https://doi.org/10.36315/2026inpact027
Psychological Applications and Trends 2026

PERCEIVED AND ACTUAL DETECTABILITY OF AI-GENERATED STUDENT WORK ACROSS THREE COHORTS

Irena Miljković Krečar, & Maja Kolega
VERN'University (Croatia)

Abstract

The everyday use of generative artificial intelligence by students has made it increasingly difficult to know who actually authored an academic homework assignment. This paper presents findings from a three-year, three-cohort mixed-methods research program conducted between 2023 and 2025, examining both perceived and actual ability to distinguish between student-authored and AI-generated academic assignments. The first research wave (2023), previously published, examined university teachers' (N = 12) ability to identify AI-generated versus student-authored assignments across multiple course types. Results showed detection accuracy only slightly above chance level (M = 53.75%), accompanied by moderately high confidence in judgments. One year later, a second wave (2024) replicated the procedure with a convenience sample of educational psychologists (N = 16), yielding comparable results: detection accuracy remained low (M = 52.31%), while confidence remained moderately high. The third research wave (2025) focused on first-year university students (N = 27) from two study programs (business informatics and cyber communication). A mixed-methods design combined an anonymous questionnaire with four focus groups. Quantitative findings indicate that students assessed teachers' detection accuracy with striking precision (M = 52.73%), closely matching actual performance observed in the two previous waves. At the same time, students estimated their own ability to produce AI-assisted assignments that would remain undetected at a substantially higher level (M = 71%). Focus-group discussions further revealed that, despite this confidence, students themselves struggled to reliably identify AI-generated work when presented with concrete examples. Participants agreed that AI can be a valuable tool to support learning, but not a substitute for producing entire assignments. Students emphasize the need to develop their own writing, research, and critical thinking skills to ensure the quality of their learning. Because students may struggle to resist the temptation of taking the easier path, institutions need to offer clear guidance and support that promotes responsible and learning-oriented uses of AI. Taken together, the findings reveal a growing asymmetry in higher education: students appear to have developed high confidence in their ability to produce AI-assisted work that remains undetectable, along with an experience-based sense that teachers generally struggle to identify such work—an impression that closely reflects teachers' actual detection performance. Although students are aware of its possible negative consequences, many still report continued and intended use of AI tools. These findings raise urgent questions for assessment design, clearer institutional guidance, and the sustainability of current approaches to academic integrity.

Keywords: *Academic integrity, AI detectability, higher education, mixed-methods research.*

1. Introduction

Students' independent work in the form of essays, projects, research tasks, reflections on readings, and similar take-home assignments is an essential part of higher education. Within the European Credit Transfer and Accumulation System (ECTS), credits are based on overall student workload and imply that students engage in independent work outside the classroom in order to achieve intended learning outcomes. Through written take-home assignments, students develop writing competence and an academic style; strengthen critical thinking, information literacy, and research skills; and foster metacognition and reflectivity, self-regulation and time management, and—no less importantly—ethical conduct and academic integrity. Because these tasks are completed outside direct supervision, they also depend on trust and transparent authorship, which makes them vulnerable to different forms of misconduct. At the same time, it is not new that some students attempt to reduce effort by using shortcuts. Research on cheating in higher education shows that academic misconduct is shaped by both individual and contextual factors, and that it requires systematic prevention rather than punishment alone (Chiang, Zhu, & Yu, 2022; Miles, Campbell, & Ruxton, 2022; Tight, 2024; Waltzer & Dahl, 2023).

