Scientific and technological research article

How to cite: Ventura Ramos, P. E., Memije Alarcón, N. Y., & Zaragoza Martínez, J. (2026). Turnitin false positives and their management in the Law Degree Program at the Autonomous University of Guerrero. *Estrategia y Gestión Universitaria*, 14, e9232. https://doi.org/10.5281/zenodo.23671197

Received: 01/09/2026
Accepted: 16/10/2026
Published: 02/11/2026

Corresponding author:

ventura-eliza31@hotmail.com

Conflict of interest: the authors declare that they have no conflict of interest, which may have influenced the results obtained or the proposed interpretations.

Perla Elisabeth Ventura Ramos [1]
Universidad Autónoma de Guerrero
https://orcid.org/0000-0001-8680-1703
ventura-eliza31@hotmail.com
México

Norma Yadira Memije Alarcón [2]
Universidad Autónoma de Guerrero
https://orcid.org/0000-0002-3402-1112
profesmx2023@gmail.com
México

Jesús Zaragoza Martínez [3]
Universidad Autónoma de Guerrero
https://orcid.org/0000-0003-3685-7578
jesusmar63@msn.com
México

Turnitin false positives and their management in the Law Degree Program at the Autonomous University of Guerrero

Falsos positivos de Turnitin y su gestión en la Licenciatura en Derecho de la Universidad Autónoma de Guerrero

Falsos positivos do Turnitin e sua gestão no Curso de Direito da Universidade Autônoma de Guerrero

Abstract

Introduction: the use of artificial intelligence detection tools such as Turnitin has become common in higher education institutions to assess the originality of academic work. However, these tools present significant limitations, particularly in identifying texts written by humans. **Objective:** to analyze false positives in AI detection by Turnitin in the Law Degree Program at the Autonomous University of Guerrero. **Method:** a mixed-methods approach was employed, including surveys, documentary analysis of similarity reports, and interviews with faculty members. **Results:** between 10% and 20% of entirely human-written texts were incorrectly flagged as AI-generated. Some educational institutions use these percentages as the sole criterion for sanctioning students or rejecting academic work. **Conclusion:** AI detectors should not be used as the sole evidence of academic misconduct. A comprehensive evaluation framework is proposed that combines human review, appeal procedures, and alternative assessment methods. This study provides empirical evidence from the Mexican and Latin American context to inform fairer and more effective institutional policies.

Keywords: Turnitin, artificial intelligence, false positives, higher education, academic integrity

Resumen

Introducción: el uso de detectores de inteligencia artificial como Turnitin se ha vuelto común en instituciones de educación superior para evaluar la originalidad de trabajos académicos. Sin embargo, estas herramientas presentan limitaciones importantes, especialmente en la identificación de textos escritos por humanos.

This is an Open Access article under the license CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Objetivo: analizar los falsos positivos en la detección de IA con Turnitin en la Licenciatura en Derecho de la Universidad Autónoma de Guerrero. **Método:** enfoque mixto que incluye encuestas, análisis documental de informes de similitud y entrevistas a docentes. **Resultados:** entre el 10% y 20% de los textos 100% humanos fueron marcados erróneamente como generados por IA. Algunas dependencias educativas utilizan estos porcentajes como único criterio para sancionar estudiantes o rechazar trabajos. **Conclusión:** los detectores de IA no deben usarse como única evidencia de mala conducta académica y se propone un marco de evaluación integral que combine revisión humana, procesos de apelación y evaluación alternativa. Este estudio aporta evidencia empírica desde el contexto mexicano y latinoamericano para informar políticas institucionales más justas y efectivas.

Palabras clave: Turnitin, inteligencia artificial, falsos positivos, educación superior, integridad académica

Resumo

Introdução: o uso de detectores de inteligência artificial, como o Turnitin, tornou-se comum nas instituições de ensino superior para avaliar a originalidade de trabalhos acadêmicos. No entanto, essas ferramentas apresentam limitações importantes, especialmente na identificação de textos escritos por seres humanos. **Objetivo:** analisar os falsos positivos na detecção de IA pelo Turnitin no curso de Direito da Universidade Autônoma de Guerrero. **Método:** foi adotada uma abordagem de métodos mistos, incluindo questionários, análise documental de relatórios de similaridade e entrevistas com docentes. **Resultados:** entre 10% e 20% dos textos totalmente produzidos por humanos foram incorretamente classificados como gerados por IA. Algumas unidades educacionais utilizam esses percentuais como único critério para sancionar estudantes ou rejeitar trabalhos acadêmicos. **Conclusão:** os detectores de IA não devem ser utilizados como única evidência de má conduta acadêmica. Propõe-se um modelo de avaliação abrangente que combine revisão humana, processos de recurso e formas alternativas de avaliação. Este estudo fornece evidências empíricas a partir do contexto mexicano e latino-americano para subsidiar políticas institucionais mais justas e eficazes.

Palavras-chave: Turnitin, inteligência artificial, falsos positivos, ensino superior, integridade acadêmica



Introduction

University management today faces the challenge of integrating artificial intelligence (AI) into the evaluation of teaching and research across each university community. Turnitin was designed to verify the originality of written work, and higher education institutions have for some time used it to screen submissions. From an educator's perspective, Arteaga (2023) notes that generative AI is applied in personalized tutoring, automated assessment, the adaptation of materials, and plagiarism identification, the last of which is a particularly significant use. All of this unfolds against a backdrop of persistent inequalities, among them gaps in infrastructure, uneven literacy levels, data-management practices, and ethical considerations. Educational institutions run into these tensions regularly.

The rapid incorporation of AI tools into Latin American higher education has changed how students are taught and how they learn. In educational institutions, it has benefited areas such as academic performance and the development of digital skills (González Torres et al., 2024; Tiglla, 2025; Cándano Pérez et al., 2025). Occasionally, however, so-called false positives occur, and these undermine the rigor and credibility of student and faculty research. From the standpoint of academic administration, the situation is considerably worse when institutions treat AI-detection or similarity percentages as the criterion for penalizing authors or simply rejecting their work. AI detectors such as Turnitin are used more and more in university administration to ensure that academic documents are authentic. These tools nonetheless have significant limitations, above all in recognizing content written by people. According to Alnaimat et al. (2025), various editorial teams have for some time relied on AI-based, in-house ethics checks to prepare the initial documentation, expand the pool of evaluators, and, above all, detect potential plagiarism long before manuscripts reach peer review, so that human readers can then examine them.

