**Abstract**

**Computational Nativism and the Politics of Linguistic Legitimacy: A Techno-orientalist Critique of AI Writing Detectors**

**By**

**ESHA ANN MARIYA**

**B.A. English Language and Literature**

**St. Teresa's College (Autonomous)**

**Ernakulam**

**Register No: AB23ENG006**

**2023 - 2026**

**Supervising teacher: Ms. Athira Babu**

This project critically examines AI writing detectors as a system that is structurally biased against non-native English speakers, repeatedly misclassifying their academic writing as AI-generated. With Techno-orientalism as the primary framework, alongside Critical Algorithm Studies and Decolonial Design Theory, the study positions AI detectors as reproducing older linguistic hierarchies. These practices are theorised as computational nativism, a term coined in this project to describe the algorithmic filters dictating the legitimacy of academic authorship on a global scale today. Through a critical analysis of the marketing strategies and popular media discourse surrounding these tools, and original survey responses from Indian scholars, the study traces the normalization of exclusionary practices. Instead of proposing a single technical remedy, the project promotes decolonial design as a transformative approach necessitating comprehensive structural changes.

# COMPUTATIONAL NATIVISM AND THE POLITICS OF LINGUISTIC LEGITIMACY: A TECHNO-ORIENTALIST CRITIQUE OF AI WRITING DETECTORS

*Project submitted to St. Teresa's College (Autonomous) in partial fulfilment of the requirement for the degree of BACHELOR OF ARTS in English Language and Literature*


**By**

**ESHA ANN MARIYA**

**Register No. AB23ENG006**
**III B.A. English Literature**
**St. Teresa's College**
**Ernakulam**
**Cochin - 682011**
**Kerala**



**Supervisor**
**MS. ATHIRA BABU**
**Assistant Professor**
**Department of English**
**St. Teresa's College**
**Ernakulam**
**Kerala**


DEPARTMENT OF ENGLISH
ST. TERESA'S COLLEGE, ERNAKULAM

**March 2026**

**Head of Department:** **Supervisor:** **External:**

# ST.TERESA'S COLLEGE (AUTONOMOUS)
# ERNAKULAM

## Certificate of Plagiarism Check for Dissertation

| | |
|---|---|
| Author Name | Esha Ann Mariya |
| Course of Study | BA English Language & Literature |
| Name of Guide | Ms. Athira Babu |
| Department | English & Centre for Research |
| Acceptable Maximum Limit | 20 |
| Submitted By | library@teresas.ac.in |
| Paper Title | Computational Nativism and the Politics of Linguistic Legitimacy: A Techno-orientalist Critique of AI Writing Detectors |
| Similarity | 1% AI - 14% |
| Paper ID | 5189139 |
| Total Pages | 41 |
| Submission Date | 2026-01-27 12:21:41 |

Signature of Student

Signature of Guide

27/01/26

Checked by
College Librarian

ST. TERESA'S COLLEGE (AUTONOMOUS) LIBRARY ERNAKULAM

* This report has been generated by DrillBit Anti-Plagiarism Software

## DECLARATION

I hereby declare that this project titled “Computational Nativism and the Politics of Linguistic Legitimacy: A Techno-orientalist Critique of AI Writing Detectors” is the record of bona fide work done by me under the guidance and supervision of Ms. Athira Babu, Assistant Professor, Department of English.

Esha Ann Mariya

Register No. AB23ENG006

III B.A. English Literature

St. Teresa’s College (Autonomous)

Ernakulam

Ernakulam

March 2026

**CERTIFICATE**

I hereby certify that this project entitled “Computational Nativism and the Politics of Linguistic Legitimacy: A Techno-orientalist Critique of AI Writing Detectors” by Esha Ann Mariya is a record of bona fide work carried out by her under my supervision and guidance.

Ms. Athira Babu
Assistant Professor
Department of English
St. Teresa’s College (Autonomous)
Ernakulam

Ernakulam
March 2026

**ACKNOWLEDGEMENT**

I take this opportunity to thank God Almighty for showering abundant blessings and grace upon me during the course of my project.

I would like to place on record my sincere gratitude to Rev. Sr. Nilima (CSST), Provincial Superior and Manager, St. Teresa's College (Autonomous), Ernakulam and Dr. Anu Joseph, Principal, St. Teresa's College (Autonomous), Ernakulam for their continued support throughout the course of my study in this institution.

I would like to express my heartfelt gratitude and appreciation to my supervisor Ms. Athira Babu for guiding my thoughts in the right direction and for helping me to express them in the best possible manner.

I extend my sincere gratitude to the Head of the Department, Dr. Tania Mary Vivera and all the other teachers of the department without whose guidance this project could never have been completed.

**CONTENTS**

# Introduction

The rise of generative AI in academic spaces has changed how students and educators approach writing. Many students today use AI to create drafts or outlines for their assignments, while many teachers also use AI to review their students' assignments. The rise in AI usage has created a corresponding rise in AI writing detection tools, specifically to determine if students are using AI to generate their own assignments. These tools claim to be able to differentiate between texts generated by humans from those generated through AI. The tools have also made the space that was previously reserved for personal reading and interpretation of student's writings by teachers, into a space that is governed solely by technology. Even students that do not utilize AI to write assignments are still required to have their submissions evaluated by these tools. The output of these tools is viewed as an authenticating document that indicates the legitimacy of an assignment as well as its origin.

At least superficially, the companies producing AI writing detection software seem to be working with the interests of originality and academic integrity. However, an examination of how these detection softwares function reveals that there is a great deal of variability in terms of their ability to recognize original work. Scholars have documented that writings produced by students for whom English is a second language (non-native English speakers or NNES) were significantly more likely to be misidentified as being created with the assistance of artificial intelligence than writings from native-English speakers; regardless of whether or not each piece was completely created by humans. (Liang et al.) Initially, corporations dismissed these results as a technical glitch and claimed that they would be able to resolve the issue by expanding their training data and re-training their algorithms. Even after these measures, the

problem continues to exist. This continued existence directs our attention to something far greater than simply a coding error.

While detection technologies appear to be neutral tools designed to assist educators in detecting plagiarism, they are actually validation mechanisms that are built upon the norms of Western-style academic English – which is fluent, has rhythm, and assumes nativization. Moreover, despite the publicly-stated intentions to correct past biases, detection technologies continue to operate as gatekeeping systems which validate certain voices while deeming others to be counterfeit. In this regard, these systems are reminiscent of the colonial era perceptions regarding indigenous writing, where indigenous writing was viewed as mimicry rather than original expression, only now they are operating in the form of algorithmic processing. The broader patterns of favoring the characteristics of English spoken by native speakers as the default markers of genuine human authorship is what this dissertation refers to as computational nativism.This idea can be better understood through the lens of techno-orientalism.