However, with the rapid development of artificial intelligence (AI), particularly ChatGPT, some forms of cheating have become easier and faster to carry out, and challenges to academic integrity are becoming increasingly significant (Cotton, Cotton, & Shipway, 2024). It quickly became apparent that AI could generate and write many types of assignments within minutes, while educators often have difficulty distinguishing students' work from text produced by ChatGPT (Busch & Hausvik, 2023; Fleckenstein et al., 2024; Miljkovic Krečar, Kolega, & Jurčec, 2024; Tyler, Coble, Gagnon, & Howard, 2025; Waltzer, Cox, & Heyman 2023). Fleckenstein et al. (2024), for example, report that current AI can produce essays that are difficult for teachers to detect, while teachers may still feel confident in their judgments. In this context, relying on "AI-like style" cues becomes problematic, because such cues are inconsistent and can also penalize honest students, including non-native speakers. In addition, reviews of AI-generated text detection tools indicate that detectors are often not reliable enough to serve as stand-alone evidence of misconduct and may be biased in real-world settings (Gotoman, Luna, Sangria, Santiago, & Barbuco, 2025).

These developments have triggered an ongoing debate in higher education: should AI use be banned, or should it be integrated into learning outcomes? Some authors argue that integration is a more realistic and educationally useful approach than blanket prohibition, provided that students are guided to use AI critically and transparently (Yu, 2023). Against this background, the present paper examines both actual and perceived detectability of AI-generated student work. We build on a three-wave research program (2023–2025) that uses comparable materials and procedures across different cohorts. The multi-year design is particularly informative because it captures whether—and how quickly—students and teachers adapt to rapidly diffusing AI tools and to emerging institutional norms, rather than treating detectability as a static, one-time phenomenon. Understanding the gap between perceived and actual detectability matters because it may shape student decisions, teacher practices, and the long-term sustainability of take-home assessment formats. If students believe that detection is weak, this can encourage risky behavior and undermine fairness; conversely, if teachers believe they can "recognize AI" but in fact cannot, then assessments of learning outcomes rest on incorrect assumptions. By combining performance data with perceptions and norms, we aim to move the discussion from "Can we detect AI writing?" toward "How should higher education respond when detection is weak—without sacrificing learning outcomes or fairness?"

2. Method

2.1. Wave 1 (2023): University teachers (in-house workshop)

Results from wave 1 have been published (Miljković Krečar, et al., 2024); therefore, only a brief description of the procedure is provided here to enable comparison with wave 2.

Based on homework prompts from teachers' courses and examples of student submissions produced before November 2022, eight task types (undergraduate and graduate level courses) were selected and compiled into 25 short texts, combining authentic student submissions with AI-generated responses (approximately 50–66% AI per task type). Most AI texts were generated with GPT-3.5 using the original teacher prompts; selected more complex tasks were generated with GPT-4. All texts were formatted uniformly. The study was conducted as an in-person classroom workshop. Teachers received folders containing texts from one task type and, for each text, indicated: (1) whether it was student-authored or AI-generated, (2) their confidence level (1–5), and (3) cues used for the decision. Folders were rotated among participants, and responses were collected in a way that minimized mutual influence. The activity lasted approximately 45 minutes, and 5–9 evaluations per paper were obtained. Prior to the task, teachers rated their expected ability to recognize AI-generated work (1–5; M = 3.0).

2.2. Wave 2 (2024): Educational psychologists (conference workshop)

Wave 2 was conducted in November 2024 as part of a workshop at an annual psychology conference, with a convenience sample of psychologists (N = 16; all female) employed in educational institutions.

To fit the workshop time constraints, five of the original eight course types were selected (6 authentic student texts plus 9 AI-generated texts). Task selection was guided by two criteria: (a) tasks that proved difficult to detect in wave 1 (to test whether they remained resistant), and (b) tasks whose content domain could be reasonably evaluated by non-specialists. The procedure mirrored wave 1: participants classified each text as student vs. AI, reported confidence on the same scale, and provided brief written cues. Prior to judging, participants rated their expected effectiveness in the task (1–5 scale; M = 2.69)

2.3. Wave 3 (2025): University students

The third research wave was conducted in December 2025 with first-year university students at VERN' University (N = 27) enrolled in two undergraduate programs (Business Informatics; Cyber Communication). The study used a mixed-methods design combining (a) an anonymous student questionnaire and (b) four focus groups.

