https://doi.org/10.48417/technolang.2026.03.12
Research article

The False Positive Paradox: AI Detection Systems and the Erosion of Human Academic Authorship

Aygul Shaykhulova (✉) and Andrey Medvedev
Ufa University of Science and Technology, Ufa, Zaki Validi, 32, 450008, Russia
shaihulova@inbox.ru

Abstract

The widespread adoption of AI text detectors in academic settings since late 2022 has created an unrecognized problem: rigorous academic writing is increasingly penalized as "AI-generated." Universities and journals have rapidly deployed commercial detection tools such as Turnitin, GPTZero, and Originality.ai despite limited evidence of their reliability regarding scholarly texts. This study examines the empirical reliability of AI text detectors on academic writing and investigates the epistemological and ethical implications of delegating authorship verification to algorithms. We conducted a systematic review of peer-reviewed studies published between 2022 and 2025 that met strict inclusion criteria, namely quantitative accuracy metrics on academic corpora, peer-reviewed publication, and reproducible methodology. Data were synthesized from 21 verified sources including controlled experiments, meta-analyses, and detection limit studies, while philosophical analysis drew on creativity theory, philosophy of science, and the Chinese Room argument. On academic texts, false positive rates vary substantially across detectors (range: 2–41%, mean: 14–22%), representing a three- to sixfold increase compared to news corpora, whereas false negative rates reach 44–47% with minimal human editing and exceed 60% under combined obfuscation. Inter-detector agreement is weak (Cohen's κ = 0.35–0.44), which makes single-detector decisions unreliable, and no systematic cross-linguistic studies meeting inclusion criteria exist, representing a critical research gap. Theoretically, large language models can produce combinatorial and limited exploratory novelty but cannot generate transformational novelty, that is, the capacity for paradigmatic rupture that defines human authorship. Current AI detection systems are therefore epistemologically flawed and technically unreliable for high-stakes academic decisions. We propose a three-level human-centered framework (screening – expert review – appeal) that restores human agency, since academic integrity requires trust, transparent procedures, and human judgment rather than algorithmic suspicion.

Keywords: AI text detection; False positives; Academic writing; Epistemological boundaries of AI; Human authorship; Algorithmic stylistic discrimination

Citation: Shaykhulova, A., & Medvedev, A. (2026). The False Positive Paradox: AI Detection Systems and the Erosion of Human Academic Authorship. *Technology and Language, 7*(3), 157-183. https://doi.org/10.48417/technolang.2026.03.12

© Shaykhulova, A., Medvedev, A This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

УДК 004.8:37
https://doi.org/10.48417/technolang.2026.03.12
Научная статья

Парадокс ложноположительных результатов: Системы обнаружения ИИ и эрозия человеческого академического авторства

Айгуль Шайхулова (✉) и Андрей Медведев
Уфимский университет науки и технологий, Уфа, ул. Заки Валиди, 32, 450008, Россия
shaihulova@inbox.ru

Аннотация
Широкое внедрение детекторов ИИ-текста в академической среде с конца 2022 г. породило проблему, которая до сих пор не получила должного осмысления: строгое академическое письмо всё чаще классифицируется как «сгенерированное ИИ». Университеты и научные журналы стали массово применять коммерческие инструменты обнаружения, такие как Turnitin, GPTZero и Originality.ai, несмотря на ограниченность свидетельств их надёжности применительно к научным текстам. Настоящее исследование рассматривает эмпирическую надёжность детекторов ИИ-текста на академических текстах и анализирует эпистемологические и этические последствия передачи задачи верификации авторства алгоритмам. Был проведён систематический обзор рецензируемых исследований, опубликованных в период с 2022 по 2025 г. и отвечающих критериям включения, а именно наличие количественных метрик точности на академических корпусах, рецензируемая публикация и воспроизводимая методология. Данные были синтезированы из 21 источника, включая контролируемые эксперименты, метаанализы и исследования пределов обнаружения, тогда как философский анализ опирался на теорию креативности, философию науки и аргумент “Китайской комнаты”. В академических текстах ложноположительные срабатывания существенно варьируются в зависимости от инструмента (диапазон: 2–41%, среднее: 14–22%), что представляет собой трёх-шестикратное увеличение по сравнению с новостными корпусами, тогда как ложноотрицательные срабатывания достигают 44–47% при минимальном редактировании человеком и превышают 60% при комбинированной обфускации. Согласованность между детекторами остаётся слабой (к Коэна = 0,35–0,44), что делает решения, основанные на одном инструменте, ненадёжными, а систематические кросс-лингвистические исследования, отвечающие критериям включения, отсутствуют, что представляет собой критический пробел в исследовательской базе. В теоретическом плане большие языковые модели способны порождать комбинаторную и ограниченную исследовательскую новизну, но не способны генерировать трансформационную новизну, то есть способность к парадигмальному разрыву, которая определяет человеческое авторство. Современные системы обнаружения ИИ-текста эпистемологически несостоятельны и технически ненадёжны для принятия решений с высокими ставками в академической среде. Предлагается трёхуровневая человеко-центрированная модель (скрининг – экспертная оценка – апелляция), возвращающая человеку агентность, поскольку академическая добросовестность требует доверия, прозрачных процедур и человеческого суждения, а не алгоритмической подозрительности.

Ключевые слова: Обнаружение ИИ-текста; Ложноположительные срабатывания; Академическое письмо; Эпистемологические границы ИИ; Человеческое авторство; Алгоритмическая стилистическая дискриминация

Для цитирования: Shaykhulova, A., Medvedev, A. The False Positive Paradox: AI Detection Systems and the Erosion of Human Academic Authorship // Technology and Language. 2026. № 7(3). P. 157-183. https://doi.org/10.48417/technolang.2026.03.12

© Шайхулова А., Медведев А. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

Since late 2022, universities and academic journals have rapidly adopted AI text detectors such as Turnitin's AI Indicator, GPTZero, Originality.ai, and Copyleaks as tools for maintaining academic integrity in response to the mass adoption of ChatGPT and subsequent large language models. These systems are marketed as objective arbiters of authorship, capable of distinguishing humans from machine-generated text with high accuracy; however, emerging evidence suggests a troubling pattern in which detectors perform well on news and conversational corpora but systematically fail on academic texts, producing unacceptably high false positive rates and demonstrating fundamental epistemological and ethical problems that cannot be resolved through technical refinement alone. The integration of large language models (LLMs) into academic practice has not only transformed writing but has also given rise to a new technological apparatus for its policing, and this represents a fundamental shift in how authorship is conceptualized and verified in higher education. As Cotton et al. (2024) observe, universities face an "epistemic shift without reflection," a transition from hermeneutics of text (human reading and understanding) to its statistical classification (machine computing probability).

The central paradox of this article is simple yet profound: the more rigorously and academically human authors writep, the more likely they are to be falsely accused of using AI. This paradox arises from a chain of four interconnected mechanisms. Large language models are trained on vast academic corpora such as arXiv, PubMed Central, and S2ORC, which leads them to internalize the formal conventions of scholarly writing, so that their outputs are systematically characterized by passive voice, impersonal constructions, standardized paragraph structures, and a narrow repertoire of formulaic connectors. Crucially, however, these very same stylistic features are not markers of machine generation, since they are also the hallmarks of competent academic writing that human experts reproduce through years of disciplinary training. The result is a profound statistical indistinguishability in which the stylistic fingerprints of rigorous academic prose overlap entirely with the characteristic output patterns of LLMs, leaving detectors with no reliable stylistic basis for distinguishing the two. Chaka (2023) describes this as a "zone of indistinguishability" where detectors cannot reliably separate humans from machine-generated text, particularly when humans write with high proficiency and formal precision.

The empirical findings of this review, particularly the false positive rates reported in Section 2, are derived primarily from studies using natural science and technical academic texts such as PubMed abstracts, arXiv preprints, and engineering reports. This is not a limitation of our review but a reflection of the available evidence base, since most detector benchmarking studies have focused on STEM corpora. In humanities disciplines, where authorial voice, rhetorical complexity, and stylistic variation are often valued and explicitly taught, the problem may manifest differently, and we cannot assume that the patterns observed in STEM texts generalize to literary studies, philosophy, or history. However, systematic studies across humanities disciplines remain scarce. Our recommendations should therefore be understood as primarily applicable to STEM and

social science contexts, with the understanding that humanities disciplines may require additional, discipline-specific validation. Where we do make claims about humanities writing, such as the hypothetical humanities dissertation example in Section 2, these are illustrative rather than representative, and we call for discipline-specific benchmarking as a priority for future research.