Techno-orientalism is a cultural discourse that considers both the East and contemporary technologies to be productive yet shallow (no individuality); efficient yet unsympathetic (lacking in emotional texture). In this way, the concepts of the East and the machine are combined: mechanical, second rate and less than a human entity. AI writing detectors repeat the same ideology. While processing NNES writing, many times these systems interpret it as formulaic or machine-like because of how much it diverges from the default patterns of English written in the West. Even when corporations promise recalibration to address the issue, the fact that these tools continue to exhibit bias demonstrates that their underlying assumptions about language use are built into the system and hence remain unchanged. Under slogans like “preserve humanity in writing”

or “ensure transparency,” detectors imply that NNES linguistic expression sits closer to AI generated texts.

The Techno-orientalist perspective further develops with the concept of Linguistic Imperialism to contextualize a longer historical context of English positioning itself as the universal, rational and academically legitimate language; explaining why deviations from that model can so quickly be viewed as algorithmic output. Critical Algorithm Studies expands on this analysis: since algorithms themselves are not neutral or objective (meaning they represent social and technical constructs of culture, ethics and politics), they assist in allocating epistemological authority to particular identities and/or groups at the expense of others, and in determining whether an individual's writing represents an authentic writing voice or a questionable one. The intersection of these three theoretical perspectives illustrates how, through their evaluative algorithms, AI writing detectors continually reproduce biases and how NNES scholars are disproportionately impacted by misclassifications. Decolonial design theory is another framework that complements these perspectives. It critiques the colonial legacies inherent in modern technology and seeks to rebuild those technologies in ways that foreground marginalized knowledge systems and value collaborative, culturally-based design.

The dissertation is structured as follows. Chapter 1 develops the theoretical base from which the dissertation proceeds: Techno-orientalism and its associated concepts. Chapter 2 outlines the central argument and provides a critical analysis of the discourse that surrounds AI writing detectors (specifically their marketing claims) and the tensions between those claims and the global debates about misclassification and the potential harms of it. Supplementing it is an original study on a survey done to contextualise the

perceived experience and perception of NNES scholars who have been vulnerable to misclassification by algorithms and are pressured to use them by institutions. Chapter 3 then focuses upon the ethical implications of these systems. The problem of persistent bias despite the repetition of intervention is tackled, forming the basis for a discussion of decolonial design theory in order to consider possibilities for reform that are based upon ideas of just futures. While it does not offer specific technical solutions, the dissertation discusses conceptual routes to thinking about the design of systems other than within the dominant Eurocentric norms, and developing new approaches that are rooted in justice, plurality and the linguistic realities of a world that exists beyond the West.

This dissertation is of interest for an academic world that increasingly depends on AI-based tools to assess authorship, authenticity and intellectual credibility, as it draws attention to a structural issue that corporations usually downplay. The dissertation also aims to add to current discussions about ethical and political aspects of language and authorship with regard to AI systems in educational settings.

# Chapter 1

## How Writing Comes to be Judged and What that Reveals about Language and Power

The concept of Techno-orientalism emerges at the intersection of orientalist discourse and late 20th century narratives of technological progress. The term was introduced in 1995 by David Morley and Kevin Robbins in their essay "Techno-Orientalism: Japan Panic," published in *Spaces of Identity*. Written in the context of Japan's rapid technological ascent, they describe how Japan began to be framed as a Techno-other and how this characterisation unsettled Western cultural and economic stability. (Morley and Robins 147) This early articulation becomes a precursor to the concerns later developed by Lisa Nakamura.

> "Cybertypes are the images of race that arise when the fears, anxieties, and desires of privileged Western users (the majority of Internet users and content producers are still from the Western nations) are scripted into a textual/graphical environment that is in constant flux and revision." (Nakamura 6)

Although Nakamura does not explicitly use the term Techno-orientalism, her critique of how racial images circulate in supposedly open and universal digital spaces aligns closely with this theory.

Techno-orientalism gains a fuller theoretical formalization in Roh, Niu and Huang's 2015 edited volume *Techno-Orientalism* which defines it as

> "…the phenomenon of imagining Asia and Asians in hypo- or hypertechnological terms in cultural productions and political discourse. Techno-Orientalist imaginations are infused with the languages and codes of the

> technological and the futuristic. These developed alongside industrial advances in the West and have become part of the West's project of securing dominance as architects of the future, a project that requires configurations of the East as the very technology with which to shape it." (Roh et al. 2)

Asia is envisioned simultaneously as technologically hyperadvanced and always othered; both as an emotionally opaque space of machine-like precision, and a technologically precise one. This image resonates with Edward Said's argument that "Orientalism was ultimately a political vision of reality whose structure promoted the difference between the familiar (Europe, the West, "us") and the strange (the Orient, the East, "them")." (Said 43)

The earlier conceptualizations of Techno-orientalism were concerned with how the West perceived the East through images, stories, and fantasies of culture. However, in today's world, those have been translated into forms that are operationalized inside of institutional frameworks, through methods (like procedures, tools, metrics, protocols), in what constitutes day-to-day administrative work in academic governance. This marks a shift as the object is no longer representation alone but academic work itself. The figure of the Techno-other escapes the easier logic of a stereotype and instead leads to an epistemic problem centred on knowledge, trust, and authorship.

Even more importantly, this scepticism is not caused by evidence, but anticipated by it. Bias works preemptively. Human authorship is preemptively distrusted. This move from imagining the East and then appraising its academic work is pointing toward another shift in how academic value will be judged. Techno-orientalist features of machinic, hyper-efficient and automatised writing discursively operate through discourse by means of language, policy and institutional convictions especially in AI writing

detection systems. NNES writing is therefore trapped in a condition of structural vulnerability where academic qualities such as fluency, coherence, and stylistic neutrality are not perceived as indices of perfection but as proof of inauthenticity. Competence itself flips into incrimination. And whilst these institutional systems appear neutral and technical, they still exercise racialised assumptions even through forms of violence that are subtle, bureaucratic and deniable.

One insight is common across all these texts – discourses about technology and progress are never neutral. They are racialized, mythologized and rooted in longer histories of Western cultural imperialism. Roland Barthes defines myth as a type of speech that naturalizes ideology. (Barthes 107) When looked at this way, the narratives that currently represent algorithmic systems as objective, scientific and inevitable are myths. By mythologising AI writing detectors, these systems are positioned as bodiless truth arbiters. The cultural narratives that shape this logic, or the dominant ideology that underpins it, become obscured, making way for the assumption that the judgment of a computational system is inherently neutral.

But Techno-orientalist myth-making actively reproduces authority. Such AI writing detectors are produced as objective, scientific and neutral judges of authorship despite the fact that they are interpretive and political bodies. This way of framing makes contestation difficult, for how can you argue with neutrality? Consequently, other interpretations of textual variation such as systemic differences in language use or non-native expression are being shut down. The credibility of these systems does not derive primarily from their proven accuracy or reliability, but instead from the fact that they correspond to the institutional values of efficiency, objectivity and standardization. With the increasing strength of algorithmic authority, human assessment is pushed to the

background. Intention, context, explanation become even more diluted. As a text is flagged, there is less and less space to contest suspicion; not because it is impossible to resist but because the architecture of the system favors machine judgment over human explanation.