2.3.1. Questionnaire (quantitative component). The anonymous questionnaire assessed: (1) frequency and modes of using generative AI tools for homework, (2) prior negative experiences (e.g., suspected plagiarism), (3) perceived clarity of institutional and secondary-school guidance regarding acceptable AI use, (4) students' estimates of teachers' ability to detect AI-assisted work (0–100%), (5) perceived teacher accuracy in detection (categorical item), and (6) students' estimates of their own ability to produce AI-assisted work that would remain undetected (0–100%). Open-ended questions captured perceived teacher detection strategies, future intentions, and perceived sanctions.

2.3.2. Focus groups (qualitative component). Four focus groups (approximately 6–7 students per group, intentionally mixed across programs) followed a semi-structured protocol. Students first attempted to classify a set of short assignments as student-authored vs. AI-generated, explained their criteria, and then discussed acceptable vs. unacceptable AI use across assignment types (essay/reflection, research tasks, technical/computational tasks), including proposals for "fair and reasonable" institutional rules. Questionnaire data were analyzed descriptively (frequencies, means, and standard deviations). Focus-group data will be reported as a thematic analysis focusing on (a) perceived detection cues and (b) students' normative boundaries of AI use by assignment type. Participation was voluntary and confidential, and all students provided informed consent prior to the focus-group discussion. They could withdraw at any time without any consequences.

3. Results

3.1. Professors' detection accuracy and confidence

Across both evaluator cohorts, the ability to distinguish student-authored from AI-generated homework remained close to chance level. In wave 1, overall detection accuracy was M = 53.75%, with moderately high confidence (M = 3.76) and in wave 2, overall accuracy across five selected task types was M = 52.31%, with comparable confidence (M = 3.51; Table 1).

When restricting wave 1 results to the same five task types used in wave 2, teachers' accuracy decreased to M = 49.40% while confidence remained high (M = 3.81), indicating weak calibration (high confidence despite near-chance performance).

A further practical concern is the asymmetry between correctly identifying AI-generated texts and avoiding false accusations. In wave 2, accuracy for classifying AI-generated texts fell slightly below chance (AI texts: M = 48.82%), while accuracy for classifying genuine student texts was higher (student texts: M = 57%), implying a non-trivial risk of both missed AI cases and erroneous "AI" labeling of authentic student work depending on task characteristics and cues used.

Table 1. Detection accuracy and confidence across three waves.

Wave (year)	Sample (N)	Task (short)	Objective & Perceived accuracy (%)	Confidence in accuracy / undetectability (M /%)
Wave 1 (2023)	Teachers (12)	Classify texts (student vs. AI) across 8 task types	AI-text: 47.94 student-text: 63.75 TOTAL: 53.75	3.76
Wave 2 (2024)	Educational psychologists (16)	Same classification task; 5 selected task types	AI-text: 48.82 student-text: 57 TOTAL: 52.31	3.51
Wave 3 (2025)	Students (27)	Questionnaire + Focus groups	Perceived teachers accuracy: 52,7	Confidence in self undetectability: 71%

In both workshop cohorts, evaluators relied mainly on surface linguistic cues—style, vocabulary level, and perceived naturalness of phrasing—when classifying texts as student-authored vs. AI-generated. Teachers in wave 1 often associated "student" authorship with personal voice and typical student errors, whereas "AI" was linked to overly polished structure and "too professional" language; wave 2 comments showed a similar pattern and frequently noted uncertainty.

3.2. Students' perceptions

Students most frequently described AI use as "help" in the form of clarifying tasks, generating directions/structure, and saving time; a smaller portion explicitly mentioned using AI to generate direct solutions. Most students reported having used AI for homework at least occasionally: 55.6% reported using

them very often (15/27), 33.3% a few times (9/27), and 11.1% never (3/27). Across the sample, students estimated teachers' ability to detect AI-assisted work at approximately chance level (M ≈ 52%, Median = 50%). In contrast, students estimated their own ability to produce AI-assisted work that would remain undetected substantially higher (M = 71.22, SD = 21.29; Median = 70). This pattern indicates a pronounced asymmetry between perceived teacher detectability and students' perceived "undetectable performance" capability. When asked how accurate teachers are in detecting AI use, the modal response was that teachers are only partially accurate (15/27), followed by "mostly accurate if they try" (8/27), while a minority believed detection is not accurate (4/27). As teacher detection strategies, students most often referenced automated detectors/software and oral follow-up checks. Suggested sanctions most often involved grade penalties (e.g., failing the assignment), resubmission requirements, and warnings; a minority mentioned disciplinary exclusion.