In two experimental studies with groups of teachers, both highly experienced and novice, none of the participants could reliably detect texts copied from AI (Fleckenstein et al., 2024). These authors show that, given only very brief instructions, AI generates content that is hard to identify, which casts doubt on assessment in any educational setting.

Turnitin (2023c) states that its AI writing assessment is meant to help instructors spot text that may have been produced by a generative AI tool and should not serve as the sole basis for adverse decisions about a student. Deeper analysis and human judgment, together with an institution's own academic regulations, are essential for determining whether misconduct has actually occurred.

When choosing a detector, it is worth recognizing that using several at once can lead to more careful and better-grounded decisions, especially in higher education, where the language is often technical and can trigger false positives (Hyatt et al., 2025). False positives generate conflict in academic settings and call into question the effectiveness of institutional policies that depend solely on these systems. Evaluation in scholarly contexts also raises the dilemma of how to measure originality in academic work, given that students and professors alike have access to AI tools capable of generating complete texts in a matter of seconds.

Annie Chechitelli, Chief Product Officer at Turnitin, reports that false positives occur more often in documents in which less than 20% of the content is identified as AI-generated (Turnitin, 2023a).

In Mexico, the case of institutions such as the Autonomous University of Guerrero (UAGro) cannot be overlooked. Its Faculty of Law, specifically, has adopted Turnitin as a resource for identifying potential plagiarism and, more recently, for recognizing AI-generated content. The system compares the papers submitted by students and professors against an extensive database of academic documents, websites, and articles.

Pellegrina and Helmy (2025) offer, for the field of scientific integrity, an overview of what AI-generated text detectors can and cannot do. They identify weaknesses in performance, sensitivity to different types of text, and susceptibility to rewriting attacks or adversarial conditions. This comes amid recent editorial efforts to identify “paper mills” and collaborative platforms, as well as services designed specifically for image manipulation.

Since 2023, the tool has included an indicator showing the percentage of a text that was “possibly” created by artificial intelligence. According to the company itself, the detector recognizes more than 98% of the texts produced by ChatGPT, with a false positive rate below 1%. Turnitin nevertheless states that its reports should be read as signals rather than conclusive evidence (Turnitin, 2024).

According to Navarrete-Cazales and Cruz-Victoria (2026), bringing AI into education puts the tension between progress and dishonesty in plain view. On one side, using AI can be seen as comparable to using other permitted tools (online tutors, proofreading software, advanced calculators); on the other, if AI carries out substantial parts of the intellectual work, it could be considered a form of cheating. Schneider et al. (2025) examined 100 neurosurgery texts published in technical journals in 2023 and 2024 and found that one in five contained sections identified as AI-generated. Abstracts and methods sections were the parts most often affected.

The situation has exposed some gray areas. Is it wrong, for instance, for a student to turn to ChatGPT for extra clarification of a difficult topic? Almost certainly not; it could even be considered a helpful learning aid. And what about using it to produce a first draft of an essay that the student then revises and strengthens with citations? Some would see this as legitimate collaboration or assistance, while others would object that the work does not come entirely from the student. Using AI does not automatically amount to cheating or copying. Cases like these show that the judgment depends, above all, on the purpose and the method applied.

Murdock and Tadinada (2025) explain that, despite growing confidence in the technology, recent tests indicate that AI detection tools, although moderately to highly accurate at telling AI-generated text from human writing, remain unreliable and frequently produce false positives that harm students and educators alike. In scientific publishing, some of the detectors regarded as more accurate display a stronger bias against particular groups and in specific disciplines, which weighs heavily on decision-making in educational and publishing settings (Pratama,

2025).

Khan et al. (2025) point out that paraphrasing and "humanizing" tools make it harder to estimate precisely how much of a text is similar or AI-generated, as does the size of the dataset behind the text in question. Other scholars, such as Hua and Yao (2024), assert that some models built specifically for identification and attribution perform remarkably well on controlled datasets, but that rewriting and personalization tools substantially reduce detection effectiveness.

Weber et al. (2023) report that these instruments have clear limitations with certain kinds of text: human-written work translated by machine, and texts produced with AI that people later edit or paraphrase (hybrid texts).

What implications do Turnitin's false positives hold for university management?

Given this, universities have an obligation to put appeal protocols in place to protect their communities. Human review must always be part of the process, and it must be clear that no detector can serve as the only proof of academic dishonesty. Collegial review and appeal mechanisms are needed, so that students and professors alike can submit additional evidence. Professors, for their part, need to learn where plagiarism-detection software stops being useful; if they trust these tools alone, the result is imprecise. Other forms of assessment, such as interviews, examination of documents, or detailed rubrics, are a better bet. With that change, students and scholars would work in a climate of greater trust, free of the strain of living under an algorithm, and both the integrity of research and well-being within the university would be safeguarded.

In 2026, several studies carried out at Mexican universities showed that nearly one in four essays written by people was wrongly labeled as the work of an AI. Even Turnitin, one of the most reliable tools available for this task, produced 11% false positives on formal documents. Chávez-Martínez (2026) argues that, from an ethical standpoint, AI cannot be regarded as the author of an academic work, essentially because it lacks moral responsibility, cognitive intent, and the capacity for critical judgment that human beings have. Scientific authorship ultimately goes hand in hand with truth, with the ability to interpret, and with accountability for what one writes, qualities that only humans possess.

A study published in Education and Information Technologies (Springer, 2026) makes clear that Turnitin still has trouble distinguishing properly between what a person writes and what an AI generates. Even when a text is almost entirely human-authored, the system sometimes flags passages as artificial. This yields a fair number of false positives and leads many to ask whether the detector can really be trusted in a university setting.

At several universities in Mexico, nearly one in four essays written by people was mistakenly flagged as AI. Turnitin is somewhat more precise, but it still reaches an 11% false positive rate on academic work. Together, these results show how hard it is to detect AI text in Spanish, especially when the language is legal or technical (Trejo, 2026).

Studies from the United States and Europe have likewise reported false-

positive rates. Research in English-speaking contexts indicates that Turnitin had error rates between 8% and 12% on university papers. At Chinese universities, detectors flagged up to 20% false positives in academic texts that had been translated into English. Certain disciplines, law among them, are affected more severely. The language used there is technical, rigid, and bound by highly fixed conventions, and repeated references to statutes, codes, and articles of the Constitution are almost unavoidable.

Legal reasoning almost always rests on the same foundation: premise, argument, conclusion. For that reason, students new to the program write in such a heavy, formulaic way that their papers stand out to any automated filter. They are still learning to fit the pieces into a mold that is too narrow.