Kate Crawford challenges this myth directly arguing that " …AI is neither artificial nor intelligent. Rather, artificial intelligence is both embodied and material, made from natural resources, fuel, human labor, infrastructures, logistics, histories, and classifications." (Crawford 8) Then AI writing detectors can be interpreted as Techno-orientalist relics that perpetuate racialized hierarchies of the colonial past while purporting to be neutral defenders of academic integrity – a stance that is closer to a veil or curtain than an accurate description.

This phenomenon can also be examined through the concept of linguistic imperialism. Robert Phillipson defines linguistic imperialism as the structured dominance of one language over others, sustained through institutional, cultural, and economic power. (Phillipson 47) Academic English, associated with clarity, linearity, and nativeness, has long been regarded as a rational and universal standard. This is seen in the continued evaluation of postcolonial literatures in terms of English canons. Despite the status of English as a global lingua franca, it is a language which carries normative assumptions already built into it. Thus, NNES writers tend to adapt to an anglophone standard, a hierarchy that becomes computationalized within AI writing detectors.

Linguistic logs too are transformed through this hierarchy when culturally and historically contingent standards of academic English become operationalized within algorithmic systems. Based on norms developed within predominantly anglophone academic traditions, features like syntactic regularity, predictable vocabulary, stylistic

uniformity and linear argumentation are operationalized and turned into a set of benchmarks for an algorithmic judgment. Once encoded, these norms do not look like cultural preferences but function as ostensibly neutral indicators of legitimacy and quality. Linguistic difference is no longer read as variation within a plural linguistic landscape but is reinterpreted as deviation from an assumed norm – deviance which now comes to mean deficiency.

The plurality expressed by World Englishes is consecutively collapsed into a binary of writing that is either recognisable or unrecognisable. Algorithmic systems are not designed to recognise strategic code-switching, postcolonial hybridity or culturally specific rhetorical traditions and as such misclassify. While linguistic imperialism was once imposed through channels such as educational systems and institutional filters, today it is enforced through algorithmic classification – a form of enforcement that is more opaque, harder to challenge and endowed with more authority simply by virtue of its presentation as technical and objective.

Critical Algorithm Studies foregrounds the ideological and material aspects of algorithmic evaluation. This is because far from being neutral, algorithms function as sociocultural artefacts bearing values, assumptions, and biases. This framework reconceptualises AI writing detectors as infrastructures of power. From this perspective, algorithmic evaluations do not merely assess but actively shape authority, legitimacy, and results. Algorithmic bias, therefore, is not a technical glitch or accidental error but simply an inevitable consequence of systems that have been trained on historically uneven data and developed within unequal social and institutional power relations. Classification mechanisms privilege certain linguistic forms while marginalising others, and this privileging is directly built into the logic of algorithmic judgment.

"Machine learning is not the root cause of inequality, but it can and often does encode and reproduce social inequalities." (Alegria and Yeh 38) Native-speaking norms embedded within training datasets, benchmarks, and evaluative thresholds result in a narrow template for legitimate authorship. Writing that fails to fit inside this template is read as anomalous. This constitutes how computational nativism – the privileging of linguistic traits associated with native English speakers as the default marker of authentic human authorship – functions as an epistemic filter, determining which forms of writing are legible and which forms of knowledge count. This process is further entangled with data colonialism, wherein expressive human labour (especially that of NNES writers nowadays) is extracted, quantified, and imported as raw material for algorithmic training. In this extraction, context, authorial intention, and even agency are stripped away while benefits accrue disproportionately to Western corporations. When systems like these misclassify and discredit the labour of these communities, the issue exceeds being just a technical error and becomes epistemic injustice in the form of systematic devaluation of particular voices, knowledges, and modes of enunciation. Hence AI writing detectors operate as contemporary gatekeeping mechanisms that reinscribe colonial hierarchies under the guise of computational objectivity.

Decolonial design theory offers a critical framework for the reimagining of AI writing detection tools by contesting their epistemic foundations rather than merely treating bias as a technical flaw to be corrected. This perspective begins from the premise that design itself is an epistemic practice; one which is shaped by historical relations of power, knowledge hierarchies, and institutional worldviews. From this standpoint, bias and misrecognition are not accidental failures but predictable outcomes of how these systems are designed, justified, and naturalised through claims of technological neutrality. As Cruz argues in his decolonial philosophy of technology, claims of

technological neutrality function to naturalise specific worldviews while obscuring the colonial epistemologies embedded within technical systems. (Cruz) As a result, calls to improve AI writing detectors cannot be reduced to just better calibration or optimisation. They require a reorientation of the framings of legitimacy, authorship, and evaluation that structure system design.

Arturo Escobar argues that design practices must center relationality, plurality, and community autonomy rather than hegemonic technical paradigms. His concept of the pluriverse, a multiplicity of world-making practices, pushes back against techno-neocolonial narratives. (Escobar) These vantage points align with ongoing initiatives in pluriversal language technology, like the LiveLanguage approach, which advocates for integrating small and minority languages into design. Applied to AI writing detection, this framework calls to look at epistemic recognition. Whose writing is recognised? Under what conditions? According to which evaluative logics? By challenging neutrality, universality, and scalability as the default design imperatives, decolonial design opens up space for alternative modes of evaluation that acknowledge linguistic difference without construing it as risk, deviation, or deficiency. In doing so, it offers a theoretical reorientation which can open up radically new ways of thinking about authorship, recognition, and academic legitimacy.

Taken together, these frameworks do not simply accumulate into a theoretical assemblage but form a layered analytic structure that operates across cultural, linguistic, computational, and epistemic registers. They reposition AI writing detectors not as neutral technical instruments, but as cultural and literary artefacts – systems that participate in the production of meaning, value, and authority. They become legible as sites where racialisation, language politics, and epistemic power converge; where

histories of colonial classification, linguistic hierarchy, and technological myth-making are translated into computational form. In this sense, AI detectors do not merely evaluate writing; instead they reproduce regimes of recognition that determine whose language counts as human, whose knowledge appears legitimate, and whose voice is rendered suspect within contemporary academic culture.

# Chapter 2

## How AI Writing Detection is Experienced in Academia

The public-facing rhetoric of AI writing detectors such as Turnitin and now GPTZero shows how new technological systems continue to replicate older racial and linguistic hierarchies and do so while presenting themselves as neutral tools. Their assertions of safeguarding "academic integrity" and "preserving what's human" operate as cultural narratives that normalise particular assumptions about writing and language. When read with Techno-orientalism, these narratives no longer appear neutral and instead as ideological performances that hide the systems' epistemic foundations about what is considered as legitimate writing.

