagging it as AI-generated. This partial disagreement exemplifies the influence of style on classification variability.

10.31902/LL.11.2025.2

Essay D – Conversational / Narrative Style

Cindy lived with her stepmom and two stepsisters, and they were the worst. Every day they made her do all the chores around the house, from scrubbing the floors to washing the windows. She never complained, but deep down she wished for a different life. One afternoon, while she was cleaning the garden, a small bird flew down and started to sing. Cindy stopped to listen, and for the first time in a long while, she felt a little hope.
The next day, she found a note tucked under her pillow. It said, “Pack your things. Tomorrow you leave for the city.” She didn’t know who had sent it, but her heart raced with excitement. Cindy had no idea what awaited her, but she knew it was the start of something new.
Style Note: With its simple syntax, storytelling focus, and absence of academic conventions, this narrative was unanimously classified as human-authored by all four tools, supporting the finding that informal, personal writing is less prone to false positives.

Acknowledgments

The author wishes to thank colleagues and students at Pan-European University “Apeiron”, Banja Luka, for their support and for permitting the anonymized use of student essays. Constructive feedback from peer reviewers is also gratefully acknowledged.

Data Availability Statement

The data supporting this study consist of anonymized undergraduate essays. These materials are not publicly available in order to protect student confidentiality but may be shared in anonymized form upon reasonable request to the corresponding author.

Biography of the Author

Katarina Držajić Laketić is an Associate Professor of English Language and Literature and Dean of the Faculty of Philological Sciences at Pan-European University “Apeiron” in Banja Luka, Bosnia and Herzegovina. She earned her PhD in literature and her research spans modern English literature, translation studies, and applied linguistics. Her recent publications explore narrative and psychological dimensions in twentieth-century fiction, as well as the pedagogical and ethical implications of artificial intelligence in academic writing and translation practice.

10.31902/LL.11.2025.2