At the Autonomous University of Guerrero, law students regularly submit their papers to Turnitin. University professors and researchers send articles and book chapters to various journals and publishers for publication, and they also submit random samples of this work to Turnitin. When the tool rates such work as partly AI-written, some educators use those percentages to punish students, reject papers, or call faculty members' reputations into question.

This study seeks to document what is happening in the Faculty of Law of the Autonomous University of Guerrero. Among other questions, it asks the following: How many first-year students have been wrongly flagged as AI users? What characteristics do the texts that produce false positives share? How are instructors and academic units using Turnitin results?

The general objective of this study is to analyze false positives in AI detection with Turnitin among students and faculty in the Law program at the Autonomous University of Guerrero. The specific objectives are to (1) identify the prevalence of false positives among the 58 first-year students; (2) document cases of faculty researchers whose work has been wrongly flagged; (3) analyze the characteristics of the texts that produce false positives; and (4) evaluate how instructors and academic units are using Turnitin results.

Methods and materials

This study used a mixed-methods approach, combining quantitative and qualitative methods to obtain a complete picture of the problem. The design was descriptive: it describes the prevalence of false positives in AI detection with Turnitin and details how the phenomenon relates to specific characteristics of the texts. The mixed approach makes it possible, on the one hand, to quantify the frequency of false positives through analysis of Turnitin similarity reports and, on the other, to understand the experiences, perceptions, and practices of students and faculty through surveys and interviews.

The study took place at the Faculty of Law of UAGro in Chilpancingo, Guerrero, Mexico. UAGro is a public higher education institution that serves students in a range of disciplines, including law, medicine, engineering, the social sciences, and the humanities. The Faculty of Law offers the Bachelor of Laws, a nine-semester

program that prepares professionals to practice in various areas of the legal field. First-year students (first semester) take introductory courses such as Introduction to Law, Roman Law, and History of Law, along with general education courses.

As for Turnitin, UAGro holds an institutional license for the platform. Instructors can upload student work to generate similarity reports, and some academic units require this report as part of the evaluation of theses, articles, and degree projects. No detailed public information is available, however, on the specific policies that govern the use of Turnitin for AI detection in the Faculty of Law.

The study population consisted of two groups:

Table 1

Study population

GROUP 1	POPULATION	SAMPLE	INCLUSION CRITERION
First-year students	58	58	Enrolled in the first year of the Law program during the 2024-2026 academic year and having submitted at least one paper evaluated with Turnitin.
GROUP 2			
Faculty researchers	25	15	Full-time or part-time faculty member in the Faculty of Law, actively engaged in research, who submitted at least one work (article, book chapter, conference paper) evaluated with Turnitin between 2024 and 2026.

Source: Authors' own elaboration.

All 58 first-year students were included because the population is small and accessible, which made it possible to obtain complete data without sampling. For faculty, purposive convenience sampling was used: those who were available and willing to take part were selected.

Data collection instruments

- Student survey

A structured survey with 12 closed-ended and 2 open-ended questions was designed. The closed-ended items included multiple-choice options, a Likert scale (from "strongly disagree" to "strongly agree"), and yes/no questions. The open-ended questions allowed students to describe their experiences with Turnitin in their own words.

Survey content:

- Demographic data (age, sex, municipality of origin)
- Prior experience with AI tools (ChatGPT, Gemini, Copilot, etc.)
- Number of papers submitted through Turnitin during the semester
- AI percentages reported by Turnitin for their work
- Actual authorship of the work (100% human, partial AI use, full AI use)
- Instructors' reactions to AI percentages
- Knowledge of Turnitin's limitations
- Experiences of being questioned or sanctioned because of AI percentages
- Knowledge of appeal processes
- Perceptions of the fairness of the current system
- Open-ended questions on specific experiences and suggestions for improvement

The survey was administered in person in the classroom, with the instructors' authorization and the students' informed consent. Average completion time was 15-20 minutes.

- Document analysis of Turnitin reports

Students and faculty were asked for copies of the Turnitin similarity reports for their recent work. In all, 127 reports were collected: 89 from students and 38 from faculty.

For each report, the following information was recorded:

- Type of work (essay, exam, interim project, thesis, article, etc.)
- Document length (word count)
- Similarity percentage
- AI percentage
- Submission date
- Course or context in which the work was produced
- Grade received (if available)

- Instructor comments (if available)

To verify actual authorship, participants were also asked for evidence of their writing process: earlier drafts, edit history in Google Docs or Word, handwritten notes, or any other document showing how the work was developed. This evidence was crucial for identifying false-positive cases with certainty.

A documentary review was also conducted of the policies, regulations, and guidelines of UAGro and its Faculty of Law. The search covered the following topics:

- Academic integrity
- Use of plagiarism detection tools
- Evaluation of student work
- Sanction procedures for academic misconduct
- Use of artificial intelligence in academic work.

The following documents were consulted:

- General Regulations of the Autonomous University of Guerrero
- Academic Integrity Regulations (if any exist)
- Faculty of Law guidelines for the evaluation of student work
- Policies on the use of Turnitin (if any exist)
- Official statements on the use of AI (if any exist)

Comparable documents from other Mexican universities were also consulted to compare and contextualize the findings.

- Semi-structured interviews with faculty

Semi-structured interviews were conducted with 15 faculty researchers from the Faculty of Law. Each interview lasted between 30 and 45 minutes and was recorded with the participants' consent for later transcription and analysis.

Interview topics:

- Personal experience with Turnitin (as instructor and as author)
- Criteria used to assess the originality of student work
- How they use the Turnitin similarity and AI percentages
- Known cases of false positives or false negatives
- Knowledge of Turnitin's limitations
- Their academic unit's policies or guidelines on the use of Turnitin
- Sanction or appeal procedures when improper AI use is suspected
- Perceptions of the fairness and effectiveness of the current system
- Suggestions for improvement

Interviews took place in person or virtually, depending on each instructor's preference. All transcripts were anonymized to protect the participants' identity.

