Neutrosophic Sets and Systems, Vol. 100, 2026
University of New Mexico

The Percentage Fallacy and a Single-Valued Neutrosophic Framework for Indeterminacy in AI Detector Verdicts on Academic Authorship

Oscar Caicedo-Camposano[1]*, Mayra Jara-Sen San[1], César Caicedo-Camposano[2], Ariana Cruz-Posligua[2] and Maikel Leyva-Vázquez[3]

1 Universidad Técnica de Babahoyo, Ecuador
2 Universidad Regional Autónoma de Los Andes, Ecuador
3 Universidad Bolivariana del Ecuador, Ecuador
* Correspondence: ocamposano@utb.edu.ec

Abstract: Artificial intelligence (AI) content detectors have entered university academic assessment as authorship verification instruments, frequently without formal regulation or consideration of their documented error rates. This article proposes a single-valued neutrosophic (SVN) framework to formalize the epistemological indeterminacy that characterizes the verdicts of these systems. We argue that every AI detector collapses an implicit neutrosophic triple ⟨ T, I, F⟩ into a one-dimensional scalar, eliminating the indeterminacy component I and producing what we call the percentage fallacy. To resolve this indeterminacy, we propose a neutrosophic score function $S = T + \alpha \cdot D - I - F$ that incorporates the demonstrated cognitive mastery D through oral defense or direct evaluation, and we formalize the Principle of Cognitive Primacy (PCP) as the prevailing criterion of academic authorship. The systematic literature review documents false positive rates exceeding 60% for academic writing produced by non-native English speakers, including Ecuadorian and Latin American students. We conclude that, when an institution has a regulation limiting AI use in academic documents and applies a detector as a verification tool, the cognitive mastery demonstrated by the student over that document must operate as the primary criterion to resolve the algorithmic ambiguity, since it is the only evidence that captures the indeterminacy component I that the detector's scalar suppresses.

Keywords: single-valued neutrosophic sets; AI detectors; false positives; academic integrity; Principle of Cognitive Primacy; algorithmic bias; academic authorship; higher education.

1. Introduction

The emergence of large language models (LLMs) in the university ecosystem has destabilized established assumptions about academic intellectual production. Higher education institutions have responded with uneven speed, deploying commercial AI content detectors as control instruments without regulatory frameworks that establish their validity as evidence of infraction in most cases [4, 7]. Tools such as GPTZero, Turnitin AI, Copyleaks, and OpenAI Classifier have been incorporated into faculty workflows under an implicit premise that this article subjects to formal scrutiny: that a numerical percentage of artificial authorship probability constitutes sufficient evidence to determine whether a student used AI in producing an academic work. We demonstrate that this premise is mathematically incorrect.

The central problem we address is not whether students use AI or not, but whether a scalar produced by an algorithm constitutes epistemologically valid evidence to determine authorship. We argue that the answer is negative, and that this error can be formalized with precision through the theory of single-valued neutrosophic sets (SVN) [8, 9], since academic authorship under algorithmic uncertainty is irreducibly a problem of indeterminacy. Neutrosophic theory is the only mathematical framework that treats indeterminacy as an independent component, not derived from the truth and falsity components [9, 10].

In the Ecuadorian context, the Organic Law of Higher Education (LOES) [19] and the Academic Regulations of the Higher Education Council (CES) [22] contain no specific provision on the use of AI detectors as authorship verification instruments. This absence does not enable any evaluative practice: when institutions establish rules limiting AI use in producing academic documents, the verification process activated must guarantee that the

evidence employed is epistemologically valid. The principle of due process requires that any disciplinary action be supported by sufficient evidence, not by statistical probabilities not validated for the specific application context.

This article makes five contributions. First: we reformulate the output of an AI detector as a single-valued neutrosophic triple ⟨ T, I, F⟩ . Second: we define a neutrosophic score function S(work) = T + α·D − I − F that incorporates demonstrated cognitive mastery D. Third: we formulate the Principle of Cognitive Primacy (PCP). Fourth: we document the paradox of historical false positives. Fifth: we analyze the implications of the PCP for institutional contexts with and without specific regulation on AI use.

2. Neutrosophic Preliminaries

We present the formal definitions necessary for the development of the proposed framework. Neutrosophic theory was introduced by Smarandache [9] as an extension of Zadeh's fuzzy logic [12] and of Atanassov's intuitionistic sets [11], with the fundamental difference that the three components of a neutrosophic element are mathematically independent of each other.