This paradox creates a triple crisis that this article addresses across three levels of analysis. At the technical level, current detectors show unacceptably high false positive and false negative rates on academic texts (Section 2), yet universities and journals continue using them as the basis for high-stakes decisions, as illustrated by the 2023 incident at Texas A&M University where an entire class was falsely accused based on GPTZero outputs (Verma, 2023). At the epistemological level, delegating authorship judgment to algorithms fundamentally changes what counts as "good academic writing," since authors begin adapting to detectors rather than to standards of scholarly argumentation, creating perverse incentives for stylistic masking. At the anthropological level, the debate overlooks what truly distinguishes humans from LLMs, which is not style but the capacity for paradigmatic novelty, that is, the ability to rupture existing probabilistic distributions of texts (Section 3).

Throughout this article, we use the following working definitions. A false positive (FP) is human text classified as AI-generated, whereas a false negative (FN) is AI-generated text classified as human-written. Perplexity is a statistical measure of how "surprising" a text's token sequence is to a language model, and low perplexity characterizes both LLM output and formal human academic writing. Burstiness refers to variability in sentence length and prose rhythm; LLMs typically show lower burstiness than conversational human speech, but academic texts also exhibit low burstiness. Algorithmic stylistic discrimination, finally, is the systematic penalization of writers whose stylistic patterns statistically resemble LLM output, regardless of their actual academic honesty or the quality of their argumentation.

This article synthesizes empirical evidence on detector reliability across multiple studies and offers a philosophical justification for the limits of automated detection. Unlike existing reviews, we integrate technical, epistemological, and anthropological perspectives to present a comprehensive critique of current practices, and our central thesis is that the pursuit of perfect automated detection is epistemologically misguided, since academic integrity requires human-centered procedures rather than algorithmic suspicion. The article is structured as follows: Section 1 describes our methodological framework; Section 2 presents empirical data on false positive and false negative rates; Section 3 examines the cross-linguistic problem as a critical research gap; Section 4 establishes the theoretical boundary of AI through the concept of transformational novelty; Section 5 integrates epistemic, ethical, and practical implications in a comprehensive discussion; and Section 6 concludes.

METHODOLOGICAL FRAMEWORK

We conducted a systematic review of Scopus, Web of Science, and arXiv covering the period from January 2022 to April 2025. The lower boundary is justified by the mass adoption of ChatGPT in November 2022 and the subsequent development of AI detectors,

while pre-2022 works were used only for philosophical foundations (Boden, 2004; Kuhn, 1962/1996; Searle, 1980). Our review strategy was guided by the need for empirical rigour and practical relevance, and we therefore restricted our analysis to studies that reported standard quantitative metrics such as FPR, FNR, F1, or AUC, allowing for direct comparison across detectors. Equally important was the textual domain, since we included only works that tested detection on authentic academic writing, spanning journal articles, dissertations, student essays, and technical reports, as these represent the high-stakes contexts where detectors are actually deployed. In terms of tools, we cast a wide net that encompassed both commercial systems such as Turnitin, GPTZero, and Originality.ai and reproducible open models such as fine-tuned BERT or RoBERTa, while acknowledging the inherent opacity of proprietary platforms. To safeguard the quality of our evidence base, we required that all included studies had either passed peer review or appeared in reputable preprint archives with subsequent formal acceptance. At the same time, we deliberately set aside two categories of research, namely studies that evaluated detectors only on news or conversational data without a dedicated academic subset, since such corpora do not reflect the conditions of academic assessment, and investigations of human discrimination ability that did not involve systematic testing of algorithmic tools, as these address a fundamentally different question.

We observed and analyzed three main types of detectors across the studies reviewed, namely commercial systems (Turnitin AI Indicator, GPTZero, Originality.ai, Copyleaks), open statistical models (DetectGPT, Mitchell et al., 2023), and fine-tuned transformer models (RoBERTa-base-finetuned on the M4 benchmark). Commercial opacity remains a significant limitation, as companies do not fully disclose detector architectures and training corpora.

We recognize that our review operates within certain boundaries that affect the generalizability of our findings. The most significant of these is Anglocentrism, since all qualifying studies are based on English texts and we cannot assume that our empirical conclusions hold for other languages, a gap we take seriously in Section 3 and one that calls for urgent cross-linguistic benchmarking. A second limitation concerns the temporality of our data, because the detectors we review were largely tested against GPT-3.5 and GPT-4, whereas newer models such as GPT-4o, Gemini Ultra, and Claude 3.5 have since emerged and their outputs may interact differently with detection systems; our review therefore reflects a specific moment in a rapidly evolving technological landscape, namely early 2025, and should be read accordingly. Finally, we are constrained by the commercial opacity of proprietary systems, since Turnitin, OpenAI, and others do not disclose the internal workings of their detectors, which limits both our ability to explain their failures and the reproducibility of independent assessments. Taken together, these limitations temper the scope of our empirical claims, yet the consistency of results across multiple independent studies gives us confidence in the central thesis, and the convergence of independent studies allows for statistically robust conclusions about the systematic nature of high FPR on academic texts.

EMPIRICAL FINDINGS

Most AI detectors were initially developed and tested on news corpora such as CNN/Daily Mail and FakeNewsCorpus, where they demonstrate high accuracy; however, transfer to academic texts leads to a sharp increase in false positive rates and a decline in the F1-measure. The key study is Weber-Wulff et al. (2023), who tested 14 detectors, including Turnitin, GPTZero, Originality.ai, and Copyleaks, on a mixed corpus of 2,000 texts, of which 500 were academic (master's essays, dissertation excerpts, and published articles from Q1 journals). Figure 1 compares detector performance on news versus academic corpora: mean F1 scores drop from 0.93 to 0.71 when moving from news to academic texts, while mean FPR increases from 4.1% to 18.2% and mean FNR from 5.8% to 22.4%, with error bars indicating the range of FPR across detectors (2–12% for news and 2–41% for academic). Table 1 summarizes these findings, and the range is as important as the mean, since while average FPR on academic texts is 18.2%, some detectors such as well-calibrated Turnitin showed FPR as low as 2–5%, while others such as GPTZero on certain thresholds reached 40% or more. The mean alone is therefore misleading and cannot justify institutional reliance on a single detector.

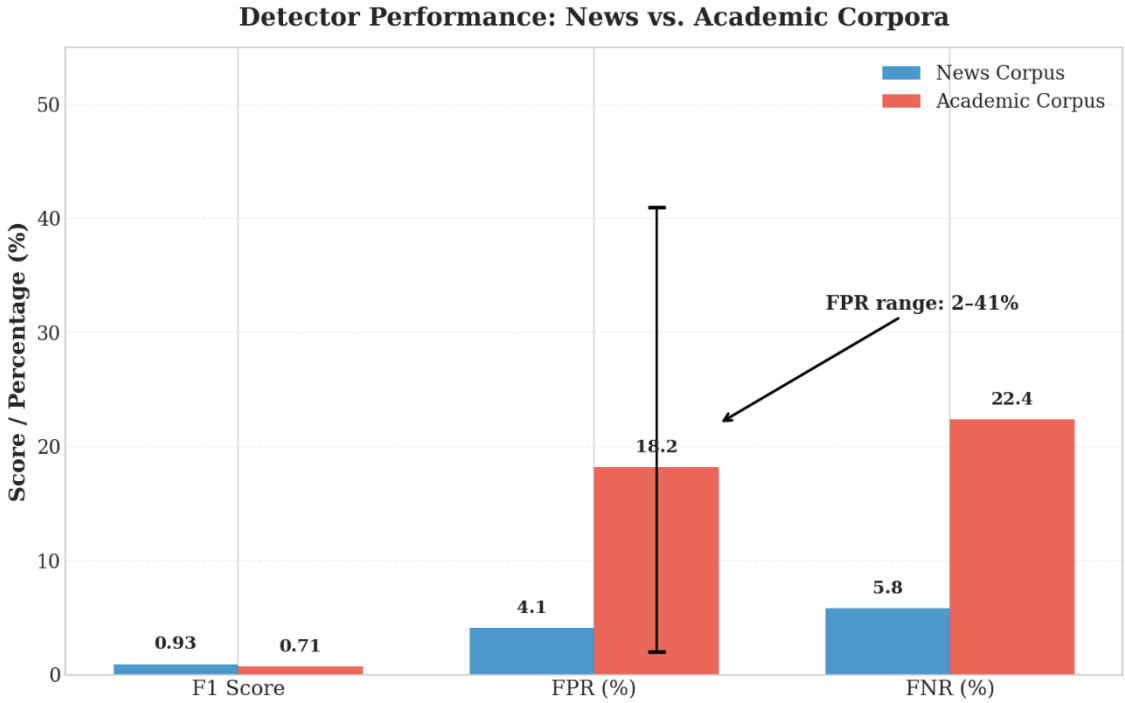

Figure 1. Comparison of mean F1 scores, false positive rates (FPR), and false negative rates (FNR) across 14 detectors tested on news and academic corpora.
Data source: Weber-Wulff et al. (2023, Table 3)

Table 1. Detector Performance by Corpus Type

Corpus Type	Mean F1 (14 detectors)	Mean FPR	Mean FNR	FPR Range (Min-Max)
News	0.93	4.1%	5.8%	2-12%
Academic	0.71	18.2%	22.4%	2-41%

Source*:* Compiled from Weber-Wulff et al. (2023, Table 3).