The focus-group analysis highlights what students themselves see as most relevant in the context of AI. Students emphasized that AI-supported text is hard to distinguish from human writing: *"If a student makes an effort, it is impossible to recognize whether it is their work or AI. Only if they literally copy the text without additional technical editing and without checking the data, then it is easy to see. Or if the professor knows the student well, then they can recognize whether it is the student's work or not."* Still, participants described several cues they associate with AI output, especially related to voice and coherence. They contrasted human writing with AI by stating that *"students write with a more emotional tone*," while AI was described as polished but emotionally flat: "*There are no emotions, as if it were a script."* Regarding surface features of writing quality, students associated human work with more visible mistakes and less consistent formatting, while AI was seen as "*too neat*".

Students also drew relatively clear boundaries of acceptable AI use across task types. They described AI as acceptable for generating ideas, supporting structure, improving language, and early exploration, but emphasized that students must write and think independently. For computational tasks, students generally agreed that AI can support calculations, but interpretation must remain the student's responsibility, and outputs should be verified. Views were mixed regarding programming. Some participants warned that AI-generated code can be difficult to understand and debug: *"In many cases, the code AI writes will either be a bit more advanced, or the student will need a very long time to navigate it because they did not write it themselves; but when they wrote it themselves, they immediately know where an error is and can fix it quickly."* For issues of major social importance— "*especially ethical topics such as abortion*"—participants felt that relying on AI is not appropriate because it lacks depth and quality; in their view, students must develop and articulate such arguments themselves. Finally, some students questioned the importance of writing skills today: *"Today it is more important to know how to communicate and present well; writing is not that important—this is the past now"*, while others emphasized time pressure: *"AI saves time, and time is the most important thing today "*. Participants also expressed a strong preference for keeping creativity tasks for humans: *"No matter how much AI advances and produces excellent work, it will never surpass a human."* In terms of academic integrity, students argued that the core issue is not whether AI was used, but whether the student contributed and understands the topic. They suggested introducing a course on advanced and responsible AI use, so they can benefit from available tools while simultaneously developing their own competencies: *"There are already various tutorials on how to use it to an extent we cannot even imagine."* Students are aware that AI-related competencies will be increasingly demanded in the labour market, and they expect universities to teach them more advanced ways of using AI for different purposes—not only to write seminar papers: *"Let's be realistic—by the time we enter the labour market, AI will already replace us in part."*

Overall, participants questioned their own self-discipline and predicted that more than 70% would *"take the path of least resistance"* if there were no consequences, while still stressing that they do not want to obtain a degree without genuinely learning. They preferred assessment designs that make learning visible and reduce unsupervised take-home writing—such as in-class drafting of seminar papers, more practical and project-based tasks, and closer supervision through progress monitoring and timely feedback.

4. Conclusion

Our findings strongly suggest that teachers' ability to detect AI-assisted writing has not improved since generative AI became widely accessible. Across two measurement points one year apart, evaluators classified student versus AI-generated homework at near-chance levels while reporting moderate confidence in their judgments. This aligns with a growing body of evidence showing that educators struggle to reliably distinguish AI from student writing (e.g., Fleckenstein et al., 2024, Miljković Krečar & Pavlin-Bernardić, 2026), particularly when judging AI-generated texts (Tyler, Coble, Gagnon, & Howard, 2025).