Definition 1 (Neutrosophic set [9, 10]). Let X be a universe of discourse. A neutrosophic set A on X is defined as:

$$A = \{\langle x, T\_A(x), I\_A(x), F\_A(x)\rangle : x \in X\}$$

where T_A(x), I_A(x), F_A(x) : X → $]^{-}0, 1^{+}[$ are the truth, indeterminacy, and falsity membership functions, respectively, with:

$$^{-}0 \leq T\_A(x) + I\_A(x) + F\_A(x) \leq 3^{+}$$

The independence between T, I and F distinguishes this structure from classical fuzzy logic, where the condition T + F ≤ 1 is mandatory, and from intuitionistic sets [11], where T + F + H = 1. In neutrosophic theory none of these restrictions applies, which makes it possible to represent genuinely indeterminate states of knowledge without forcing a distribution between truth and falsity.

Definition 2 (Single-valued neutrosophic set — SVN [8]). An SVN over X is a triple A = ⟨ T_A, I_A, F_A⟩ with T_A, I_A, F_A ∈ [0, 1] and:

$$0 \leq T\_A + I\_A + F\_A \leq 3$$

SVNs are the practical restriction of neutrosophic sets to single values in [0, 1], which makes them directly operational in decision-making problems with real uncertainty [15, 17, 21]. Each component is quantified independently, without the value of one determining or restricting the value of the others.

Definition 3 (Neutrosophic score function [15, 16]). Let A = ⟨ T, I, F⟩ be an SVN. The score function of A is defined as:

$$s(A) = (T - I - F) / 3, \quad s(A) \in [-1, 1]$$

This function penalizes indeterminacy and falsity and allows ordering of neutrosophic alternatives. Values of s close to 1 indicate high truth and low uncertainty; values close to −1 indicate high falsity or high indeterminacy.

Definition 4 (Extended neutrosophic score with cognitive mastery). Let W be an academic work assessed by an AI detector, with verdict represented as ⟨ T, I, F⟩ . Let D ∈ [0, 1] be the cognitive mastery demonstrated by the student over W, and α ∈ [0, 1] an institutional weighting parameter. The extended neutrosophic score of the work is defined as:

$$S(work) = T + a \cdot D - I - F$$

When α = 1, cognitive mastery receives the same weight as the truth component. When α = 0, cognitive mastery is discarded, which is equivalent to accepting the algorithmic verdict without human correction. We argue that α = 0 is epistemologically inadmissible in any academic assessment system that recognizes the existence of false positives.

3. State of the Art on False Positives in AI Detectors

The scientific literature published between 2022 and 2025 consistently documents the low reliability of AI content detectors, with particular concentration on the false positive phenomenon and its asymmetric distribution among writer populations.

Weber-Wulff et al. [2] evaluated fourteen AI text detection tools and concluded that none was sufficiently accurate or reliable for use in academic contexts with disciplinary consequences. Their analysis revealed systematic biases toward classifying human text as machine-generated, especially when the text presented regular syntactic structures or standardized technical vocabulary. Elkhatat, Elsaid and Almeer [3] evaluated five commercial detectors against texts produced by GPT-3.5 and GPT-4 and found significant inconsistencies in the classification of human controls: the instruments produced false positives and uncertain classifications even when the text was written by the researchers themselves.

The most relevant finding for the Latin American context comes from Liang et al. [1], who demonstrated through controlled experiment that AI detectors classify as machine-generated more than 60% of TOEFL essays written by non-native English speakers, while texts from native English-speaking students are correctly identified as human in nearly 100% of cases. This gap does not reflect actual AI use, but rather statistical regularity in the lexical and syntactic structure characteristic of formal academic writing in a second language. Detectors confuse formal competence with algorithmic artificiality.

Sadasivan et al. [6] formally demonstrated that perfect detection of LLM-generated text is, under general conditions, mathematically impossible. Their argument establishes that any detector can be defeated by simple paraphrasing, and that the error rate inevitably grows when the analyzed text comes from writers with consolidated formal styles. Mitchell et al. [14], with their DetectGPT system, achieved acceptable results only under laboratory conditions that are not reproduced in ordinary academic practice.