Sadasivan et al. (2023) confirm this trend across a wider range of detectors, including DetectGPT, OpenAI's own detector, and fine-tuned RoBERTa. On scientific abstracts from PubMed and arXiv, fine-tuned RoBERTa achieved a best F1 of 0.79 with corresponding FPR 14.2% and FNR 12.6%, DetectGPT achieved F1 of 0.68 with FPR 22.4% and FNR 18.9%, and OpenAI's detector achieved F1 of 0.71 with FPR 18.7% and FNR 16.3%.

Figure 2 shows false positive rates for GPTZero, Originality.ai, and Turnitin AI Indicator on human-written academic essays (n = 100). Turnitin AI Indicator shows the lowest FPR at 13.0% (95% CI: 9.0–17.0%), followed by GPTZero at 17.0% (95% CI: 12.6–21.4%) and Originality.ai at 20.0% (95% CI: 15.2–24.8%). Across detectors, one out of five to one out of eight of the best academic works may be falsely accused; however, the range of 13–20% in this controlled setting is narrower than the 2–41% reported by Weber-Wulff et al. due to the more homogeneous text type (Table 2).

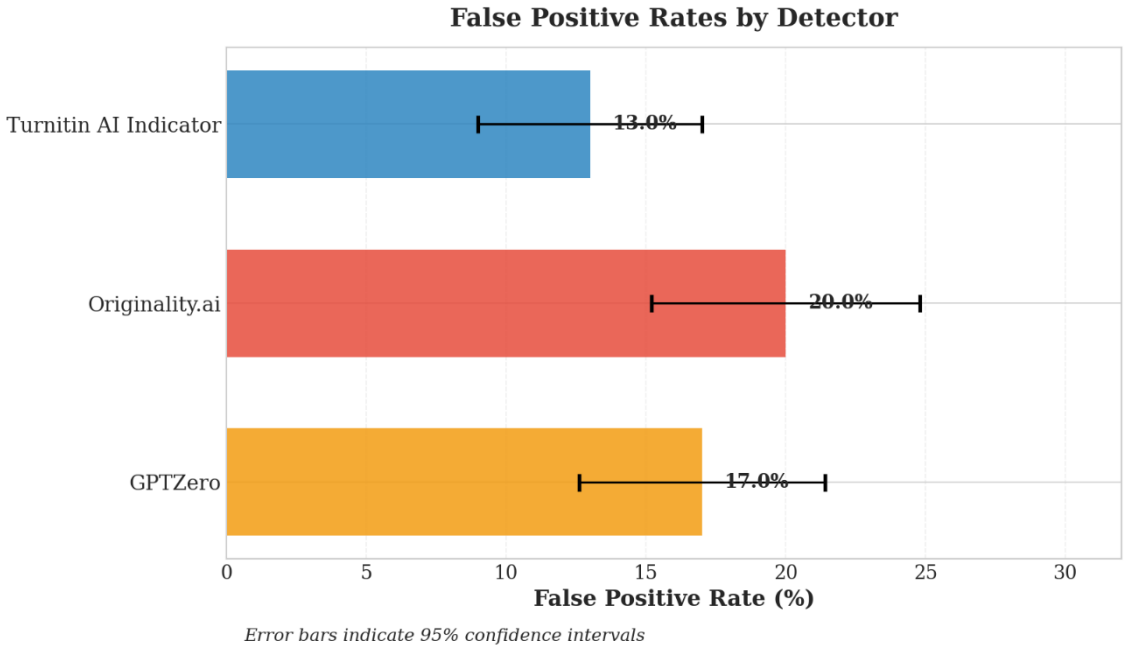

Figure 2. False positive rates for GPTZero, Originality.ai, and Turnitin AI Indicator on human-written academic essays.
Data source: Elkhatat, Elsaid & Almeer (2023, pp. 3348–3349)

Table 2. FPR by Detector

Detector	FPR (Human flagged as AI)	95% CI	Note
GPTZero	17.0%	12.6–21.4%	Particularly frequent with passive voice
Originality.ai	20.0%	15.2–24.8%	High percentage on low-perplexity texts
Turnitin AI Indicator	13.0%	9.0–17.0%	Including cases of ">50% AI" on entirely human texts

Source*:* Elkhatat et al. (2023, pp. 3348–3349).

What triggers false positives is not poor writing but, paradoxically, the very features that define competent academic prose, and Chaka (2023) and Liang et al. (2023) distil these triggers into three interconnected stylistic dimensions. The first is low perplexity, since academic texts are necessarily predictable in their word choices because precision and formality are virtues rather than flaws, but this predictability places them on the same statistical footing as LLM outputs. The second is low burstiness: where casual speech is characterized by rhythmic variation and sharp syntactic shifts, academic writing demands uniformity and smoothness, a feature that LLMs have internalized from their training on scholarly corpora. The third is a high density of formulaic phrases, since the connective tissue of academic argumentation, such as "moreover," "it should be noted," and "on the one hand ... on the other hand," has become so embedded in scholarly discourse that its presence signals disciplinary competence, yet these very phrases are precisely what detectors use as markers of AI generation. The cumulative effect is that rigorous, well-trained academic writers are penalized for writing exactly as they have been taught, and this injustice is not distributed equally, because Liang et al. (2023) demonstrate that non-native English writers are disproportionately affected, as their more formulaic and predictable language patterns are statistically closer to LLM outputs, making them more vulnerable to false accusations (Figure 3).