Procedure

Fieldwork was carried out between January and June 2026 and proceeded in the following stages:

Stage 1: Preparation (January 2026)

- Design of the instruments (survey, interview guide, document analysis form)
- Validation of the instruments by experts in research methodology
- Request for institutional permissions from the Faculty of Law
- Preparation of informed consent forms

Stage 2: Survey administration (February-March 2026)

- Visits to first-year classrooms in the Law program
- Explanation of the purpose of the study to students
- In-person administration of the survey
- Collection of signed informed consent forms

Stage 3: Collection of Turnitin reports (March-April 2026)

- Request to students and faculty for copies of their Turnitin reports
- Verification of authorship through review of drafts and process evidence
- Data entry into an Excel database

Stage 4: Faculty interviews (April-May 2026)

- Contacting faculty and scheduling interviews
- Conducting the interviews (in person or virtual)
- Transcription and anonymization of the interviews

Stage 5: Document review (May 2026)

- Search and analysis of institutional policies
- Comparison with the policies of other Mexican universities

Stage 6: Data analysis (June 2026)

- Descriptive statistical analysis of the quantitative data
- Content analysis of the qualitative data (interviews and open-ended questions)
- Triangulation of quantitative and qualitative data

Data analysis

Quantitative data: Survey and Turnitin report data were analyzed with descriptive statistics. Frequencies, percentages, means, and standard deviations were calculated for the main variables. Statistical analysis was run in SPSS version 28.

The main variables analyzed were:

- AI percentage reported by Turnitin
- Verified actual authorship (human/AI/hybrid)
- Type of work (essay, exam, project, thesis, article)
- Document length
- Grade received
- Instructor reaction (questioning, sanction, no action)

Qualitative data: The interviews and the open-ended survey questions were analyzed through thematic content analysis. Recurring themes, patterns, and emerging categories were identified, and Atlas.ti version 9 was used to code and organize the qualitative data.

The main themes identified were:

- Experiences with false positives
- Perceptions of the fairness of the system
- Instructors' evaluation criteria
- Suggestions for improvement

Data were triangulated by comparing quantitative with qualitative results and student perspectives with those of faculty. This gave a fuller picture and made it possible to validate the findings against multiple sources. The study followed the basic ethical principles for research with human subjects: informed consent, confidentiality, appropriate data use, institutional approval, and non-maleficence.

Results and discussion

Sample characteristics

Students:

Of the 58 first-year students, 34 were women (58.6%) and 24 were men (41.4%). The mean age was 19.2 years (range: 17-24). Most came from Chilpancingo and nearby municipalities (75.9%); the rest came from other regions of Guerrero (24.1%).

As for access to technology, 52 students (89.7%) reported having a personal computer at home, and 56 (96.6%) reported having internet access. Only 23 students (39.7%), however, reported having used AI tools such as ChatGPT before entering the university.

Faculty:

Of the 15 faculty members interviewed, 9 were men (60%) and 6 were women (40%). Mean teaching experience was 12.4 years (range: 3-28). All had at least one publication in a journal or book, and 11 of them (73.3%) held a doctorate.

Prevalence of false positives

Students: of the 89 Turnitin reports analyzed for students, 67 corresponded to work verified as 100% human-written (through review of drafts, edit histories, and interviews). Of these 67 fully human-written submissions, 12 (17.9%) received an AI percentage above 20% in Turnitin, which this study classifies as a false positive.

Distribution of AI percentages in 100% human-written student work:

- 0-5%: 32 submissions (47.8%)
- 6-10%: 15 submissions (22.4%)
- 11-20%: 8 submissions (11.9%)
- 21-30%: 6 submissions (9.0%)
- 31-40%: 4 submissions (6.0%)
- 41-50%: 2 submissions (3.0%)
- Over 50%: 0 submissions (0%)

The mean AI percentage in 100% human-written student work was 11.3% (standard deviation: 12.8%). The maximum was 47%, for an Introduction to Law essay that the student had written entirely on her own, without any AI. Students expressed frustration and distrust toward the system, along with fear of unjust sanctions.

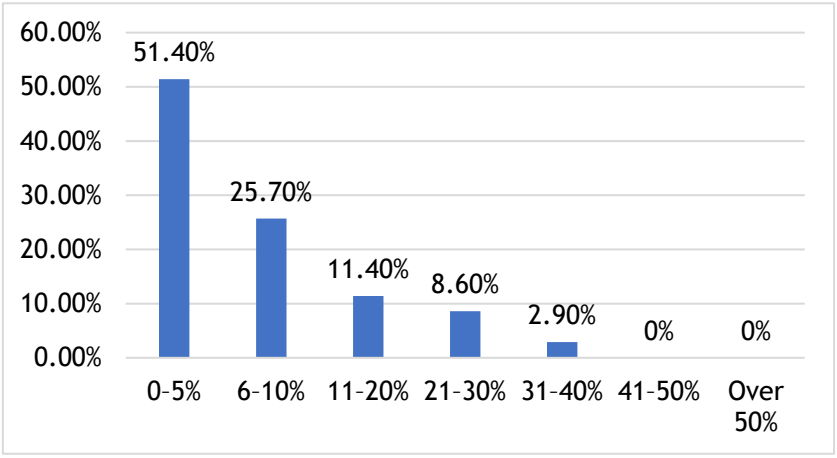
Faculty: of the 38 Turnitin reports analyzed for faculty, 35 corresponded to work verified as 100% human-written. Of these 35 submissions, 4 (11.4%) received an AI percentage above 20% in Turnitin.

Distribution of AI percentages in 100% human-written faculty work:

- 0-5%: 18 submissions (51.4%)
- 6-10%: 9 submissions (25.7%)
- 11-20%: 4 submissions (11.4%)
- 21-30%: 3 submissions (8.6%)
- 31-40%: 1 submission (2.9%)
- 41-50%: 0 submissions (0%)
- Over 50%: 0 submissions (0%)

Figure 1

Distribution of AI percentages in 100% human-written faculty work



Source: Authors' own elaboration.

The mean AI percentage in 100% human-written faculty work was 8.7% (standard deviation: 9.4%). The maximum was 36%, for an article submitted to an indexed journal of constitutional law.

Characteristics of the texts that generate false positives

Analysis of the submissions with false positives revealed several common patterns:

- Document length

Shorter submissions were more likely to produce false positives. The average word count was 1,420 for submissions with false positives, compared with 2,680 for those without.

Relationship between length and AI percentage:

- Under 1,000 words: average of 24.3% AI
- 1,000-2,000 words: average of 16.8% AI
- 2,000-3,000 words: average of 11.2% AI
- Over 3,000 words: average of 7.4% AI

This is consistent with Turnitin's own warning that short documents show a higher incidence of false positives.

- Type of work

Essays and midterm exams were more likely to produce false positives than projects or theses.