Turnitin's language operates through a kind of myth-making in the Barthesian sense: it naturalizes its evaluative capacities by tethering it to academic virtue. Phrases like "cultivating citizens of integrity" turn writing into a quantifiable moral act. Instead of helping students navigate the plurality of academic expression, such platforms frame writing as something whose authenticity can be and needs to be objectively measured. This conception of objectivity is precisely what Techno-orientalism helps to debunk: Western corporations cast their technologies as neutral guardians of legitimacy even as they inscribe cultural norms into the very infrastructure of evaluation. The upshot is a consequential shift where linguistic conformity is treated as ethical correctness and nonconformity becomes a source of risk.

GPTZero indulges this even further with a more overt mythological stance. Its promise to "preserve what's human" imagines a situation in which human expression is under threat and must be protected via technological intervention. This narrative

constructs a culturally specific ideal of the human voice that is fluent, native, unambiguous, and stylistically attuned to Anglo-academic rhythm. Techno-orientalism helps reveal this familiar binary in which the non-Western or the hybrid is rendered suspect or machinic. The detector presents itself as a saviour of the ‘correct’ (Western) linguistic norms while simultaneously claiming its universalism. Studies from PeerJ and Hadra et al. and a series of detection-bias audits show that NNES writing, especially those exhibiting multilingual rhythm, indirectness, or containing non-canonical phrasing, is disproportionately flagged as AI-generated. This fallacy does not arise from deception but from the algorithm’s narrow definition of what constitutes legitimate English. When GPTZero boasts that it has “reduced ESL (English as a Second Language) false positives,” it has also reduced linguistic diversity to a statistical problem rather than a valid form of expression. This is where Critical Algorithm Studies comes into play: detectors encode existing older hierarchies into classification pipelines and turn differences into flaws under the pretext of scientific neutrality.

The rhetoric of precision – “confidence scores,” “component models,” “false positive rates below one percent” – function as tech-mythology. The more opaque the system becomes, the more it performs transparency through numbers. This process echoes Kate Crawford’s critique of technical language and how they create the illusion of openness while obscuring the infrastructures, datasets, and cultural judgements that shape algorithmic classification. What appears as precision is, in practice, a culturally curated aesthetic of purity. The detector's moral authority derives less from demonstrable correctness than from this sustained performance of quantification of accuracy.

Turnitin’s discourse operates along a similar moral-linguistic hierarchy, if not worse than GPTZero and other platforms. By framing linguistic conformity as a part of

academic citizenship, it affirms Phillipson's argument that English has historically been associated with rationality, legitimacy and moral superiority.

The system's evaluative criteria implicitly tends to reward those whose writing resembles the linguistic style of the Inner Circle, while marks the linguistic style of the Outer and Expanding Circles as unclear or algorithmic. Kachru's model can be read in a way to defang detectors who collapse English as a singular ideal, negating its actual global plurality. A multilingual text becomes unpredictable within such systems not because it lacks authenticity but because it falls outside of the model's narrow myopic centre. This rupture is pertinent to studies on hybrid writing practices. Barrot and Aranda demonstrate that detectors fail almost entirely with texts that combine human and AI edits (a practice popular among writers refining their drafts).

What gets flagged as AI-like frequently ends up being perfectly normal linguistic strategies like declarative sentences, compressed arguments, edits with the help of grammar tools, or second-language rhetorical pacing. Systems designed around monolingual and monocultural norms are ill-equipped to recognize the nuance of writings that emerge from multilingual backgrounds and do not conform to the simplistic rigid human/machine dichotomy. This is a structural issue. The failure lies in the narrow linguistic imagination of the detectors.

The psychological ramifications of this phenomenon documented by Jiang et al. make the issue even sharper. NNES writers self-censor, simplify vocabulary, flatten arguments and avoid experimenting. They create 'safe' English to avoid drawing the algorithm's suspicion. This is computational nativism in practice: writers adjust themselves to conform to a model that never included them in the first place.

Global studies echo the same pattern. The Stanford study showing 61% of TOEFL essays flagged as AI (Liang et al.) and the contradictory scores documented by Malik and Amjad all show that authenticity is not a computational category but a cultural judgement that has been woven into software. Institutions in the Global South, such as the University of Cape Town, have already recognised this discrepancy, rejecting Turnitin's AI Score for unduly penalizing students whose linguistic styles diverge from Anglophone norms. (Mashinini) China drives the point home with a starker example: a century-old classical essay was flagged as mostly AI-written, exposing how detectors collapse when they assess the very literatures it claims to protect. (Peng)

Detectors are complicit in propagating a cultural narrative in which linguistic difference primarily associated with Asia, Africa, and Latin America becomes a site of algorithmic suspicion. The figure of the suspicious NNES student mirrors the Techno-other. Too mechanical when concise, too flowery when indirect, too polished when edited and too imperfect when unedited.

The system is never questioned; the writer is.

Media discourse amplifies this trope with claims like 'too perfect' or 'mechanical', pathologizing non-Western cadence with dishonesty. AI detectors encode a monolingual Anglocentric model of language in which deviation is deception. Their proposed fixes of de-biasing models, recalibrating thresholds and adding ESL categories do not challenge this foundational assumption that English has a single, ideal form. Instead they reinscribe colonial hierarchy on the pretext of improvement. In this way, their public narratives function as techno-myths that disguise the cultural labour being done by the system.

Far from preserving the human, these detectors preserve instead a very specific historical construction of the human – fluent, linear, standard, and disciplined. Global South scholars face the most scrutiny as their writing sits outside the narrowly constructed linguistic norm. Writing styles that are simply unfamiliar to models trained on Inner Circle English, unfamiliar to institutions accustomed to monolingualism, and unfamiliar to technological imaginations shaped by techno-orientalist anxieties about the purity of authorship in a globalized world get flagged as AI.

Along with this critical discourse analysis, this dissertation also incorporates a small-scale, focused survey (See Appendix for survey titled Deconstructing AI's gaze on Academic Writing) of 45 NNES scholars working across various academic contexts. The respondents are a mixture of research scholars, postgraduate students, early-career academics, and faculty members, all of whom regularly engage in academic writing and have direct experience with AI writing detectors.

The purpose of this survey is not to establish statistical generalisability or to re-prove arguments well-documented in existing scholarship. Instead it intends to anchor these arguments within lived academic experience foregrounding how AI writing detection is encountered, interpreted, and emotionally navigated by those who are routinely subjected to its evaluative scrutiny or compelled to interact with such tools through institutional requirements, supervisory practices, or self-checking. The scope of the survey was intentionally kept narrow to prioritise depth, specificity, and reflexive engagement over breadth or representativeness. Responses are approached as situated testimony and are read ethnographically. They are treated as contextual material that illuminates what it is like to be subject to the aforementioned processes in this dissertation.