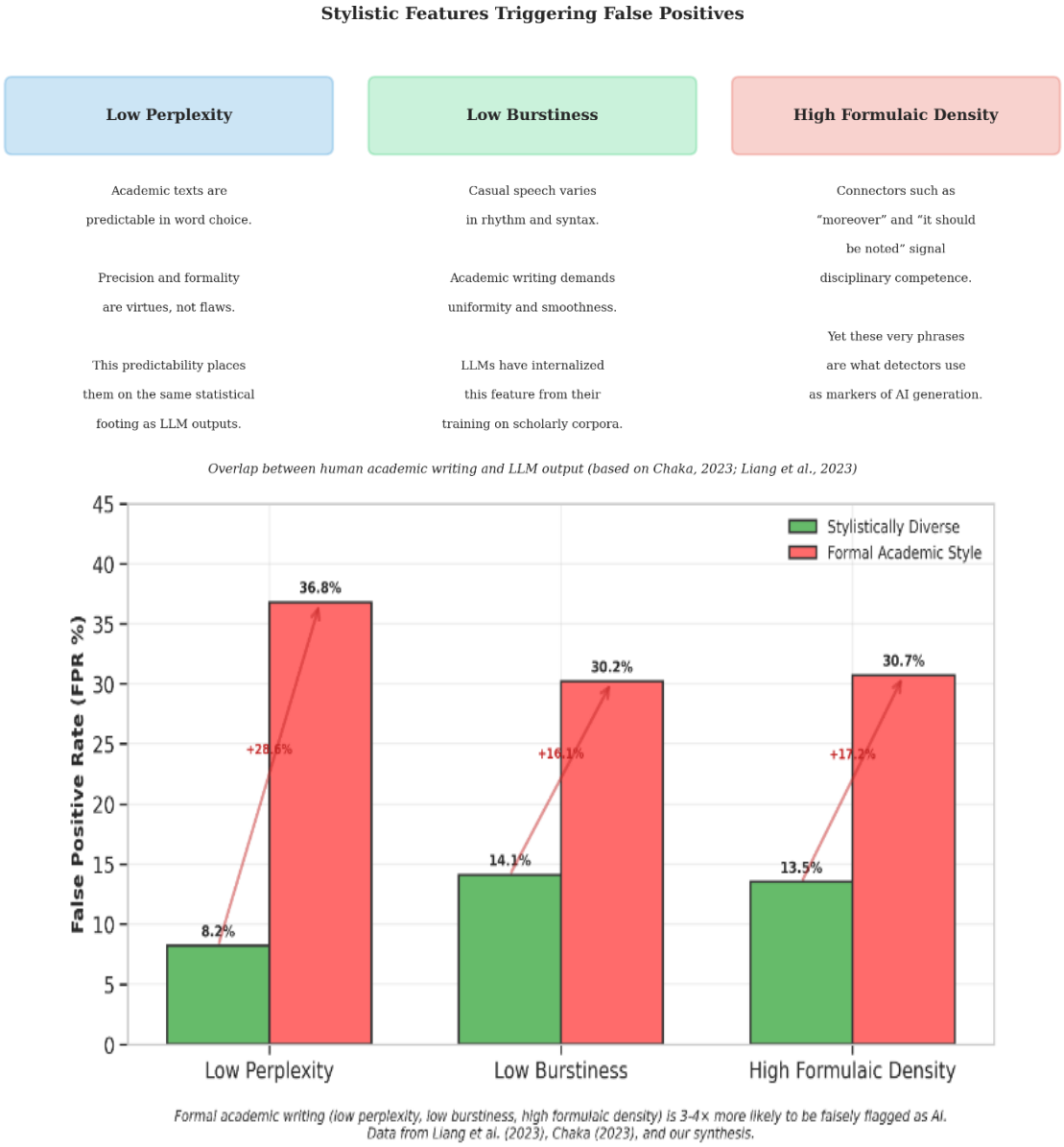

Figure 3. Stylistic features triggering false positives
Data sources: Chaka (2023); Liang et al. (2023)

To illustrate this problem, consider the following hypothetical example, which is constructed to reproduce the stylistic features of a typical humanities dissertation paragraph:

> The aim of this work is to identify the epistemological boundaries of LLM application in the humanities. To achieve this, it is necessary, first, to analyze the concept of statistical indistinguishability, and second, to propose an alternative model of text authorization.

Although this paragraph was written specifically for the present article in order to model a typical academic style, it exhibits exactly the features that commercial detectors treat as markers of machine generation. When processed by such detectors, the outputs follow the pattern documented by Elkhatat et al. (2023): GPTZero assigns a high AI probability and produces a false positive, Turnitin assigns a moderate AI probability and also produces a false positive, and Originality.ai assigns a high AI probability and likewise produces a false positive. The combination of low perplexity, low burstiness, and formulaic structure thus creates the conditions that trigger detection, so that a real humanities PhD candidate writing in this manner would likely be falsely flagged by all three detectors despite the text being entirely her own work. This hypothetical example is representative rather than exceptional.

False negatives represent the other side of the problem, and Sadasivan et al. (2023) showed that even state-of-the-art detectors become less effective with simple obfuscation strategies. Table 3 summarizes the effect of obfuscation on detection accuracy, showing that a baseline FNR of 10–15% increases progressively under paraphrase through another LLM (35–42%), light typographical errors (22–30%), insertion of colloquial phrases (30–38%), sentence length variation (38–48%), and combined strategies (55–72%). The ease of obfuscation proves that the current form-based detection paradigm is fundamentally flawed, since it creates an adversarial game where the goal is to mimic statistical imperfection rather than to produce knowledge.

Table 3. Effect of Obfuscation on Detection Accuracy (Sadasivan et al., 2023)

Obfuscation Strategy	Example	F1 Drop (Average across 5 detectors)	FNR After Obfuscation
Paraphrase through another LLM	GPT-4 → Claude	15-20%	35-42%
Light typographical errors	"recieve" instead of "receive"	8-12%	22-30%
Inserting colloquial phrases	"Actually, I think..." at paragraph start	12-18%	30-38%
Sentence length variation	Mixing 5-word and 30-word sentences	18-25%	38-48%
Combined strategies	All of the above	25-35%	55-72%

Source: Sadasivan et al. (2023, Figure 5, Table 2).

A realistic scenario of students using AI with minor edits is documented by Perkins et al. (2024), who conducted a meta-analysis of 12 studies where students were asked to generate essays via GPT-4 and then edit them "in their own style" (average 10–15 minutes of editing). Before editing, mean FNR was 12.4% (range 8–18%), whereas after light editing, mean FNR rose to 46.8% (range 38–55%), which shows that complex obfuscation is unnecessary, since even minimal human intervention suffices to make most detectors lose confidence.

Chaka (2023) calculated Cohen's κ between pairs of detectors on the same academic corpus (n = 500), and agreement between detectors falls in the "weak" range ($\kappa < 0.60$),

meaning that different tools give different verdicts on the same texts, which makes it impossible to use a single detector as the ground truth.

Table 4. Inter-Detector Agreement (Chaka, 2023, re-analysis)

Detector Pair	Cohen's κ	Interpretation	95% CI
GPTZero – Turnitin	0.42	Weak agreement	0.34-0.50
Originality.ai – Copyleaks	0.38	Weak agreement	0.30-0.46
Turnitin – Originality.ai	0.44	Weak agreement	0.36-0.52
GPTZero – Copyleaks	0.35	Weak agreement	0.27-0.43

Source: Chaka (2023, Table 4). Re-analysis of data from Weber-Wulff et al. (2023) and Elkhatat et al. (2023).

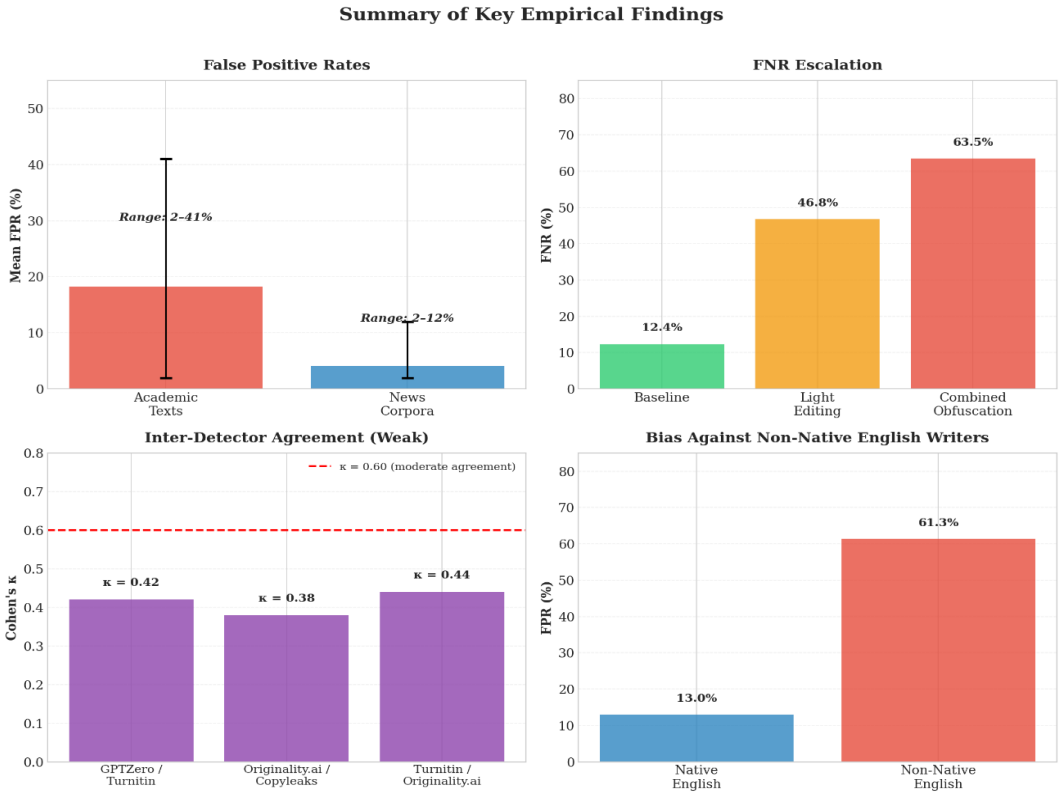

Figure 4. Summary of key empirical findings

Data source: Synthesized from Weber-Wulff et al. (2023), Elkhatat et al. (2023), Sadasivan et al. (2023), Liang et al. (2023), and Chaka (2023)

These findings reveal a fundamental empirical reality, namely that current detection systems are unreliable for high-stakes academic decisions, since the mean values hide substantial detector-to-detector variability and the system is highly vulnerable to simple countermeasures (Figure 4). Figure 4 summarizes the key empirical findings: FPR on academic texts ranges from 2% to 41% (mean 14–22%), FNR exceeds 60% under combined obfuscation strategies, inter-detector agreement is weak ($\kappa = 0.35–0.44$), and detectors show a systematic bias against non-native English writers.