Percentage of submissions with false positives, by type:

- Essays: 22.4% (11 of 49 essays)
- Midterm exams: 19.2% (5 of 26 exams)
- Interim projects: 12.5% (3 of 24 projects)
- Theses/final projects: 8.3% (1 of 12 theses)
- Faculty articles: 11.4% (4 of 35 articles)

This suggests that shorter, less structured work (essays, exams) is more prone to false positives than longer, well-developed work (theses, projects).

- Linguistic characteristics

Submissions with false positives tended to share certain linguistic features:

Use of standardized formal phrases:

- "In conclusion…" (present in 83% of submissions with false positives vs. 45% of those without)
- "In accordance with…" (76% vs. 52%)
- "According to the theory of…" (71% vs. 38%)
- "Therefore…" (68% vs. 41%)
- "It is important to highlight that…" (64% vs. 35%)

- Standardized citations of legal norms:

Submissions that quoted constitutional articles, codes, or statutes verbatim were more likely to produce false positives. Quoting Article 1 of the Constitution ("In the United Mexican States, all persons shall enjoy the human rights recognized in this Constitution…"), for example, generates spurious matches because that text appears in millions of documents.

- **Rigid structure:**

Submissions with a highly predictable structure (introduction, body, conclusion, each with standardized phrasing) were more likely to produce false positives. This is particularly common among first-year students, who are only beginning to learn academic writing.

- **Language and style**

As expected, all submissions were in Spanish. Even so, those with heavier use of Latin legal terminology (habeas corpus, amparo, res judicata, etc.) or with citations in English had slightly higher AI percentages. Submissions by students whose first language was not Spanish (for example, students from Indigenous communities) also had higher AI percentages on average, although the difference was not statistically significant because of the small sample size.

Sajid et al. (2025), however, found no major differences between English-speaking and non-English-speaking AI users, except in email writing. Another important point in plagiarism detection is the use of paraphrasing and rewriting,

which draws on advances in semantic analysis and on combinations of methods covering textual and nontextual features.

The review of the reports showed that legal texts with abundant quotations of legal norms and repetitive structures were the most likely to be classified as suspicious. In several cases, Turnitin labeled as "AI-generated" paragraphs that consisted of transcriptions of constitutional articles or statutes, which exposes a technical limitation of the detector.

Instructors' reactions to student use of AI:

When a submission had an AI percentage above 20%, instructors' reactions varied:

Distribution of reactions (n=16 submissions with >20% AI):

- Verbal questioning of the student: 10 cases (62.5%)
- Request to redo the work: 6 cases (37.5%)
- Grade reduction: 4 cases (25.0%)
- Rejection of the work: 2 cases (12.5%)
- No action (accepted without questioning): 3 cases (18.8%)
- Review of authorship evidence: 5 cases (31.3%)

Notably, in 3 of the 16 cases (18.8%), the instructor took no action despite the high AI percentage. In interviews, these instructors said that they "know the student's writing style" or that "the content is too good to have come from AI," which suggests that they exercise professional judgment beyond the Turnitin percentage.

Faculty use of Turnitin

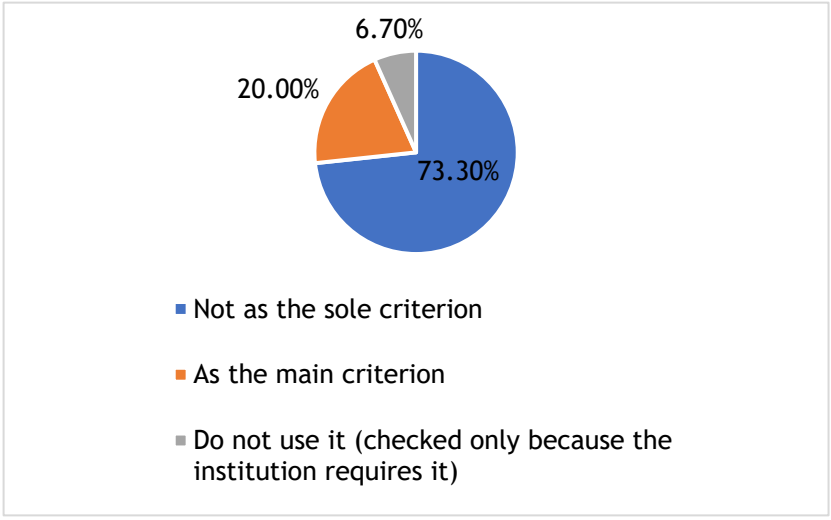
In the interviews, all 15 instructors described how they use Turnitin in their evaluations:

Use of the similarity percentage:

- I use it as a reference, but not as the sole criterion: 11 instructors (73.3%)
- I use it as the main evaluation criterion: 3 instructors (20.0%)
- I do not use it; I only check it because the institution requires it: 1 instructor (6.7%)

Figure 2

Use of Turnitin (similarity percentage) as an evaluation criterion by professors



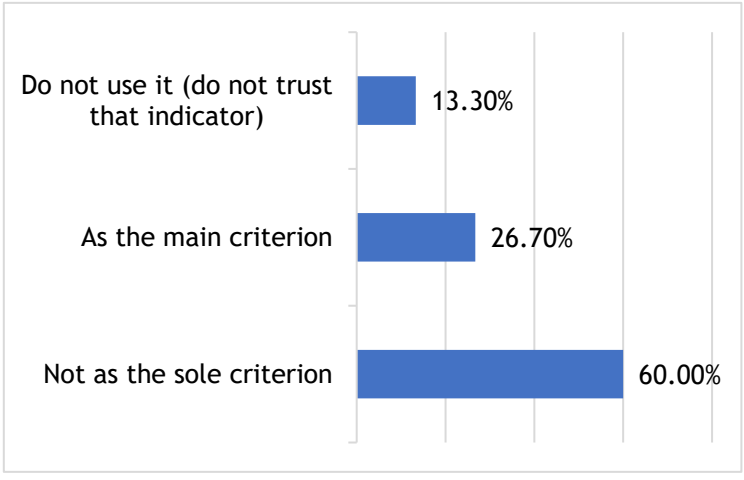
Source: Authors' own elaboration.