Almost all participants (44 out of 45) explicitly acknowledged that their languacultural background shapes how they write academically, with the majority describing this influence to be moderate or significant – an observation that reiterates the theoretical claims outlined before. (See Figure 1) Rather than being constructed as a deficit or weakness, this influence is conceptualised as a normal state of affairs of academic writing.

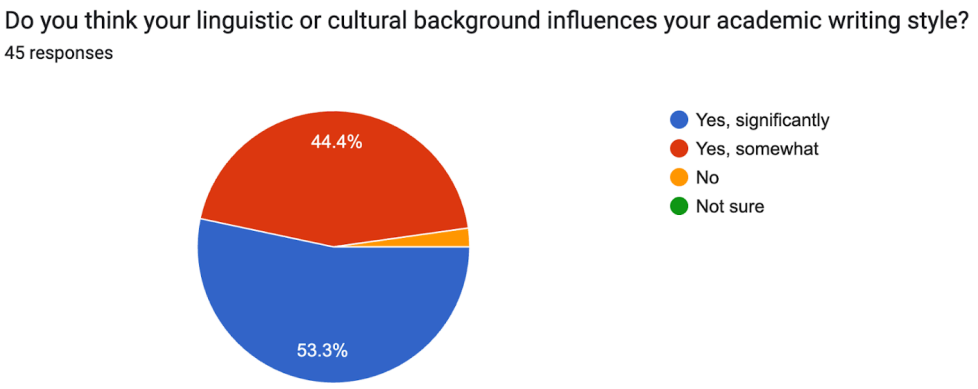


Figure 1. Perceived influence of linguistic and cultural background on academic writing among survey respondents

However, when respondents reflect on voice, this relationship becomes more negotiated. Many feel that their academic writing reflects their academic identity only partially (See Figure 2), and this implies that academic voice is not essentially a matter of self-expression, but also one of accommodation entailing continuous assessments about how much of oneself can appear on the page.

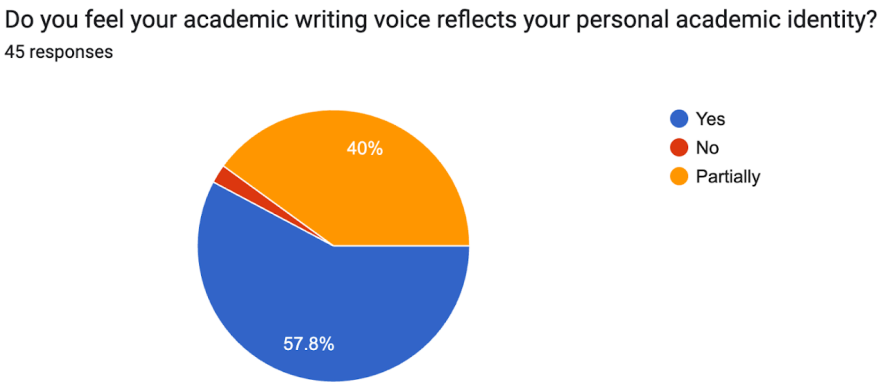


Figure 2. Relationship between academic writing voice and self-perceived academic identity.

This tension becomes apparent in responses to questions about constraint. Few respondents described experiencing constant pressure to conform to a global standard of academic English, while others noted that such pressure appears intermittent and situational, emerging in specific contexts rather than operating uniformly. Similarly, advice to 'westernise' one's writing emerges as a significant but not universal experience.

A substantial proportion report having used AI writing detection tools, indicating that these systems are no longer exceptional or marginal but have become a routine part of academic practice. Nevertheless, familiarity does not translate into confidence. Despite regular use, trust ratings cluster largely in the lower to middle ranges (See Figure 3), implying that AI detectors are approached with a grain of salt. Notably, this skepticism persists even though relatively few respondents disclosed having their writing formally falsely flagged as AI-generated. Distrust, therefore, does not stem primarily from experiences of punishment or penalisation, but from the opaque logic of these systems regarding how they read, interpret, and judge writing.

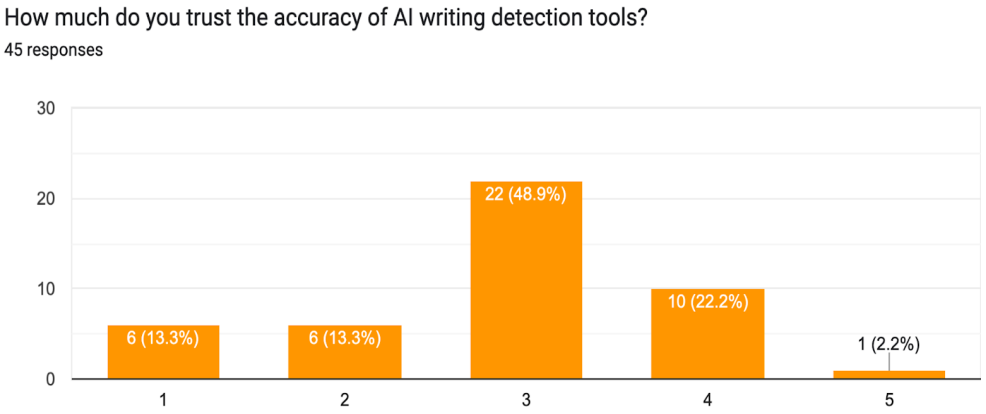


Figure 3. Respondent perceptions of the accuracy of AI writing detection tools on a 5-point Likert scale

This opacity is significant in case of responses concerning creative or experimental language, which a considerable number of respondents describe as either misclassification as AI-generated (12 out of 45) or uncertainty about whether such writing might get flagged (15 out of 45). Even when writers don't face direct accusations they are left second-guessing themselves.

The impact of AI writing detectors manifests in the anticipatory adjustments respondents claim they make to their writing practices. A large section reports consciously modifying tone, voice, and idiomatic expressions due to the perceived risk of being flagged, even when such changes are not grounded in prior experiences. (See Figure 4) This produces a striking imbalance. While relatively few respondents say that they have been formally falsely flagged, a much larger proportion describe altering their writing pre-emptively. This asymmetry is a reminder of how power operates in advance – through self-censorship and internalised regulation – rather than primarily through

overt punishment. In this context, AI writing detectors function less as reactive evaluative mechanisms and more as pre-emptive behavioural regulators.

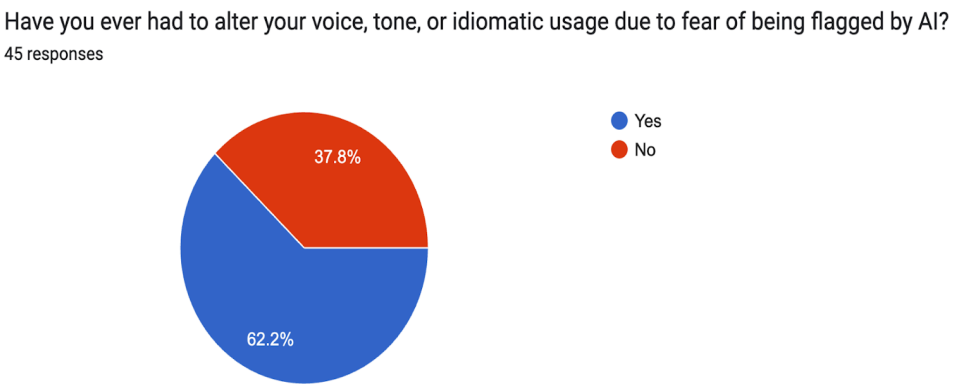


Figure 4. Self-reported modification of academic writing in response to AI detection concerns.