One year later, students appear to have learned this "weak point" of the system: they estimated teacher detectability at a similarly moderate level while expressing high confidence in their own ability to

produce undetected AI-assisted work. In other words, the system has not become better at "seeing" AI in text, whereas students have become better at working around that expectation. This is why a forensic mindset—policing "AI-like" style—offers little practical value. Institutional policies and procedures, especially assessment practices, should assume low detectability and respond by strengthening the link between submitted work and learning. Concretely, this means designing assignments where the process is visible (staged drafts or checkpoints), where students can be asked to account for their choices (short oral follow-ups or brief reflections tied to course content, including—where appropriate—transparent descriptions of how AI was used), and where tasks are sufficiently contextualized that generic AI output is less useful without genuine engagement. Alongside task design, policy can help by being specific rather than moralizing: define what counts as acceptable assistance for a given task, how disclosure should look, and how grading criteria treat AI-supported work. This shifts the emphasis from "catching" to making expectations workable while protecting fairness—an approach more consistent with institutional integrity values (Fishman, 2014; Cotton, Cotton, & Shipway, 2024) than an arms race focused on chasing unreliable stylistic 'tells' that can be easily adjusted by prompting and revision.

Limitations should also be stated openly: samples were small and convenience-based, and the wave 1 and wave 3 cohorts came from a single institution. Because both AI models and student practices change quickly, the next step is replication with larger samples and longitudinal tracking of detectability, norms, and classroom adaptations over time.

References

Busch, P. A., & Hausvik, G. I. (2023). *Too Good to Be True? An Empirical Study of ChatGPT Capabilities for Academic Writing and Implications for Academic Misconduct.* Paper presented at the Twenty-ninth Americas Conference on Information Systems, Panama.

Chiang, F. K., Zhu, D., & Yu, W. (2022). A systematic review of academic dishonesty in online learning environments. *Journal of Computer Assisted Learning, 38*(4), 907-928. https://doi.org/10.1111/jcal.12656

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Fishman, T. (2014). The Fundamental Values of Academic Integrity. Second Edition (International Center for Academic Integrity). Retrieved November 12 2025, from: https://www.academicintegrity.org/wp-content/uploads/2017/12/Fundamental-Values-2014.pdf

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, *6*, 100209. https://doi.org/10.1016/j.caeai.2024.100209

Gotoman, J. E. J., Luna, H. L. T., Sangria, J. C. S., Santiago Jr, C. S., & Barbuco, D. D. (2025). Accuracy and reliability of AI-generated text detection tools: a literature review. *American Journal of IR*, *4*(1), 1-9. https://doi.org/10.54536/ajirb.v4i1.3795

Miles, P. J., Campbell, M., & Ruxton, G. D. (2022). Why students cheat and how understanding this can help reduce the frequency of academic misconduct in higher education: A literature review. *Journal of Undergraduate Neuroscience Education*, *20*(2), 150-160. https://doi.org/10.59390/LXMJ2920

Miljković Krečar, I., Kolega, M., & Jurčec, L. (2024). Perception of ChatGPT Usage for Homework Assignments: Students' and Professors' Perspectives. *IAFOR Journal of Education, 12*(2), 33-60. https://doi.org/10.22492/ije.12.2.02

Miljković Krečar, I., & Pavlin-Bernardić, N. (2026, in press). Appropriate and inappropriate use of artificial intelligence language models by students: Teachers' (self-)perceptions and experiences across educational levels. *Croatian Journal of Education*.

Tight, M. (2024). Challenging cheating in higher education: a review of research and practice. *Assessment & Evaluation in Higher Education*, *49*(7), 911–923. https://doi.org/10.1080/02602938.2023.2300104

Tyler, D., St. George Coble, S., Gagnon, A., & Howard, J. (2025). Assessing educators' ability to identify AI-generated student submissions: An experimental vignette study. *Journal of Criminal Justice Education*. Advance online publication. https://doi.org/10.1080/10511253.2025.2506435

Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Human Behavior and Emerging Technologies*, 1-9. https://doi.org/10.1155/2023/1923981

Waltzer, T., & Dahl, A. (2023). Why do students cheat? Perceptions, evaluations, and motivations. *Ethics & Behavior, 33*(2), 130–150. https://doi.org/10.1080/10508422.2022.2026775

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, *14*, 1181712. https://doi.org/10.3389/fpsyg.2023.1181712