The historical paradox represents the most compelling argument against the epistemological validity of detectors. Khalil and Er [13] documented that tools such as GPTZero assign artificial-origin probabilities greater than zero to texts written before the existence of modern LLMs. This result demonstrates that detectors do not identify artificial authorship: they identify linguistic statistical regularity. An instrument that assigns artificial-origin probability to nineteenth-century academic prose cannot operate as valid evidence of authorship in the twenty-first century.

Giray [18] documents that false accusations derived from false positives have caused verifiable damage to the academic trajectory of students and researchers, constituting a form of algorithmic discrimination whose consequences are asymmetric depending on the language of writing and the level of stylistic formality. Foltýnek, Meuschke and Gipp [25] note that detection systems for non-original production require, as a minimum validity condition, error rates sufficiently low for the specific application context, a condition that current AI detectors do not satisfy. Kasneci et al. [20] argue that the integration of AI in education requires assessment frameworks that recognize the complexity of technologically mediated learning, rather than control instruments of unproven validity.

Cotton, Cotton and Shipway [7] argue that academic integrity in the era of LLMs requires a reconceptualization of verification mechanisms, with emphasis on assessment of process over textual product. Bozkurt [5] notes that human-AI co-authorship challenges conventional academic authorship frameworks and demands criteria that go beyond inspection of the final artifact. Perkins [4] documented that faculty perceive detectors as convenient but epistemologically fragile control instruments, and notes the need for alternative verification mechanisms based on direct demonstration of knowledge.

4. The Percentage Fallacy: Neutrosophic Formalization

We call the percentage fallacy the practice of taking the scalar output of an AI detector as sufficient evidence to determine artificial academic authorship. The formal argument we develop shows that this practice implies a mathematical simplification that eliminates precisely the most relevant information for a fair decision.

Let W be an academic work. An AI detector applies a function $f : W \to [0, 1]$ on W that produces a scalar p, interpreted as the probability that W is of artificial origin. From the neutrosophic perspective, the authorship of W under algorithmic uncertainty should be represented as a triple $\langle T, I, F \rangle$ where T is the degree of evidence of verifiable human authorship, I is the degree of indeterminacy inherent to the algorithmic verdict, and F is the degree of evidence of artificial authorship.

Proposition 1 (Epistemological insufficiency of the detection scalar). The scalar output p of an AI detector implies the assumption $T = 1 - F$, which corresponds to classical fuzzy logic and not to neutrosophic logic. This assumption eliminates by construction the indeterminacy I and produces an epistemologically incomplete representation of authorship.

Proof. If the detector produces $p = F$ as the probability of artificial origin, and presents $(1 - p)$ as the probability of human origin, then $T := 1 - F$. This identity collapses the independence between T and F and

implicitly establishes I = 0. But I = 0 would mean that no degree of indeterminacy exists in the verdict, which contradicts the empirical evidence documented [1, 2, 3, 6], which shows incorrect, contradictory, and systematically biased classifications. Therefore, the scalar representation of the detector does not capture the actual epistemic structure of the problem. □

The direct consequence of Proposition 1 is that no AI detector, by its fundamental statistical design, can produce sufficient evidence of artificial authorship. It can produce risk indicators that justify additional verification, but it does not constitute proof.

Proposition 2 (Primacy of the extended score). If I > 0 in the neutrosophic representation of the verdict of an AI detector, then S(work) = T + α·D − I − F is strictly greater than the assessment based exclusively on p = F when α > 0 and D > 0.

Proof. By hypothesis, I > 0. For any work with D > 0, the extended score satisfies:

$$S = T + a{\cdot}D - I - F \;>\; T - I - F \geq -F - I \;>\; -1$$

The difference between S and the scalar criterion −F is:

$$\Delta = S - (-F) = T + a{\cdot}D - I$$

For Δ > 0, T + α·D > I is required. Given that the documented values of I in AI detectors are significant [2, 3] and cognitive mastery D can approach 1 through direct oral defense, Δ > 0 is the usual condition for students who genuinely know their work. □

5. The Principle of Cognitive Primacy (PCP): Formal Framework

Definition 5 (Cognitive mastery). Let W be an academic work and s a student. The cognitive mastery D(s, W) ∈ [0, 1] of s over W is a measure that quantifies the ability of s to answer, explain, expand, and relate the content of W under direct assessment conditions, through a structured, contradictory oral defense protocol conducted by a human evaluator.