Perception of bias intensifies this effect. Many respondents believe that detectors penalise multilingual, hybrid, or non-standard academic styles (25 out of 45); a few remain uncertain about whether native English norms are systematically privileged (11 out of 45). Although there are some dissenting voices, the prevalence of uncertainty itself is consequential. It reflects the opaqueness of these tools and particularly the absence of clear criteria for what is legitimate writing. Implying that uncertainty is produced by the detectors themselves.

Respondents' accounts of engagement with AI writing detectors are not only shaped by the systems themselves but also by the institutional contexts in which they are deployed. Institutional policies governing AI detection as well as the awareness of these policies are significantly inconsistent. While some respondents report the existence of formal guidelines, a comparable number state that no such policies are in place or they are unaware of them altogether. This ambiguity becomes more pronounced around

questions of consent and choice. Participation operates as a default condition rather than an informed or voluntary choice.

There is a parallel lack of training and institutional support, with nearly all respondents reporting no formal instruction, guidance, or explanation of how AI writing detection tools work. The cumulative effect of this lack results in the displacement of responsibility onto the shoulders of individual scholars. They are left to navigate opaque systems on their own, expected to anticipate risk and manage suspicion without institutional clarity or protection.

When asked to characterise their overall experience with these systems as empowering or alienating, responses span unevenly across empowerment, ambivalence, and alienation. Instances of alienation are detailed, concrete, and grounded in specific experiences of frustration, stylistic constraint, and observed harm, whereas expressions of empowerment tend to be vague, conditional, and instrumental. Institutional opacity and algorithmic opacity together reconfigure academic evaluation, normalising uncertainty and urging writers to adapt without clear rules, stable recognition, or transparent standards of legitimacy.

# Chapter 3

## Ethical Challenges and Decolonial Reimagining of AI Writing Evaluation

This chapter shifts the trajectory of the dissertation into the ethical terrain by addressing why the NNES experiences constitute an urgent issue within contemporary higher education academic institutions. Selvam and Vallejo conclude that AI systems carry forward the biases built into the data that trains them. (Selvam and Vallejo) Computational inequalities emerge from structural inequalities, and older hierarchies materialize inside algorithmic decision-making. Their impact is not theoretical. At the heart of these issues is a repetitive pattern, not an isolated error. When misclassification follows a discernible pattern, it produces what can be called procedural injustice. NNES scholars are increasingly suspected, required to defend the authenticity of their work, and subjected to evaluative standards that their native English speaking peers are not. In academic contexts where originality and authorship carry high stakes, such differential treatment undermines the principle of equal exposure and erodes trust in evaluative systems.

Opacity exacerbates these concerns. Transparency is often cited as a central value in responsible tools; detectors largely operate without it. They produce percentages like ‘63% AI,’ ‘28% mixed’ or ‘confidence scores’ but do not explain how those numbers came to be. Even recent features which explain why a text is labeled AI-generated give only vague reasons such as ‘mechanical transitions,’ without showing where they occur or why they matter. Authority gets transferred from human evaluators, who can justify and revise their assessments, to algorithmic systems whose reasoning remains inaccessible. For NNES writers, whose language usage may naturally differ from dominant expectations, this ambiguity becomes a trap. Their ordinary choices appear

suspicious for reasons they cannot identify or challenge. This dynamic atrophies student autonomy, leaving writers in a continually defensive stance where they must continually justify the authenticity of their work against a system whose judgment is unquestionable. Ethical evaluation requires contestability; systems that foreclose explanation also foreclose justice.

Questions of consent and privacy add a further layer of ethical concern. Many institutions require students to submit their writing to third-party AI detection services as a criteria of assessment, often without a genuine opportunity to consent. Their intellectual labour is transformed into data which is stored, assessed and circulated under terms that are largely incoherent to those who produce it. As researchers have noted, AI systems frequently gather data without meaningful consent, producing long-term risks such as profiling, surveillance, and misappropriation. (Debnath et al.)

From an ethical perspective, coerced participation is not the same as consent. When students are required to submit their intellectual labour to opaque commercial systems in order to be assessed, participation ceases to be voluntary and the outcome becomes to resemble involuntary data extraction. What is framed institutionally as procedural evaluation operates as a form of exploitation in practice. This is especially consequential for NNES scholars, who may already experience heightened vulnerability within academic institutions. Their writing is not only misinterpreted but also funneled into platforms over which they have no control, undermining both autonomy and ownership. The harm accumulates as NNES scholars are disproportionately misread, denied access to the rationale behind those misreadings, and constricted to avoid challenging the system that produces them.

The problem transcends questions of technical accuracy. What prevails is a mode of computational discrimination that extends into autonomy, privacy, agency, and the right to interpret and represent one's own work. The issue, then, is not simply whether detectors can be made more precise, but whether the models within which they operate respect fundamental principles of fairness, and epistemic dignity.

A productive solution begins with recognising that the flaws of AI writing detectors are less isolated technical problems and more symptoms of a deeper epistemic narrowing. Modern AI systems, to a greater extent, rely on enclosures that privilege certain linguistic norms while concealing the colonial assumptions behind them. (Mollema) Now a decolonial approach requires the opposite – opening the system up so that plural linguistic worlds can exist without being forced into a single standard. This direction calls for a decolonial AI that is rooted in awareness of power, context, and cultural difference. (Mohamed et al.) Fairness, in this sense, is not achieved from fine-tuning the existing system but from rethinking how we define authorship, legitimacy, and the value of linguistic diversity in the first place.

A decolonial redesign also involves addressing who gets to shape these technologies. Cruz argues that the philosophy of technology has long been Western-centric, defining what counts as a good system through narrow, exclusionary frameworks. This helps explain why detectors keep reproducing linguistic inequality. They are built on assumptions that leave out Global South practices of writing and knowing. (Cruz) Decolonizing AI writing detection means inviting NNES writers, educators as well as Global South institutions into conversations around training data, evaluation criteria, error tolerances, and governance. This approach aligns with

Mohamed, Png, and Isaac's notion of "reverse pedagogies," where historically marginalised communities take an active role in shaping it. Detectors, in this view, must not simply permit linguistic variety but treat it as central to how writing is understood. (Mohamed et al.)