Use of the AI percentage:

- I use it as a reference, but not as the sole criterion: 9 instructors (60.0%)
- I use it as the main evaluation criterion: 4 instructors (26.7%)
- I do not use it; I do not trust that indicator: 2 instructors (13.3%)

Figure 3

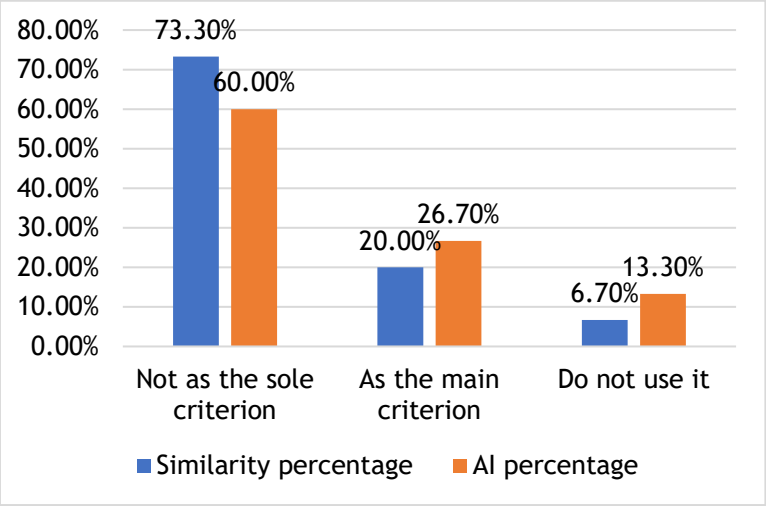
Use of Turnitin (AI percentage) as an evaluation criterion by professors



Source: Authors' own elaboration.

Figure 4

Comparison of Turnitin use as an evaluation criterion for similarity percentages and AI percentages



Source: Authors' own elaboration.

Thresholds considered "acceptable":

- Under 20% AI: 12 instructors (80.0%)
- Under 30% AI: 2 instructors (13.3%)
- Under 10% AI: 1 instructor (6.7%)

These data show that most instructors (73.3% for similarity, 60.0% for AI) use Turnitin as a reference rather than as the sole criterion. Even so, 4 instructors (26.7%) do use the AI percentage as their main criterion, which is problematic given the evidence of false positives. The professors noted that Turnitin reports are at times used as the only basis for rejecting work, which creates ethical and academic conflicts. Some acknowledged that they lack specific training in interpreting the results, which raises the likelihood of erroneous decisions.

Institutional policies

The documentary review revealed that:

- UAGro has no specific academic integrity regulation that explicitly mentions the use of AI.
- The documents consulted contained no institutional policy on acceptable similarity or AI thresholds in Turnitin.
- There is no formal appeal process for students who are challenged on the basis of Turnitin percentages.
- Some units have informal internal guidelines (for example, "less than 20% similarity is acceptable"), but these are neither documented nor uniform.

One instructor commented: “Every professor uses Turnitin however they like. There are no clear guidelines. Some are very strict, others more flexible. This creates a lot of inconsistency and, at times, injustice.” From the standpoint of university leadership, the core of the problem lies in the specific ways these tools are applied. Using such detectors can lead to unjust conclusions about plagiarism, with adverse effects on students’ academic outcomes and on the reputation of faculty authors.

Today it is undeniable that artificial intelligence is an integral part of society. Thanks to recent advances in generative AI, it has entered education at remarkable speed, although it is essential to weigh both the benefits AI could offer in higher education institutions and the dangers it may bring (Ortiz et al., 2025). The present findings agree with those of Merlo and Rinaldi (2025), who investigated the impact of Turnitin use on master’s-level work, looking at both similarity percentages and AI use. This study is consistent with theirs regarding the reduced use of such tools and the decline in non-original texts.

These findings are consistent with recent studies in other settings. Alonso (2024), for instance, observes that it is common to use artificial intelligence to search study databases and then paraphrase lightly without acknowledging the original authors. This suggests that gaps between an author’s grasp of a subject and the quality of the material presented point to a possible reliance on AI to appropriate other people’s ideas improperly.

In addition, studies conducted in the United States indicate that relying solely on these systems can lead to unfair penalties and to a loss of trust in institutions (Deep, 2026). Erol et al. (2025) report that teachers’ confidence in technological tools for analyzing students’ evaluative texts is growing. At the same time, the evidence on AI-detection tools has not yet reached full reliability, and false positives keep appearing that place students and professors at a disadvantage, which shows that depending exclusively on these automated instruments is not feasible.

This study agrees with Ruiz-Tirado et al. (2024), who argue that verifying the accuracy of data, together with promoting a robust ethics in how technologies are applied, becomes essential work for ensuring that the knowledge obtained is genuine and meaningful. The findings of this research confirm that AI detection with Turnitin produces false positives, a real and considerable problem in the Faculty of Law of the Autonomous University of Guerrero. Of the 67 submissions written entirely by students, 12 (17.9%) showed AI rates above 20%, which is defined as a false positive under the parameters of this study. Among faculty, 4 of 35 submissions (11.4%) produced false positives.

It should be noted that the figures in this study are higher than the 1% false-positive rate Turnitin reports (Turnitin, 2023a). Several causes may account for the gap. First, Turnitin states its false-positive rates at the document level, for documents in which the AI percentage exceeds 20%, whereas the present study covered all fully human-written submissions regardless of the percentage shown. Second, Turnitin was developed mostly with English-language data, and various studies have found error rates 15% to 20% higher in Spanish. Finally, the specific field

of law, with its technical terminology and standardized references to legal norms, could raise the likelihood of false positives.

The findings indicate that false positives in AI identification with Turnitin are not isolated cases but a frequent phenomenon that erodes confidence in academic evaluation processes. This agrees with international studies that have recorded similar limitations.

Fernández and Ruiz (2023), for example, found in Spain that documents with specific technical terminology showed false-positive proportions above 15%, in line with the results obtained in the Law program at the Autonomous University of Guerrero. Similarly, Soto (2024) reported in Chile that law and social science students were more likely to be flagged by AI systems, owing to strict regulatory language and the repetition of legal approaches.

In the English-speaking world, Deep (2026) points out that relying solely on these tools could weaken trust in institutions and lead to unjust punishments, which makes it urgent for universities to develop policies that combine technology with human oversight. Recent studies in educational ethics indicate that excessive dependence on automated tools can create an atmosphere of distrust and excessive surveillance, contrary to the principles of critical education and student autonomy (Pérez, 2021). In this context, university administrators should take a proactive stance in regulating these technologies, implementing specific protocols that prevent arbitrary decisions.

The results are also in line with Montesinos and Guaña-Moya (2026), who recently reported that, in documents combining human effort and automation, content creation together with human intervention during drafting lowers the detectability of a text's origin compared with purely AI-generated settings. These findings likewise coincide with recent studies published in Scopus-indexed outlets that warn of the dangers of relying solely on AI detectors.