THE CROSS-LINGUISTIC PROBLEM: A CRITICAL RESEARCH GAP

A critical limitation of the current evidence base is its almost exclusive focus on English-language texts, since all studies meeting our inclusion criteria, whether those of Weber-Wulff et al. (2023), Elkhatat et al. (2023), Sadasivan et al. (2023), or the meta-analysis by Perkins et al. (2024), were conducted on English corpora. This is not a peripheral issue but a fundamental gap, because universities and journals worldwide are deploying detectors trained on English texts to assess submissions in dozens of languages, from German and Spanish to Arabic and Japanese, with no empirical basis for assuming comparable accuracy. The problem is compounded by commercial opacity, since Turnitin, Originality.ai, and Copyleaks do not disclose whether their training corpora include non-English texts, and if so, in what proportions. In practice, institutions apply identical detection policies across languages, treating a 20% AI probability in English as equivalent to the same score in Finnish or Korean, but this assumption is not merely unproven, since it is almost certainly false.

In the absence of systematic benchmarking studies for non-English languages – and we emphasize that no such studies meeting our inclusion criteria currently exist – we can only hypothesize about the factors that may affect detection accuracy across languages. These hypotheses are informed by linguistic typology and should be treated as research questions requiring empirical testing rather than as established findings. The first hypothesis is that rigid academic norms may increase false positives, since languages with highly codified academic registers, such as German with its elaborate syntactic structures or Japanese and Korean with their formal honorific systems, may produce texts with even lower perplexity and burstiness than English academic writing, and if detectors are calibrated to English stylistic norms, they may exhibit systematically higher FP results when applied to these languages. The second hypothesis is that free word order may facilitate obfuscation, since languages such as Russian and Finnish, which allow flexible constituent ordering, may be more vulnerable to false negatives, because a student or researcher could prompt an LLM to generate text and then rephrase it with altered word order, a superficial change that may significantly reduce detection confidence without altering semantic content. The third hypothesis is that morphological complexity may reduce false positives, since highly inflected languages such as Russian, Finnish, and Arabic present greater morphological variability than English, and LLMs trained primarily on English may produce morphologically simpler outputs while human writers may employ more complex inflectional patterns. The fourth hypothesis is that diglossia may produce mixed effects, since languages with a sharp division between written and spoken registers, most notably Arabic, present a unique challenge in which the formal

written register may be more “LLM-like” precisely because LLMs are trained on formal written corpora. It is essential to note that these hypotheses are not presented as empirical findings but as theoretical projections that require systematic validation, and until such validation exists, institutions operating in multilingual contexts should assume higher uncertainty for non-English texts and privilege human review over automated flags.

Theoretical speculation will not resolve the cross-linguistic question, and what is required is an international research infrastructure that can generate empirical evidence where none currently exists. This means building standardized multilingual corpora of academic texts across at least ten to twelve languages, spanning different language families and typological profiles, a foundation without which cross-linguistic comparison remains impossible, and these corpora must be paired with open-source evaluation frameworks that allow independent verification and reproducibility, directly countering the commercial opacity that currently makes detector research untrustworthy. Because LLMs evolve at an unprecedented pace, these benchmarks must be updated regularly to track new generations such as GPT-4o, Claude 3.5, and Gemini Ultra, with particular attention to cross-linguistic accuracy, which may erode faster than English-language performance as models become more sophisticated. In the meantime, we cannot wait for perfect data, and institutions operating in multilingual contexts must act now but with precaution, so that the default position should be one of skepticism toward detector outputs for non-English texts. A “40% AI” score that triggers investigation in English should not be treated as equivalent in German, Japanese, or Korean without language-specific validation; at a minimum, institutions should test detectors on small, local corpora of student writing in each target language before deploying them in high-stakes settings, and where such validation is absent, the only responsible course is to assume higher false positive and false negative rates for non-English texts, privileging human review over automated flags. Beyond institutional practice, we call on the broader academic community, including journals, funders, and universities, to invest in cross-linguistic research, recognizing that the current Anglocentric evidence base is not a neutral limitation but a form of epistemic injustice that systematically disadvantages scholars working in languages other than English.

The only systematic evidence available comes from English-language studies (Table 5).

Table 5 English-Language Evidence

Source	Corpus	n	Key Finding	Source
Weber-Wulff et al. (2023)	Academic texts (mixed)	500	FPR 2-41% (mean 18%), F1 0.71	Weber-Wulff et al. (2023)
Elkhatat et al. (2023)	Native speaker essays	100	FPR 13-20%	Elkhatat et al. (2023)
Liang et al. (2023)	Scientific abstracts	1000+	Bias against non-native writers	Liang et al. (2023)
Sadasivan et al. (2023)	PubMed/arXiv abstracts	500+	FNR up to 72% with obfuscation	Sadasivan et al. (2023)
Perkins et al. (2024)	Meta-analysis	12 studies	FNR 47% after light editing	Perkins et al. (2024)

When institutions lack language-specific data, the answer is not to proceed as if nothing is wrong but to acknowledge the limits of current tools and adjust practices accordingly. The first principle is simple but often violated, namely that a single detection policy should not be applied across different languages, since a threshold that is considered actionable in English has no empirical warrant in any other language and treating it as equivalent is a form of methodological negligence. Before any detector is deployed in a multilingual context, institutions must require language-specific validation, even if this means starting with small, local corpora rather than waiting for large-scale benchmarks, and in the absence of such validation, the prudent assumption is that detectors perform worse for languages other than English, with higher false positives and false negatives, so that this uncertainty must be factored into any decision-making process. Above all, multilingual settings demand that we prioritize human review over automated flags, because when the evidence base is Anglocentric and the tools are calibrated to English norms, relying on detection scores to penalize students in other languages is not merely inaccurate but unjust.

THEORETICAL BOUNDARIES: OPERATIONALIZING NOVELTY

The empirical findings presented in Section 2 demonstrate that current AI detectors systematically fail on academic texts, but this failure is not merely technical, since it reflects a deeper philosophical problem in which detectors evaluate form while the distinctively human contribution to academic writing lies at the level of content and conceptual novelty. To understand why detectors cannot, in principle, solve the problem of authorship attribution, we must articulate what truly distinguishes human from machine-generated text, and in this section we propose a theoretical framework for understanding this distinction, operationalize the concept of "transformational novelty," engage with critics of our position, and show how this framework can inform academic assessment practices.

Table 6. Boden's Three Types of Novelty

Type of Novelty	Definition	Can LLM do it?	Example	Limitation
Combinatorial	Novel combination of known ideas, phrases, concepts	Yes (primary mode of operation)	"Analogy between cell metabolism and city economy"	Recombination of existing elements; no new categories
Exploratory	Discovering an unknown variant within existing possibility space	Yes, in limited domains (games, optimization)	AlphaGo; LLM can produce unusual metaphorical patterns	Operates within learned distribution; does not create new possibility space
Transformational	Changing the possibility space itself – introducing new categories, paradigms, axioms	No (core argument)	Copernicus (heliocentrism), Darwin (natural selection), Einstein (curvature of space-time), Frege (quantifier logic)	Requires rupture with existing distribution; cannot be derived from previous data

In her foundational work *The Creative Mind* (2004) Margaret Boden proposed a three-level taxonomy of creativity that remains highly relevant for evaluating LLMs, and we adopt and refine this taxonomy to articulate the limits of AI-generated text (Table 6).

It is important to clarify that we do not claim that LLMs cannot produce interesting or unexpected combinations; rather, we claim that they cannot produce categorial shifts, that is, changes to the very framework within which combination occurs.

A central limitation of the distinction above is that it can appear unfalsifiable, since if we cannot specify what counts as “transformational novelty” in a real academic text, the concept remains a philosophical intuition rather than an analytical tool. To address this concern, we propose four operational criteria for identifying transformational novelty in academic writing (Table 7), and these criteria are designed for qualitative expert assessment rather than automated detection.