Institutional practices, too, need to move toward trust and not suspicion. Rendering detectors as de facto judges undermine academic integrity and creates an atmosphere of worry and distrust. Emphasis should be on a model where AI scores are seen as starting points for conversation, not as final evidence of misconduct. (Giray et al.) A trust-based framework supplemented by clear institutional policies, mandatory human review, educator training, and availing appeal routes returns interpretive authority back to human hands. Paired with a decolonial design ethos, this development offers a holistic answer: detectors need to be rebuilt to accommodate linguistic plurality, and institutions need governance policies that uphold dignity, autonomy, and equity. Only through this combination can AI writing detectors move toward ethical, inclusive, and genuinely decolonial practice.

# Conclusion

This dissertation set out to investigate an issue that appears simple on the surface but carries deep consequences. AI writing detectors consistently misclassify NNES writing as AI-generated, and this pattern continues even after repeated technical 'fixes.' The fundamental questions of why detectors remain biased and why technical improvements fail needs an explanation that goes beyond computation. Using Techno-orientalism as the guiding framework, supported by linguistic imperialism, critical algorithm studies, and decolonial design theory, this dissertation argues that detector bias is not a technical mistake but an inherent characteristic of the epistemic foundations shaping ideas of language, writing, and authenticity.

The analysis chapters examine discourse and design of detectors through two popular platforms - Turnitin and GPTZero. Their public narratives about linguistic purity, neutrality, and technical objectivity belie their universalism. Critical discourse analysis reveals these models being tethered to Western academic English. This standard becomes the silent benchmark for human writing, against which deviation is rendered suspicious. AI systems position linguistic differences as deficiency, turning non-normative English into a computational anomaly. These outcomes reflect the hierarchies that are already embedded in training data and design choices. These chapters also extend to the ethical domain and lived experience, revealing a layered structure of harm. Misclassification patterns, obscure methods of functioning, and institutional mandates to use detectors converge to produce a state of sustained vulnerability. Survey responses highlight that students have no knowledge regarding how their writing is judged, cannot ascertain why a passage is labelled "AI-like," and often cannot opt out. Patterns of self-censorship, stylistic flattening, and fear of institutional consequences highlight the uneven burden

placed on NNES writers who must navigate systems that treat their linguistic identity as forms of risk and not legitimate linguistic expression.

Detector bias persists because they rely on narrow, Western-centric assumptions about legitimate English. Misclassification is not a minor statistical aberration but is an outcome of monolingual training methodologies and what this dissertation names as computational nativism: the privileging of linguistic traits associated with native English speakers as the default marker of authentic human authorship.

Technical solutions alone cannot resolve these problems because they leave the underlying epistemology intact – systems developed within a closed linguistic frame. Decolonial thinking calls for shifting these tools away from gatekeeping and towards systems that acknowledge linguistic plurality rather than discipline it. This means design principles centred on transparency, human interpretation, pedagogical guidance, and an expanded understanding of World Englishes. While no single technical blueprint is proposed, this dissertation makes the ethical demand clear: design must move from control to care, from suspicion to recognition.

The dissertation repositions AI writing detection as a cultural, ethical, and epistemic problem rather than a purely technical one. It introduces Techno-orientalism into the discursive politics of AI writing detection, where scholarship remains limited, and demonstrates how linguistic imperialism persists in algorithmic form. By bringing in decolonial design theory, the study reframes detector bias as a question of epistemic justice rather than computational error.

The dissertation also recognises its limitations. The survey sample is modest, no technical audit of detectors is undertaken, and the focus remains largely on Anglophone contexts. Nevertheless these boundaries also open possibilities for further research with

larger quantitative studies, comparisons across Global South institutions, collaborative development of decolonial language models, ethnographic research on student experiences, and evaluations of multilingual detection systems.

Ultimately, this dissertation argues that detectors do not solely misclassify texts but they misrecognize people. They misrecognize World Englishes, global scholarship, and the full range of human linguistic expression. Creating a more ethical academic future, then, calls for dismantling these enclosures and imagining AI systems that recognise, rather than erase, the complexity of human language.

Works Consulted

Alegria, Sharla, and Catherine Yeh. “Machine Learning and the Reproduction of Inequality.” *Contexts*, vol. 22, no. 4, 2023, pp. 34 - 39. *Sage Journals*, https://journals.sagepub.com/doi/10.1177/15365042231210827. Accessed September 2025.

Barrot, Jessie S., and M. R. R. Aranda. “Efficacy of AI-Text Detection Tools in Distinguishing Student-Produced, AI-Edited, and AI-Generated Essays.” *Tech Know Learn*, 2025. *Springer Nature Link*, https://link.springer.com/article/10.1007/s10758-025-09884-0. Accessed December 2025.

Barthes, Roland. *Mythologies*. Edited by Annette Lavers, translated by Annette Lavers, Vintage, 1993. *Internet Archive*.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2022. *Essra*.

Cruz, Cristiano Codeiro. “Decolonizing Philosophy of Technology: Learning from Bottom-Up and Top-Down Approaches to Decolonial Technical Design.” *Philosophy & Technology*, vol. 34, 2021, pp. 1847 - 1881. *Springer Nature Link*, https://link.springer.com/article/10.1007/s13347-021-00489-w#Abs1. Accessed December 2025.

Debnath, Rishab, et al. “AI and Privacy: Ethical Concerns in Data Collection and Surveillance.” *International Journal for Multidisciplinary Research (IJFMR)*, vol. 6, no. 6, 2024. *International Journal For Multidisciplinary Research*, https://www.ijfmr.com/research-paper.php?id=32150. Accessed October 2025.

Escobar, Arturo. *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. Duke University Press, 2018. *Academia.edu*.

Giray, Louie. “The Problem with False Positives: AI Detection Unfairly Accuses Scholars of AI Plagiarism.” *The Serials Librarian*, vol. 85, no. 5-6, 2024, pp. 181-189. *Taylor & Francis Online*, https://www.tandfonline.com/doi/abs/10.1080/0361526X.2024.2433256. Accessed January 2026.

Giray, Louie, et al. “Beyond Policing: AI Writing Detection Tools, Trust, Academic Integrity, and Their Implications for College Writing.” *Internet Reference Services Quarterly*, vol. 29, no. 1, 2025, pp. 83 - 116. *Taylor and Francis Online*.

GPTZero. *AI Detector - Free AI Checker for ChatGPT, GPT-5 & Gemini*, https://gptzero.me/.

Hadra, Mohammad, et al. “Evaluating the Accuracy and Reliability of AI Content Detectors in Academic Contexts.” *Research Square*, vol. Preprint (Version 1), 2025. *Research Square*, https://www.researchsquare.com/article/rs-7359956/v1. Accessed December 2025.

Jiang, Yang, et al. “Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers?” *Computers & Education*, vol. 217, 2024. *ScienceDirect*, https://www.sciencedirect.com/science/article/abs/pii/S0360131524000848. Accessed October 2025.