The measure D incorporates at least three verifiable dimensions: the ability to reformulate the central ideas of the work without consulting the text; the ability to answer in-depth questions on the topic addressed; and the ability to identify the primary sources cited and explain their relevance. A trained evaluator can assign D through an explicit rubric, which makes the PCP operationalizable in concrete institutional contexts.

Principle of Cognitive Primacy (PCP). Let W be an academic work on which an AI detector has produced a verdict ⟨ T_d, I_d, F_d⟩ with I_d > 0. Let D(s, W) be the cognitive mastery demonstrated by student s in defense. The PCP states:

$$If\ S(work) = T\_d + a{\cdot}D(s,W) - I\_d - F\_d \;>\; 0,$$

then W must be considered a work of verified human authorship, regardless of the value of F_d. Consequently, no verification process on AI use can be resolved exclusively with the scalar p = F_d.

The epistemological foundation of the PCP rests on the distinction between process evidence and product evidence. An AI detector examines only the final textual artifact. But academic authorship is an attribute of the intellectual process that generated that artifact, not of the artifact itself. A text may present statistical characteristics similar to those of AI-generated text because the student writes in a language that is not their mother tongue [1], because they have deeply internalized the patterns of formal academic discourse [18], or because their style is technically rigorous [2]. None of these causes implies AI use. Only direct assessment of the cognitive process resolves this indeterminacy.

The PCP is also consistent with the principle of due process: the burden of proof falls on the accuser, and the accused has the right to present exculpatory evidence. The positive demonstration of cognitive mastery is precisely that evidence, and its epistemological weight is superior to that of a statistical percentage whose Type I errors exceed 60% in populations comparable to the Ecuadorian one [1].

6. Illustrative Numerical Case

We present a hypothetical case that illustrates the application of the neutrosophic framework with concrete and internally verifiable numerical values.

Let s_1 and s_2 be two students who submit academic works W_1 and W_2 respectively. Both obtain the same percentage from the AI detector: $p = 0.65$ (65% probability of artificial origin). Under conventional scalar assessment, both students would receive an identical institutional response.

Step 1: Neutrosophic representation of the detector verdict.

Since the detector only produces the scalar $p = F$, and implicitly assumes $T = 1 - F$, the initial triples are identical. Correcting with $I > 0$ according to documented error rates [2, 3]:

$$W_1 : (T_1, I_1, F_1) = (0.25, 0.10, 0.65)$$

$$W_2 : (T_2, I_2, F_2) = (0.25, 0.10, 0.65)$$

Step 2: Oral defense and quantification of cognitive mastery.

$$D(s_1, W_1) = 0.90 \quad \textit{(high mastery, consistent and autonomous responses)}$$

$$D(s_2, W_2) = 0.15 \quad \textit{(low mastery, superficial and inconsistent responses)}$$

Step 3: Calculation of the extended neutrosophic score with $\alpha = 1$.

$$S(W_1) = 0.25 + 1 \times 0.90 - 0.10 - 0.65 = 0.40 > 0 \rightarrow \textit{verified authorship}$$

$$S(W_2) = 0.25 + 1 \times 0.15 - 0.10 - 0.65 = -0.35 < 0 \rightarrow \textit{requires additional investigation}$$

s_1 obtains a positive score that, by the PCP, indicates verified human authorship despite the detector percentage. s_2 obtains a negative score that generates a case warranting additional investigation. This numerical example demonstrates that two students with the same AI detection percentage may have radically different situations regarding their authorship, and that only the SVN framework with the variable D can formally capture this difference.

To illustrate the historical paradox, consider an academic text H written in 1994. If a contemporary detector is applied to H, the result would be $F_H > 0$ despite H being produced before the existence of any LLM. The neutrosophic representation requires $I_H > 0$, which immediately activates the PCP: the verdict is indeterminate and cannot be used as evidence of infraction.

7. Discussion

7.1. Reliability of AI Detectors and the Percentage Fallacy

The analysis formalized in Propositions 1 and 2 shows that every AI detector that produces a scalar as output structurally eliminates neutrosophic indeterminacy, which invalidates its use as sufficient evidence of artificial authorship. This is consistent with what was stated by Weber-Wulff et al. [2], who document that none of the fourteen detection tools evaluated in their study reached sufficient levels of accuracy or reliability for academic contexts with disciplinary consequences. According to these authors, detectors present systematic biases that make them inadequate as authorship verification instruments, which confirms that the scalar output of these systems cannot operate as conclusive evidence by itself.