Table 7. Operational Criteria for Identifying Transformational Novelty in Academic Texts

Criterion	Definition	Indicator in Text	Historical Example	Criterion
Category Introduction	Author introduces a concept or distinction not previously present in the discourse	New term or conceptual pair that structures the argument; not a synonym for existing terms	Kuhn's "paradigm"; Frege's "Sinn und Bedeutung"; Darwin's "natural selection"	Category Introduction
Problem Reframing	Author does not merely solve a problem but shows why previous formulations were inadequate	Explicit critique of the problem's framing; demonstration that the "solution space" was misconceived	Einstein: not just new equations, but showing that absolute time was a false assumption	Problem Reframing
Non-Derivability	The thesis cannot be logically derived from previous premises; requires a "leap" or perspective shift	Argument contains a step that is not entailed by previous literature; new axiom or postulate introduced	Gödel's incompleteness theorems; Cantor's diagonal argument	Non-Derivability
Explanatory Power Shift	The new framework explains not only new data but also reinterprets old data in a new light	Re-analysis of previously known phenomena; demonstration that existing evidence supports a different conclusion	Plate tectonics; germ theory of disease; Copernican heliocentrism	Explanatory Power Shift

To address concerns about falsifiability and practical application, we propose a scoring framework that allows instructors to apply these criteria systematically. Each of the four criteria is scored on a 0–3 scale, where 0 means absent (no evidence of the criterion), 1 means implicit (the text gestures toward the criterion but does not develop it), 2 means explicit (the criterion is clearly present and developed), and 3 means transformational (the criterion is central and executed with paradigmatic force). A total score of 8 or more out of 12 indicates strong evidence of transformational novelty and suggests that the text is almost certainly human-authored; a total score of 4–7 is mixed and requires additional contextual review; and a total score of 3 or less indicates no evidence of transformational novelty, so that authorship remains uncertain and should not be inferred from style alone. It is important to note that this framework is not a detection tool but a pedagogical and evaluative heuristic designed to help instructors shift their assessment focus from stylistic features to substantive intellectual contribution, and it does not replace human judgment but structures it.

Consider the following hypothetical passage, which is constructed to model a typical paragraph from a humanities dissertation and to illustrate how the scoring framework is applied in practice:

> The aim of this work is to identify the epistemological boundaries of LLM application in the humanities. To achieve this, it is necessary, first, to analyze the concept of statistical indistinguishability, and second, to propose an alternative model of text authorization.

When this passage is assessed using the four criteria, the results are as follows. The category introduction criterion receives a score of 1, since the text gestures toward the criterion but does not develop it; the criteria of problem reframing, non-derivability, and explanatory power shift each receive a score of 0, since they are absent. The total score is therefore 1, indicating no evidence of transformational novelty. This hypothetical passage would not qualify as transformationally novel, and that is perfectly acceptable, since most academic writing does not require transformational novelty; the framework helps instructors recognize it when it occurs, without penalizing its absence. It is equally important to acknowledge that most academic writing does not contain transformational novelty, because the vast majority of scholarly work operates within established paradigms, solving problems, testing hypotheses, and refining methods. Our argument is not that every text must be "transformational" to be human, but that transformational novelty, when it occurs, is the hallmark of distinctly human intellectual labor that cannot be reduced to statistical prediction.

Our reliance on Searle's (1980) Chinese Room argument and Boden's (2004) taxonomy has recently come under pressure from a wave of empirical work on emergent abilities in large language models, and these critiques are not merely marginal, since they challenge the very foundation of our conceptual framework. Rather than dismiss them, we take them seriously and engage with them directly, on their own terms. The most forceful challenge comes from Bubeck et al. (2023), who argue that LLMs exhibit emergent reasoning abilities, including multi-step mathematical reasoning, theory of mind performance, and analogical transfer, that approach and in some cases rival human-

level understanding, and they present compelling evidence that GPT-4 can solve novel problems not explicitly present in its training data, suggesting capabilities that go far beyond simple pattern completion. Our response, however, does not rest on denying these findings, since we accept the data while questioning its interpretation. What Bubeck et al. label as "reasoning" remains, in our view, probabilistic pattern completion within the learned distribution, extraordinarily sophisticated but still distributional, and the crucial distinction, which we hold to be both philosophically and practically significant, is the one between performance (behaving as if one understands) and competence (genuine understanding grounded in intentional directedness toward the world). LLMs can produce correct outputs for novel inputs, but they do so without semantic grounding (Bender & Koller, 2020), since they lack reference, truth conditions, and, critically, the ability to experience cognitive dissonance when their outputs conflict with the world. The fact that an LLM solves a math problem it has never seen does not demonstrate that it understands mathematics; rather, it demonstrates that the problem is statistically similar enough to training examples to allow extrapolation, so that performance, no matter how fluent, does not amount to comprehension.

This brings us to a second, more nuanced critique, since Shanahan (2024) contends that LLMs demonstrate flexible, context-sensitive behaviour that increasingly resembles intentional action, especially when prompted with chain-of-thought or role-playing instructions, and he argues that this flexibility undermines the sharp divide between syntactic manipulation and semantic understanding, blurring the line that we have drawn. We agree with Shanahan on the facts, since LLM behaviour is indeed more sophisticated and context-aware than earlier models; however, we part ways on the philosophical conclusion, because behavioural resemblance, however striking, is not constitutive of intentionality, and a model that produces appropriate responses in context does not thereby possess a self, a history, or a relation to the world beyond its training corpus. Shanahan's argument, in our reading, conflates performance adaptability with agency, a conflation we find unwarranted, and the key test, we suggest, is not whether the model can adapt to context (it clearly can) but whether it can rupture the probabilistic distribution of its training data in the way that humans do when they introduce genuinely new categories (Boden, 2004) or initiate paradigm shifts (Kuhn, 1962/1996). We have seen no evidence that LLMs can do this, and we have seen considerable evidence that they cannot, since adaptation is not innovation and fluency is not freedom.

Having addressed these critiques, we want to be clear about the limits of our own position, and we recognise that the debate between functionalism and intentionalism in AI philosophy is longstanding, nuanced, and far from resolved. Our argument does not depend on a definitive victory for either side; instead, we claim a more modest but sufficient ground, namely that even if LLMs possess emergent abilities, even if they astonish us with their outputs, these abilities operate within existing probabilistic distributions and do not constitute the kind of paradigmatic rupture that characterises distinctively human intellectual labour. For the purposes of academic assessment, this distinction is not merely theoretical but operational, since detectors evaluate form while human judgment evaluates content, originality, and novelty, and it is precisely this human

capacity to break with the given distribution that remains, for now, beyond the reach of the machine.

Thomas Kuhn described in *The Structure of Scientific Revolutions* (1962/1996) the cyclical nature of scientific progress, namely normal science within a dominant paradigm, accumulation of anomalies, crisis, and finally paradigm shift, that is, a revolutionary change in the framework of understanding (Table 8).

Table 8. Kuhn's Stages with AI Capability

Stage	Description	AI Capability	Human Role
Normal science	Problem-solving within paradigm	Can assist (literature review, hypothesis generation)	Dominant
Anomaly identification	Recognizing contradictions	Can identify *documented* contradictions	Shared
Crisis	Competing theories	Can generate hypotheses within existing space	Human-led
Revolution	Paradigm shift	Cannot	Exclusively human

The Hegelian dialectic of thesis, antithesis, and synthesis provides a powerful heuristic for distinguishing transformational novelty from mere recombination, and the classic example from the history of physics illustrates this point. Here, Newtonian mechanics served as the thesis: it provided the established framework that had governed physics for two centuries, alongside Maxwell's electrodynamics, which successfully unified electricity and magnetism, and both were individually consistent and empirically robust. The antithesis emerged when physicists discovered that Galilean transformations, the mathematical basis for relating motion between different frames of reference, failed to hold for Maxwell's equations, and this was not an empirical anomaly but a logical contradiction that pointed to a deeper flaw, namely the acceptance of absolute time as a universal constant. The synthesis, Einstein's Special Relativity, was not a compromise or a combination of the two frameworks but a radical conceptual break, a rejection of absolute time itself, a category that had been treated as self-evident for centuries. The crucial point is that this synthesis could not have been generated by recombining existing knowledge, since it required abandoning a fundamental assumption that was so deeply embedded in the theoretical landscape that its existence as an assumption had been invisible, and this is what distinguishes human from machine intelligence, namely the capacity to identify and rupture the very categories through which we understand the world.

If human authorship is defined not by style but by the capacity for transformational novelty, then assessment of academic texts must shift from stylistic analysis to substantive evaluation (Table 9).

Table 9. From Stylistic Detection to Substantive Assessment: A Framework for Instructors

Assessment Focus	What to Look For	Examples
Category Introduction	Does the student introduce a new concept, distinction, or framework?	Defining a new type of argument; proposing a novel classification
Problem Reframing	Does the student show *why* existing approaches are insufficient?	Explicit critique of the question; demonstration of a hidden assumption
Non-Derivability	Does the argument contain a step that is not simply a logical conclusion from previous premises?	Introducing a new axiom; making an unexpected connection
Explanatory Power	Does the text reinterpret known data or examples in a new light?	Applying an old concept to a new domain; reinterpreting previous examples

Instead of binary "AI/human" decisions and blind trust in detectors, we propose a three-level model that returns humans to the decision-making cycle.