Liang, Weixin, et al. “GPT detectors are biased against non-native English writers.” *Patterns*, vol. 4, no. 7, 2023, p. 100779. *Cell Press*, https://www.cell.com/patterns/fulltext/S2666-3899(23)00130-7. Accessed October 2025.

Malik, Muhammad Abid, and Amjad Islam Amjad. “AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by

ChatGPT, Perplexity, and Gemini?" *Journal of Applied Learning & Teaching*, vol. 8, no. 1, 2025.

Mashinini, Gugulethu. "UCT scraps flawed AI detectors | UCT News." *UCT News*, 24 July 2025, https://www.news.uct.ac.za/article/-2025-07-24-uct-scraps-flawed-ai-detectors. Accessed December 2025.

Mohamed, Shakir, et al. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *Philosophy & Technology*, vol. 33, 2020, pp. 659 - 684. *Springer Nature Link*, https://link.springer.com/article/10.1007/s13347-020-00405-8. Accessed January 2026.

Mollema, W. J. T. "Decolonial AI as Disenclosure." *Open Journal of Social Sciences*, vol. 12, 2024, pp. 574 - 603. *arxiv*, https://arxiv.org/abs/2407.13050. Accessed December 2025.

Morley, David, and Kevin Robins. *Spaces of identity : global media, electronic landscapes and cultural boundaries*. Routledge, 1995. *Monoskop*.

Nakamura, Lisa. *Cybertypes: Race, Ethnicity, and Identity on the Internet*. Routledge, 2002. *Internet Archive*.

Peng, Dannie. "Explainer | AI content detector: why does China dismiss it as 'superstition tech'?" *South China Morning Post*, 7 June 2025, https://www.scmp.com/news/china/science/article/3313319/ai-content-detector-why-does-china-dismiss-it-superstition-tech. Accessed December 2025.

Phillipson, Robert. *Linguistic imperialism*. OUP Oxford, 1992. *Internet Archive*.

Pratama, Ahmad R. "The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication." *PeerJ: Computer Science*, vol. 11,

no. e2953, 2025. *PeerJ*, https://peerj.com/articles/cs-2953/. Accessed September 2025.

Roh, David S., et al., editors. *Techno-Orientalism: Imagining Asia in Speculative Fiction, History, and Media*. Rutgers University Press, 2015. *Internet Archive*.

Said, Edward W. *Orientalism*. Penguin, 1995. *Monoskop*.

Selvam, Muthu, and Rubén González Vallejo. "Ethical and Privacy Considerations in AI-Driven Language Learning." *LatlA*, vol. 3:328, 2025.

Turnitin. *APAC | Empower Students to Do Their Best, Original Work*, https://in.turnitin.com/.

# Appendix: Survey Questionnaire

## Deconstructing AI's gaze on Academic Writing

Section 1:

Hello

I am Esha Ann Mariya, a final year BA English Language and Literature student from St. Teresa's College, Ernakulam. As part of my final year project, I am conducting a study examining potential cultural and linguistic biases in AI writing detection tools, with a focus on how these systems assess texts written by non-native English-speaking scholars, particularly from the Global South. This survey collects demographic information, personal experiences, and short writing samples to evaluate how these systems perform across different linguistic backgrounds.

If you have any experience with AI writing detection tools and are willing to share your experiences and writing samples, I kindly request you to participate in this survey. The data will be used exclusively for academic purposes in my project and will not be shared beyond this context. This form will take approximately 10 minutes to complete.

Thank you for your valuable time and participation in this study.

Do you agree to participate in this survey and allow your responses (including your writing sample) to be used for academic research purposes?

- Yes, I consent.
- No, I do not consent.

Section 2: Academic Background & Writing Practices

What is your current academic level?

- Postgraduate
- Research Scholar
- Faculty
- Postdoctoral
- Other: _____

Which academic discipline or field is your research in?

_____

How frequently do you write in English for academic purposes?

- Always
- Often
- Sometimes
- Rarely
- Never

Have you received any formal training in academic writing in English?

- Yes
- No

Have you published or submitted articles to English-medium academic journals?

- Yes
- No

Do you think your linguistic or cultural background influences your academic

writing style?

- Yes, significantly
- Yes, somewhat
- No
- Not sure

Do you feel your academic writing voice reflects your personal academic identity?

- Yes
- No
- Partially

Do you feel pressure to conform to a global standard of academic English that limits your creativity?

- Yes, often
- Sometimes
- Rarely
- Never

Have you ever been advised to adjust or ‘westernize’ your writing style for academic purposes?

- Yes
- No

Section 3: Experiences with AI Writing Detection Tools

Have you ever used an AI writing detection tool?

- Yes
- No

How much do you trust the accuracy of AI writing detection tools?

Do not trust at all

- 1
- 2
- 3
- 4
- 5

Completely trust

Have you ever had your writing falsely flagged as AI-generated by such a tool?

- Yes
- No

If yes, please provide a short excerpt of the flagged text. (min. 70 words)

_____

If your writing was flagged, did it affect your submission or evaluation in any way?

- Yes
- No
- Other:

Has an AI detector ever misinterpreted your creative or experimental use of language as AI-written?

- Yes
- No
- Not sure

Have you ever had to alter your voice, tone, or idiomatic usage due to fear of being flagged by AI?

- Yes
- No

Do you think AI detectors penalize multilingual or hybrid styles of academic writing?

- Yes
- No

Do you believe AI detectors are more effective at recognizing native English writing than non-native writing?

- Yes
- No
- Not sure

Do you think AI detectors normalize a single global standard of academic English?

- Yes

- No
- Not sure

Are there clear institutional policies explaining how AI detection results are used and interpreted?

- Yes
- No
- Not sure

In your academic context, do you feel you can opt out of such AI-based surveillance practices?

- Yes
- No
- Not sure

Do you think these tools are used punitively rather than supportively in your context?

- Yes
- No
- Not sure

Does your institution provide training or awareness sessions about AI detectors and their implications?

- Yes
- No

- Not sure

Do you know what happens to your uploaded text after it is processed by an AI detector?

- Yes
- No

Were you ever asked for consent before your writing was checked with an AI detector?

- Yes
- No

Do you believe your submitted work could be stored or used to train AI models without your consent?

- Yes
- No
- Not sure

Would you feel comfortable sharing your writing with these tools if you were given full transparency?

- Yes
- No
- Maybe

Section 4: Perceptions, Bias, and Recommendations

Do you believe AI detectors threaten your academic integrity or ownership of your work?

- Yes
- No
- Not sure

Overall, would you describe your experience with AI detectors as empowering or alienating?

_____

What changes would you suggest to make AI writing detectors more inclusive and fair?

_____

How should academic institutions in India approach the use of AI writing detection tools?

_____

Do you think AI literacy should be included in research training?

_____

Do you have any additional comments or reflections about your experiences with AI writing detection tools?

_____