Along the same lines, Elkhatat, Elsaid and Almeer [3] found that five commercial detectors produce inconsistent classifications even when the text was written entirely by the researchers themselves, generating false positives in the human controls. This coincides with Proposition 1: if detectors classify as artificial a text produced by academic evaluators, the indeterminacy component I is structurally greater than zero, which collapses any scalar interpretation of the result. In contrast to the practice implicit in some institutional guidelines that cite the Turnitin or GPTZero percentage as a direct verification criterion, the present analysis shows that this practice eliminates, by mathematical construction, the most relevant information for a fair decision.

Sadasivan et al. [6] establish, through a formal argument, that perfect detection of LLM-generated text is mathematically impossible under general conditions, since any detector can be defeated by simple paraphrasing and the error rate grows structurally when the text comes from writers with consolidated styles. This is consistent with Proposition 2: the extended score S surpasses the scalar criterion precisely in cases where the formal style of the student artificially raises the value of F. On the other hand, Mitchell et al. [14] argue that their DetectGPT system achieves acceptable performance under controlled laboratory conditions; however, the authors themselves acknowledge that these conditions are not reproduced in ordinary academic practice, so the actual discriminatory

capacity of the detector remains insufficient to ground decisions with consequences on the academic trajectory of the student.

The historical paradox documented by Khalil and Er [13], who observed that detectors such as GPTZero assign artificial-origin probabilities greater than zero to texts produced before the existence of modern LLMs, reinforces this argument from an empirical angle. According to these authors, the phenomenon evidences that detectors do not identify artificial authorship but linguistic statistical regularity, which converges with the central thesis of this article: the instrument detects frequency patterns, not cognitive processes, and this limitation is irreducible in the current design of these systems.

7.2. Linguistic Bias and Algorithmic Discrimination in the Latin American Context

The analysis reveals that AI detectors structurally penalize those who write in languages other than English or in formal academic styles. This concurs directly with the findings of Liang et al. [1], who documented that more than 60% of TOEFL essays written by non-native English speakers were classified as AI-generated, compared to correct human identification rates close to 100% for native English texts. According to these authors, the gap does not reflect actual AI use but lexical-syntactic statistical regularity proper to the formal academic register in a second language, which turns the detector into an instrument that discriminates on linguistic competence, not on authorship.

In the Ecuadorian context, where academic works are produced in Spanish and students build their writing within the conventions of Latin American academic discourse, the phenomenon described by Liang et al. [1] translates into false positive rates potentially higher than those documented for English, given that commercial detectors are trained predominantly with corpora in that language. This is consistent with what Giray [18] points out, who documents that researchers with formally rigorous writing styles recurrently obtain higher false detection rates. According to Giray [18], the false accusations derived from these errors have generated verifiable damages to the trajectory of students and academics, which reinforces the need for the PCP as a mechanism that allows correcting that bias through direct measurement of cognitive mastery.

In contrast to the reading that AI detectors are neutral teaching support tools — a reading underlying some institutional educational policy documents — the present analysis shows that algorithmic neutrality is an illusion when the instrument was designed and validated on a linguistic corpus that does not represent the population that receives it. Cotton, Cotton and Shipway [7] recognize that detectors can serve as an initial alert signal, but argue that academic integrity in the LLM era requires privileging assessment of the intellectual process over inspection of the textual artifact. This concurs with the PCP: demonstrated cognitive mastery D is precisely that process evidence which the authors point to as a superior criterion, and its formal incorporation into the function S grants it the operational weight that equitable assessment requires.

7.3. Cognitive Mastery as Verification Criterion under Institutional AI-Use Regulation

The central argument of this article does not aim to create sanctioning mechanisms for those who use AI, but rather to establish what must be verified when an institution already has regulations limiting or regulating that use and applies a detector as a verification tool. In that scenario, the problem is precise: if the detector issues an ambiguous verdict on a document presumably generated with AI, what evidence can resolve that ambiguity? This concurs with what Perkins [4] poses, who points out that the absence of clear criteria on what constitutes valid evidence of infraction transfers to faculty discretion a high-impact decision on the student's trajectory. According to this author, that discretion generates systematic inequities that disproportionately affect students with less institutional capacity to challenge.