Tsigaris and Teixeira (2026), in an article published in ScienceDirect, emphasize that most of the results produced by AI detectors are erroneous and that these systems tend to penalize unfairly students whose writing has the structured, repetitive forms typical of academic prose.

The results of Aquino et al. (2026) diverge from these findings. Those authors assessed how far certain systems for detecting AI-generated text, along with similarity percentages, can be trusted in educational institutions in Ecuador, and their results show high rates of false negatives and almost no real possibility of disciplining those who engage in the practice. The study draws attention to the distrust such situations breed and to the real prospect of a complete return to pencil-and-paper writing and, above all, to oral defenses.

Angelier (2026), in AI and Ethics (Springer), stresses that AI detection systems are biased and produce identification errors that compromise fairness in assessment, and suggests the need for a framework that integrates technology with human review. The present study offers evidence from Mexico and Latin America, where the use of legal language, the standardization of citations of legal norms, legal practices, and the rigidity of academic structure raise the likelihood of errors. This indicates that false positives are not merely a technical challenge but also a

cultural and disciplinary matter, which calls for university management strategies adapted to context.

The discussion indicates that the issue is not limited to technology; it also concerns university administration, namely how institutions understand and use the results these tools provide. Yanangómez et al. (2026) recently carried out a study to evaluate the reliability of several AI detectors applied to academic work in educational settings. They conclude that some AI detectors produce false positives and have significant limitations in accuracy, and therefore should not be used as definitive evidence in academic evaluation processes. These authors suggest that the tools be complemented with human review, contextual analysis, and pedagogical criteria before any judgment is formed about the authenticity of a student's work.

The present study is aligned with Chechitelli's recommendations for dealing with false positives (Turnitin, 2023b):

- Being informed before making decisions: it is crucial to weigh the likelihood of a false positive in advance and to settle on an approach for how to proceed when evaluating a result.
- Adopting a positive mindset: in a field so full of novelty and unfamiliar ground, educators should extend to their students the presumption of good faith.
- Being candid and transparent: it is essential to acknowledge in advance that false positives can occur, so educators and students alike need to be ready to keep communication clear and honest.

Conclusions

The texts that generate false positives in Turnitin in the Law program at the Autonomous University of Guerrero share common characteristics. Among them are document length (fewer than 1,000 words), the essay as a type of work, and the use of standardized formal phrases and citations of legal norms. Submissions with a highly predictable structure (introduction, body, conclusion) and texts that make heavier use of Latin legal terminology or English-language citations also stand out. In this sense, the data gathered in this study support the position that AI detectors should not be used as the sole evidence of academic misconduct, and a comprehensive evaluation framework combining human review, appeal processes, and alternative assessment is proposed.

References

Alnaimat, F., Feras AlSamhori, A. R., Hamdan, O., Seiil, B., & Qumar, A. B. (2025). Perspectives of Artificial Intelligence Use for In-House Ethics Checks of Journal Submissions. *Journal of Korean Medical Science*, 40(21), 1–11. https://doi.org/10.3346/JKMS.2025.40.E170

Alonso, R. A. (2024). Hacia un marco ético de la inteligencia artificial en la educación. Teoría de la Educación. *Revista Interuniversitaria, 36*(2), 79-98. https://revistas.usal.es/tres/index.php/1130-3743/article/view/31821

Angelier, V. (2026). *The pitfalls of AI detection in academic writing: Bias, false positives, and the need for inclusive assessment*. Aland Ethics. Springer. https://doi.org/10.1007/s43681-026-01376-w

Aquino, R. M., Estrella, N. P., Iza, G. P., & Pérez, M. I. (2026). Evaluación de sistemas de detección de IA y transformación del aprendizaje en la educación secundaria de Cotopaxi. (2026). *Revista Científica Y Tecnológica VICTEC*, 7(13), 237-247. http://server.istvicenteleon.edu.ec/victec/index.php/revista/article/view/251

Arteaga, J. E. (2023). Perspectivas de los docentes sobre la IA generativa: implicaciones para el aprendizaje personalizado en América Latina. *Ethos Scientific Journal, 1*(1), 56–69. https://doi.org/10.63380/ESJ.V1N1.2023.22

Cándano Pérez, M., Del Solar-Duarte, N., Remeseiro Reguero, R., Bracho-Fuenmayor, P. L., Baeza Ugarte, C. G., & Guzmán Sánchez, M. (2025). Aprendizaje basado en el uso de herramientas de IA: experiencia de realidad aumentada en la asignatura de "Nociones de Economía" con estudiantes de derecho de la Universidad Tecnológica Metropolitana del Estado de Chile. *Revista De Educación Y Derecho*, (2-Extraordinario), 331–353. https://doi.org/10.1344/REYD2024.2-Extraordinario.49196

Chávez-Martínez, O. (2026). ¿Cómo detectar textos científicos generados con inteligencia artificial? *Revista Médica del Instituto Mexicano del Seguro Social, 64*(1), e6952. https://doi.org/10.5281/zenodo.17477984

Deep, P. D. (2026). Artificial intelligence and academic integrity in higher education: Student perceptions, institutional responses, and the limits of AI detection. *International Journal of Studies in Education and Science (IJSES), 7*(3), 397-420. https://doi.org/10.46328/ijses.7882

Erol, G., Ergen, A., Gülşen Erol, B., Kaya Ergen, Ş., Bora, T. S., Çölgeçen, A. D., Araz, B., Şahin, C., Bostancı, G., Kılıç, İ., Macit, Z. B., Sevgi, U. T., & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica, 167*(214), 1–12. https://doi.org/10.1007/S00701-025-066224/FIGURES/

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

González Torres, V. H., Bracho-Fuenmayor, P. L., Baldevenites, E. V. L., Carrillo Guerrero, M. V., & Santander Erazo, R. D. (2024). Immersive learning in the metaverse: a review of evidence on pedagogical effectiveness and implementation gaps in Higher Education. *Metaverse Basic and Applied Research*, 3. 97. https://doi.org/10.56294/mr2024.97

Hua, H., & Yao, C. J. (2024). Investigating generative AI models and detection techniques: impacts of tokenization and dataset size on identification of AI-generated text. *Frontiers in Artificial Intelligence, 7*(1469197). https://doi.org/10.3389/FRAI.2024.1469197/BIBTEX