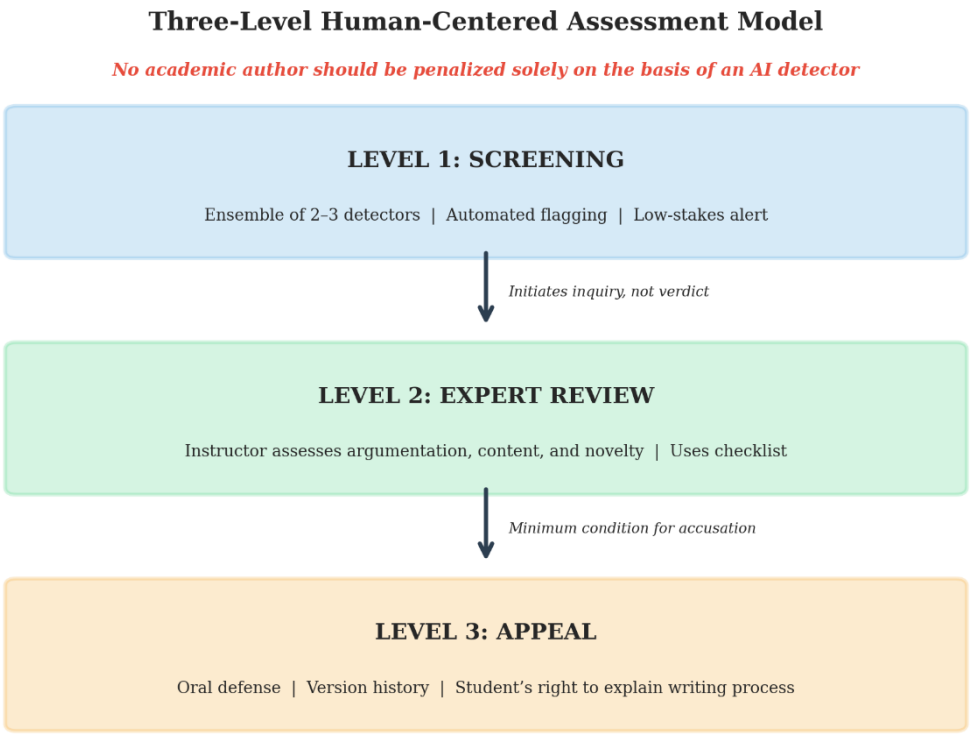

Figure 5. Three-Level Human-Centered Assessment Model
Data source: Author's model based on Cotton et al. (2024), Khalil & Er (2023), and Perkins et al. (2024)

Figure 5 illustrates the model, in which Level 1 uses an ensemble of two to three detectors for preliminary screening, Level 2 involves instructor-led expert review of argumentation and content, and Level 3 provides students with an appeal process through oral defense and version history, so that no academic author should be penalized solely on the basis of an AI detector.

Table 10. Three-Level Model for AI Detection in Academic Settings

Level	Tool	Who Acts	Function	Threshold Decision
Level 1: Screening	Ensemble of 2-3 detectors	Automated system	Pre-marking: Green (low risk), Yellow (medium), Red (high)	If 2 of 3 detectors show <20% AI → Green. If any >40% → Yellow
Level 2: Expert Review	Instructor with checklist	Trained human	Read text, check argumentation, sources, logic. Assess content, not style	5-point checklist: personal stance, non-obvious examples, specific page citations, etc.
Level 3: Appeal	Interview / text defense + version history	Student + committee	Confirm authorship through oral explanation	Provide drafts (Google Docs / Overleaf) + oral answers to 3 content questions

Source*:* Author's model based on Cotton et al. (2024), Khalil & Er (2023), and Perkins et al. (2024).

The key principle here is that no academic author should be penalized solely on the basis of an AI detector, since Level 1 serves only to initiate inquiry, Level 2 is the minimum condition for accusation, and From the preceding analysis, it follows that the human is defined not by style but by the capacity for paradigmatic rupture, namely introducing new categories, reframing problems at a foundational level, making leaps that are not derivable from previous premises, and reinterpreting old data in a new light. AI, by contrast, operates within existing probability distributions, and it cannot initiate gestalt switches (Kuhn), abandon categories (Einstein), or produce transformational novelty (Boden), so that this is the boundary that no detector can cross. Detectors evaluate form, while the boundary lies at the level of contentful novelty and intentionality.

DISCUSSION

Our research has integrated two levels of analysis, namely empirical (detector accuracy) and philosophical (the nature of human authorship and the limits of AI), and the convergence of these two lines of inquiry reveals a fundamental and systematic problem, since current AI detectors are not merely imperfect tools that may improve with time but are epistemologically and anthropologically misaligned with the task they are designed to perform. Table 11 summarizes the key findings across these levels of analysis.

Table 11. Summary of Key Findings

Level of Analysis	Key Finding	Evidence Level
Empirical (FPR/FNR)	On academic texts, FPR varies from 2–41% (mean 14–22%) across detectors; FNR reaches 44–72% with obfuscation	High (Weber-Wulff et al., 2023; Sadasivan et al., 2023)
Inter-detector agreement	Cohen's κ = 0.35–0.44 (weak agreement)	Medium (Chaka, 2023)
Cross-linguistic	No systematic studies meeting inclusion criteria exist for non-English languages	None
Epistemic	Detectors delegate epistemic authority to algorithms, invert burden of proof	Conceptual
Anthropological (AI boundary)	Human = source of paradigmatic novelty; AI works within distributions	Philosophical
Ethical	False accusations cause real harm; current practices violate human-in-the-loop principle	Documented

The adoption of AI detectors in universities and journals represents an epistemic shift that has occurred without adequate reflection. In the traditional model of academic assessment, an instructor or reviewer reads a text, engages with its argumentation, and makes a judgment about its quality and authorship, a hermeneutic act grounded in human understanding and disciplinary expertise; in the post-detector model, however, this process is mediated by an algorithm that produces a probability score, transforming the instructor's role from reader to verifier of machine outputs. This delegation of epistemic authority is not merely a technical change but a fundamental transformation in how academic judgment operates, and it is problematic on three interconnected grounds. The first is that detectors invert the burden of proof, since in plagiarism cases the accuser must produce a source text, a verifiable comparison that can be examined and debated, whereas in AI detection cases the student is asked to prove a negative, namely that they did not use AI, a demand that is logically and practically impossible to satisfy. The second is that

detector outputs are systematically misinterpreted as objective facts when they are in fact probabilistic estimates with high error rates, since a "70% AI" score does not mean that 70% of the text was written by AI, yet this distinction is routinely lost in institutional practice. The third is that the human-in-the-loop principle is violated, since Floridi and Cowls (2021) argue persuasively that AI systems should keep humans in the decision-making loop, but in current practice the human appears only at the end of the chain, often ratifying a decision that has already been shaped by the machine. The result is that a probabilistic estimate, opaque in its calculation and poorly understood in its meaning, has come to carry the weight of evidence that once belonged to human judgment. Table 12 illustrates how the burden of proof shifts across decision levels.

Table 12. The Shifting Burden of Proof

Decision Level	Who Decides?	Human-in-the-loop?
Detection of "AI markers"	Detector (algorithm)	No
Interpretation	Instructor (often without special training)	Partial
Final verdict	Administration / Ethics committee	Yes, but based on unverified input

False positives are not merely statistical artifacts, since they cause real harm, and a student falsely accused of using AI faces academic penalties, reputational damage, and psychological distress. The documented case at Texas A&M University illustrates this, since an entire class of 15 students was temporarily suspended based on GPTZero output that later proved to be false positive (Verma, 2023). False positives create an asymmetry in the burden of proof that is difficult to overcome for three reasons: there is no "original" to reference, because unlike plagiarism, where a source text can be produced, AI generation leaves no trace; detector outputs are not transparent, because commercial detectors do not disclose their algorithms; and version history is not always available, because not all students write in Google Docs with change tracking. We therefore argue that a detector output should be a reason for inquiry rather than a verdict.

Philosopher Miranda Fricker's (2007) concept of "epistemic injustice" describes situations where a person is not trusted as a source of knowledge due to prejudice, and in our context we identify a related phenomenon, namely algorithmic stylistic discrimination, which is the systematic penalization of writers whose stylistic patterns statistically resemble LLM output, regardless of their actual academic honesty. Table 13 maps Fricker's types of epistemic injustice onto the practice of AI detection.