Bozkurt [5] argues that the irruption of generative AI forces a reconsideration of concepts such as authorship and academic responsibility, and that institutions must adopt frameworks that recognize the collaborative nature of contemporary production. While the present analysis agrees with Bozkurt [5] that conventional authorship frameworks are insufficient for the current context, it differs in focus: whereas that author orients the discussion toward the redefinition of shared authorship norms, the proposed PCP does not debate whether AI use is legitimate or not, but establishes the operational criterion to determine when the student demonstrates that the knowledge contained in the document belongs to them, regardless of the tools used during preparation. The positive demonstration of high D(s,W) constitutes the evidence of intellectual authorship that no algorithmic percentage can provide.

Kasneci et al. [20] argue that the integration of AI in education requires frameworks that balance pedagogical opportunities with ethical safeguards, and warn that purely restrictive institutional responses can hinder the development of competencies that students will need in their professional life. This concurs with the PCP argument: verifying cognitive mastery does not discourage the reflective use of AI in the learning process, but rather shifts academic responsibility to the right place. When the student can sustain a solid defense on the content of their work — regardless of which tools they used during preparation — the assessment system fulfills its pedagogical purpose of verifying real appropriation of knowledge.

Zawacki-Richter et al. [27] document that AI applications in higher education have been adopted fundamentally from the technological perspective, with scant faculty participation in the design of validation criteria. This concurs with the underlying critique of this article: AI detectors entered university assessment processes without specific pedagogical validation for each use context, reproducing the pattern of technological primacy over educational criterion that those authors identify as a structural risk. The function $S = T + \alpha \cdot D - I - F$ reverses this logic by placing the parameter α — decided by the institution based on its pedagogical objectives — as a regulator of the relative weight between the algorithmic verdict and direct human evidence.

Foltýnek, Meuschke and Gipp [25] establish that detection systems for non-original production require, as a minimum condition of validity in academic contexts with disciplinary consequences, error rates sufficiently low for the specific population of use. This condition is equivalent, in neutrosophic terms, to requiring that I be significantly close to zero before accepting the verdict as evidence. Given that the studies of Liang et al. [1], Weber-Wulff et al. [2] and Elkhatat et al. [3] convergently demonstrate that this condition is not met for any tool available in the Latin American context, the conclusion is direct: when an institution has regulations limiting AI use and decides to verify the authorship of a questioned document, the PCP must operate as a mandatory criterion of resolution, since it converts the indeterminacy I of the detector into an empirically answerable question through direct human assessment.

8. Conclusions

This article has demonstrated through neutrosophic formalization that the output of every AI detector collapses an implicit neutrosophic triple $\langle T, I, F \rangle$ into a one-dimensional scalar, eliminating the indeterminacy I by design. This elimination is epistemologically incorrect because the empirical evidence consistently documents that $I > 0$ for any academic text evaluated with currently available tools.

The Principle of Cognitive Primacy that we propose formally resolves this indeterminacy by incorporating demonstrated cognitive mastery D as the central decisive variable through the extended score function $S = T + \alpha \cdot D - I - F$. This principle does not propose new sanctions for AI use: it establishes what must be verified when an institution already has regulations limiting that use and employs a detector as a verification tool. The answer is that the cognitive mastery demonstrated by the student over the questioned document — and not the algorithmic percentage — must operate as the primary criterion of resolution.

The systematic literature review confirms three facts: AI detectors present false positive rates exceeding 60% for the writing of non-native English speakers; these tools assign artificial-origin probabilities to historical texts produced before the existence of LLMs; and no available tool reaches sufficient accuracy levels for use in processes with disciplinary consequences.

In the Ecuadorian context, the absence of specific CES regulations on the use of AI detectors implies that any sanction grounded exclusively on algorithmic percentages lacks adequate normative basis. The PCP offers the criterion that institutions can adopt to guarantee due process: the verdict of a detector identifies a case that warrants verification, but it is direct oral defense that resolves the indeterminacy.

As future lines of work, we propose extending the framework to group assessments through SVN aggregation operators [15, 17], the development of standardized oral defense protocols with formal cognitive mastery metrics calibrated for Latin American university contexts, and the empirical validation of the parameter α through case studies in Ecuadorian universities.