Hyatt, J. P., Bienenstock, E. J., Firetto, C. M., Woods, E. R., & Comus, R. C. (2025). Using aggregated AI detector outcomes to eliminate false positives in STEMstudent writing. *Advances in Physiology Education*, 49(2), 486–495. https://doi.org/10.1152/ADVAN.00235.2024

Khan, H. U., Naz, A., Alarfaj, F. K., & Almusallam, N. (2025). Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques. *Scientific Reports*, 15(20366), 1–26. https://doi.org/10.1038/S41598-025-082087;SUBJMETA

Merlo, C. E., & Rinaldi, S. M. (2025). El uso de Turnitin para la evaluación de la originalidad académica: un balance del primer año de implementación en una universidad argentina. *Revista Científica RES NON VERBA*, 15(1), 49–68. https://doi.org/10.21855/resnonverba.v15i1.999

Montesinos, L., & Guaña-Moya, J. (2026). Detección de plagio y autoría en la era de los LLM: precisión y efectos pedagógicos. *Ethos Scientific Journal*, 4(1), 284-309. https://doi.org/10.63380/esj.v4n1.2026.352

Navarrete-Cazales, Z., & Cruz-Victoria, J. C. (2026). Autoría e integridad académica ante la inteligencia artificial generativa en educación superior. *Revista Educación Superior y Sociedad (ESS)*, 37(2), 260-283. https://doi.org/10.54674/ess.v37i2.1174

Ortiz, J., Quintana Álvarez, J. H., Granda Dihigo, A., & Gómez León, Y. (2025). La Inteligencia Artificial en Educación Superior: análisis de riesgos y oportunidades. *Revista Cubana De Educación Superior, 44*(especial 2), 74-84. https://revistas.uh.cu/rces/article/view/12335

Pellegrina, D., & Helmy, M. (2025). AI for scientific integrity: detecting ethical breaches, errors, and misconduct in manuscripts. *Frontiers in Artificial Intelligence*, 8(1644098). https://doi.org/10.3389/FRAI.2025.1644098/BIBTEX

Pratama, A. R. (2025). The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science*, 11(e2953). https://doi.org/10.7717/PEERJ-CS.2953/SUPP-3

Ruiz-Tirado, S. G., Morán-Ortega, S. A., Simental-López, L. M., & Tirado-López, A. B. (2024). Barreras de la Inteligencia Artificial generativa en estudiantes de educación superior. Percepción docente. *RITI Journal, 12*(25). 26-37 https://doi.org/10.36825/RITI.12.25.003

Sajid, M., Sanaullah, M., Fuzail, M., Malik, T. S., & Shuhidan, S. M. (2025). Comparative analysis of text-based plagiarism detection techniques. *Plos One*, 20(e0319551). https://doi.org/10. 1371/JOURNAL.PONE.0319551

Schneider, D. M., Mishra, A., Gluski, J., Shah, H., Ward, M., Brown, E. D., Sciubba,

D. M., & Lo, S.-F. L. (2025). Prevalence of Artificial Intelligence-Generated Text in Neurosurgical Publications: Implications for Academic Integrity and Ethical Authorship. *Cureus*, 17(2). https://doi.org/10.7759/CUREUS.79086

Tiglla, B. (2025). Impacto de la Inteligencia Artificial Generativa (IAG) en la personalización del aprendizaje en universidades latinoamericanas. *Alpha International Journal*, 3(1), 18–30. https://doi.org/10.63380/AIJ.V3N1.2025.55

Trejo, D. (2026, julio 29). *Falsos positivos de IA en examen UNAM abren vacío jurídico*. Global Media. https://www.globalmedia.mx/articles/falsos_positivos_de_ia_en_examen_unam_abren_vacio_juridico

Tsigaris, P., & Teixeira, J. A. (2026). AI detecting AI in academic writing: Why most AI detector findings are false. *Computers in Human Behavior Reports, 14*, 100354. https://doi.org/10.1016/j.chbr.2026.100354

Turnitin. (2023a). *Actualización sobre la detección de escritura con IA por la Directora de Producto Turnitin*. https://latam.turnitin.com/blog/actualizacion-sobre-deteccion-de-escritura-con-ia-por-la-directora-de-producto-turnitin

Turnitin. (2023b) *¿Qué significan los falsos positivos en nuestra función de detección de escritura con IA?* https://www.turnitin.com.mx/blog/que-significan-los-falsos-positivos-en-funcion-de-deteccion-de-escritura-con-ia

Turnitin. (2023c). *Detección de escritura de IA en la vista de reporte clásica*. https://guides.turnitin.com/hc/es/articles/28457596598925-Detecci%C3%B3n-de-escritura-de-IA-en-la-vista-de-reporte-cl%C3%A1sica-espa%C3%B1ol

Turnitin. (2024). *AI writing detection: Guidance for educators*. Turnitin Official Documentation. https://www.turnitin.com

Weber, D., Anohina, A., Bjelobaba, S., Foltýnek, T., Guerrero Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AIgenerated text. *International Journal for Educational Integrity, 19*(26), 1–39. https://doi.org/10.1007/s40979-023-00146-z

Yanangómez, P. A., Díaz, E. R., Ojeda. L. M., Yaguachi, C. B., & Carrillo, M. C. (2026) Falsos positivos en detectores de Inteligencia Artificial en trabajos académicos: análisis comparativo de plataformas digitales en contextos educativos. *ASCE MAGAZINE* 5(2) 1976-1999 https://doi.org/10.70577/asce.v5i2.861

About the main author

Perla Elisabeth Ventura Ramos: she received her education at the Autonomous University of Guerrero, Mexico, where she earned a degree in Systems Engineering, a Master's degree in Computer Science, and a Ph.D. in Pedagogy. She currently serves as a professor and researcher in Educational Technology at the university. She has worked as a full-time faculty member at the Autonomous University of Guerrero's Faculty of Law, where she teaches courses on the use of Information and Communication Technologies, Computer Law, and the intersection of Law, Gender, and Power, as well as Statistics Applied to Law.

Declaration of author responsibility

Perla Elisabeth Ventura Ramos 1: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation/Verification, Visualization, Writing – original draft, and Writing.
Norma Yadira Memije Alarcón 2: Data curation, Formal analysis, Investigation, Resources, Supervision, Validation/Verification, Visualization, Writing/original draft, and Writing, review, and editing.
Jesús Zaragoza Martínez 3: Data curation, Formal analysis, Investigation, Resources, Validation/Verification, Visualization, Writing/Original draft.

Special Acknowledgments:

Financing:

Own resources