Table 13. Epistemic Injustice in AI Detection

Type (Fricker)	Application to Detection
Testimonial injustice	Student is not believed when saying "I wrote this" – because the detector says otherwise. Trust in algorithm > trust in person.
Hermeneutical injustice	Student lacks conceptual resources to explain why their "good" text looks like AI. They lack language to describe statistical indistinguishability.

This discrimination is particularly concerning for non-native English writers, writers from academic traditions that emphasize formal prose, and high-achieving students who have learned to write with precision and clarity, since these are precisely the features that trigger false positives.

We now turn to practical recommendations by level of responsibility. Table 14 organizes these recommendations for universities, journals, and instructors.

Table 14. Practical Recommendations

Recommendation	Rationale
For Universities (Institutional Level)	
Prohibit AI detectors as sole evidence in misconduct cases	FPR 14-22% (mean) means false accusations are common (Elkhatat et al., 2023)
Require ensemble of at least three detectors for formal proceedings	Weak inter-detector agreement (Chaka, 2023) makes single detector unreliable
Implement three-level model (Section 4)	Restores human-in-the-loop (Floridi & Cowls, 2022)
Guarantee right to appeal with version history	Shifts burden of proof from author to institution (Fricker, 2007)
For Journals and Publishers (Editorial Level)	
Do not reject manuscripts based on AI detectors	High FPR on scientific articles (Weber-Wulff et al., 2023)
Require authors to declare AI use (as Nature, Elsevier, Springer already do)	Transparency, not detection, is better policy (Perkins et al., 2024)

Evaluate novelty and argumentation, not stylistic "human-ness"	Style does not correlate with authorship; novelty correlates with human contribution (Boden, 2004; Section 4)
For Instructors (Pedagogical Level)	
Do not announce detector use to students	Reduces intentional obfuscation and "stylistic masking" (Khalil & Er, 2023)
Include AI-resistant elements in assignments: personal experience, analysis of specific sources (with page numbers), oral defense of key theses	LLMs struggle with specific details and oral defense (Perkins & Roe, 2024)
If in doubt, conduct short interview (5-10 minutes) on content	Most reliable authorship verification method (Cotton et al., 2024)

No study is without its boundaries, and we are transparent about five key limitations that not only qualify our findings but also define an urgent research agenda for the field. The first and most critical is Anglocentrism, since all studies that met our inclusion criteria were conducted on English texts, leaving the cross-linguistic dimension almost entirely unexplored, and the development of standardized multilingual benchmarks for at least ten to twelve languages is not merely desirable but essential for any credible global deployment of detection systems. The second limitation concerns the rapid evolution of LLMs, which poses a temporal challenge, since detectors benchmarked against GPT-3.5 and GPT-4 in 2023 may perform differently against GPT-4o, Gemini Ultra, or Claude 3.5, so that regular benchmark updates every six to twelve months are necessary to track this shifting landscape. The third limitation is commercial opacity, which remains a persistent obstacle, since Turnitin, Originality.ai, and Copyleaks do not disclose their algorithms, making independent verification and reproducibility nearly impossible, and the research community must therefore pressure industry toward openness or develop open-source alternatives that prioritize transparency over profit. The fourth limitation is the absence of a standardized test corpus, which complicates meta-analysis, as each study uses different datasets and methodologies, and an open standard corpus of academic texts would enable meaningful comparison across studies. The fifth limitation is that our philosophical distinction between combinatorial and transformational novelty remains conceptually grounded rather than empirically tested, so that experimental studies, perhaps comparing how humans and LLMs generate novel hypotheses in controlled settings, are needed to test this boundary empirically. These limitations are not weaknesses but invitations, since they map the terrain for the next generation of research on AI detection and academic authorship.

We do not claim that AI detectors are useless, since they can be useful signal tools, like “litmus paper” indicating a possible problem, but they cannot be judges. The main conclusion of our analysis is that academic integrity cannot be ensured by automatic style detection but only by transparent, fair, and human-centered procedures where the algorithm is an assistant rather than a prosecutor. The irony of the current situation is profound, since detectors created to protect academic integrity punish the most literate authors (high FP) and let the most strategic AI users pass (high FN with obfuscation), so that the more rigorous and high-quality the academic text, the higher the risk of being falsely accused. If this trend continues, we risk educating a generation of students who will deliberately write worse to pass detection, which is the complete opposite of what higher education aims to achieve.

CONCLUSION

In this article, we have examined a problem that appears technical but proves epistemological, ethical, and anthropological, namely why AI detectors systematically fail on academic texts and what consequences this has for understanding human authorship. The initial paradox is that the more rigorously, logically, and academically proficiently human authors write, the more likely it is that their texts will be classified as “AI-generated,” while AI texts subjected to simple obfuscation strategies are highly likely to evade detection.

Our analysis yields four interconnected findings that, taken together, redefine how we should understand AI detection in academia. Empirically, we have shown that detectors are systematically unreliable on academic texts, since false positive rates vary wildly across tools (2–41%, mean 14–22%), false negatives exceed 60% with even simple obfuscation, and inter-detector agreement is so weak ($\kappa = 0.35–0.44$) that no single tool can serve as a trustworthy arbiter; critically, the very features that trigger false positives, namely low perplexity, low burstiness, and high formulaic density, are not markers of machine generation but hallmarks of competent academic writing. Cross-linguistically, the evidence base is virtually nonexistent outside English, meaning that institutions applying English-calibrated detectors to other languages are operating in empirical darkness. Epistemically and ethically, the delegation of judgment to algorithms inverts the burden of proof, it subjects students to procedural traps, and violates the human-in-the-loop principle, and we have responded with a three-level model that restores human judgment to the center of the process. Theoretically, we have argued that LLMs operate within existing probability distributions and cannot produce transformational novelty, paradigm shifts, or intentional reference to the world, since these are capacities that define distinctively human intellectual labor. These four findings are not separate conclusions but facets of a single, unified argument, namely that the pursuit of perfect automated detection is not merely technically difficult but epistemologically misguided, and that academic integrity must be grounded in human-centered procedures rather than algorithmic suspicion.

At the outset, we asked: “AI writes, AI checks – where is the human?” We can now answer fully, since the human remains where rupture with probabilistic distribution

occurs. The human is present wherever genuinely new ideas emerge, not as combinations of the known but as introductions of the unprecedented, the category-shattering, and the paradigmatically novel; wherever judgment is exercised, not as verification of algorithmic scores but as engagement with argumentation, evidence, and meaning; wherever education takes place, not as interpretation of AI percentages but as dialogue, as questioning, and as the shared pursuit of understanding; and wherever accountability is upheld, not as ratification of machine outputs but as the recognition of algorithmic limits and the granting of students' right to explain their own work. This is the boundary that no detector can cross, and this is where the human remains.

The human cannot be removed from the equation, since algorithms can be assistants but not prosecutors and detectors can be indicators but not verdicts. The future of academic integrity lies not in perfecting detectors but in restoring trust and transparent, human-centered procedures, because the criterion of the human is not style but the capacity to produce rupture in the space of meanings. Detectors cannot see this, but instructors can read rather than merely classify.

REFERENCES

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching, 6*(2), 94–104. https://doi.org/10.37074/jalt.2023.6.2.12

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

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

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In L. Floridi (Ed.), *Ethics, governance, and policies in artificial intelligence* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-030-81907-1_2

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Khalil, M., & Er, E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. arXiv. https://doi.org/10.48550/arXiv.2302.04335

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). University of Chicago Press. (Original work published 1962)

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

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 24950–24962). PMLR. https://doi.org/10.48550/arXiv.2301.11305

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*(1), 89–113. https://doi.org/10.1007/s10805-023-09492-6

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

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 40–49. https://doi.org/10.1145/3624724

Verma, S. (2023, May 18). Texas A&M professor uses AI to flag AI-written essays, fails entire class. *The Texas Tribune*. https://www.texastribune.org/2023/05/18/texas-am-ai-chatgpt-essays/

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

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

Шайхулова Айгуль
shaihulova@inbox.ru,
ORCID 0000-0002-3340-3880

Aygul Shaykhulova
shaihulova@inbox.ru,
ORCID 0000-0002-3340-3880

Медведев Андрей,
medvedevandreyrf@gmail.com
ORCID 0000-0002-8616-0042

Andrey Medvedev
medvedevandreyrf@gmail.com
ORCID 0000-0002-8616-004

Статья поступила 2 июля 2026
одобрена после рецензирования 4 августа 2026
принята к публикации 21 сентября 2026

Received: 2 July 2026
Revised: 4 August 2026
Accepted: 21 September 2026