References

[1] Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; Zou, J. GPT detectors are biased against non-native English writers. Patterns 2023, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

[2] Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Pokorný, O.; Šigut, P.; Waddington, L. Testing of detection tools for AI-generated text. International Journal for Educational Integrity 2023, 19(1), 26. https://doi.org/10.1007/s40979-023-00146-z

[3] Elkhatat, A. M.; Elsaid, K.; Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity 2023, 19(1), 17. https://doi.org/10.1007/s40979-023-00140-5

[4] Perkins, M. Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice 2023, 20(2), 7. https://doi.org/10.53761/1.20.02.07

[5] Bozkurt, A. GenAI et al.: Cocreation, Authorship, Ownership, Academic Ethics and Integrity in a Time of Generative AI. Open Praxis 2024, 16(1), 1–10. https://doi.org/10.55982/openpraxis.16.1.654

[6] Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-generated text be reliably detected? arXiv:2303.11156 2023. https://doi.org/10.48550/arXiv.2303.11156

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

[8] Wang, H.; Smarandache, F.; Zhang, Y. Q.; Sunderraman, R. Single valued neutrosophic sets. Multispace and Multistructure 2010, 4, 410–413.

[9] Smarandache, F. Neutrosophy: Neutrosophic Probability, Set, and Logic. American Research Press: Rehoboth, NM, USA, 1998.

[10] Smarandache, F. A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability and Statistics, 4th ed.; American Research Press: Rehoboth, NM, USA, 2005.

[11] Atanassov, K. T. Intuitionistic fuzzy sets. Fuzzy Sets and Systems 1986, 20(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

[12] Zadeh, L. A. Fuzzy sets. Information and Control 1965, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[13] Khalil, M.; Er, E. Will ChatGPT get you caught? Rethinking of plagiarism detection. In Learning and Collaboration Technologies. HCII 2023. Lecture Notes in Computer Science, vol. 14040; Zaphiris, P., Ioannou, A., Eds.; Springer: Cham, Switzerland, 2023; pp. 475–487. https://doi.org/10.1007/978-3-031-34411-4_32

[14] Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; Finn, C. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 24950–24962, 2023.

[15] Ye, J. A multicriteria decision-making method using aggregation operators for simplified neutrosophic sets. Journal of Intelligent and Fuzzy Systems 2014, 26(5), 2459–2466. https://doi.org/10.3233/IFS-130916

[16] Ye, J. Single valued neutrosophic cross-entropy for multicriteria decision making problems. Applied Mathematical Modelling 2014, 38(3), 1170–1175. https://doi.org/10.1016/j.apm.2013.07.020

[17] Biswas, P.; Pramanik, S.; Giri, B. C. TOPSIS method for multi-attribute group decision-making under single-valued neutrosophic environment. Neural Computing and Applications 2016, 27(3), 727–737. https://doi.org/10.1007/s00521-015-1891-2

[18] Giray, L. The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. The Serials Librarian 2024, 85(5–6), 181–189. https://doi.org/10.1080/0361526X.2024.2433256

[19] Asamblea Nacional del Ecuador. Ley Orgánica de Educación Superior (LOES). Registro Oficial Suplemento 297, August 2, 2018.

[20] Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel,

T.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 2023, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[21] Hechavarría-Hernández, J. R.; Leyva Vázquez, M. Y.; Smarandache, F. Neutrosophic stance detection and fsQCA-based necessary condition analysis for causal hypothesis assessment in AI-enhanced learning. Neutrosophic Computing and Machine Learning 2025, 40.

[22] Consejo de Educación Superior del Ecuador (CES). Reglamento de Régimen Académico. Resolución RPC-SO-08-No.111-2019. CES: Quito, Ecuador, 2019.

[23] OpenAI. GPT-4 Technical Report. arXiv:2303.08774 2023. https://doi.org/10.48550/arXiv.2303.08774

[24] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS 2017) 2017, 30.

[25] Foltýnek, T.; Meuschke, N.; Gipp, B. Academic plagiarism detection: A systematic literature review. ACM Computing Surveys 2019, 52(6), 112. https://doi.org/10.1145/3345317

[26] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; et al. Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS 2020) 2020, 33, 1877–1901.

[27] Zawacki-Richter, O.; Marín, V. I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education — where are the educators? International Journal of Educational Technology in Higher Education 2019, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

[28] Gehrmann, S.; Strobelt, H.; Rush, A. M. GLTR: Statistical detection and visualization of generated text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2019); Association for Computational Linguistics: Florence, Italy, 2019; pp. 111–116. https://doi.org/10.18653/v1/P19-3019

Received: Jan 9, 2026. Accepted: Aug 7, 2026