When Human Writing Looks Like AI

Understanding Turnitin False Positives and why the algorithm can't always tell the difference.

Prof Mphahlele Ramashego Shila

University of South Africa

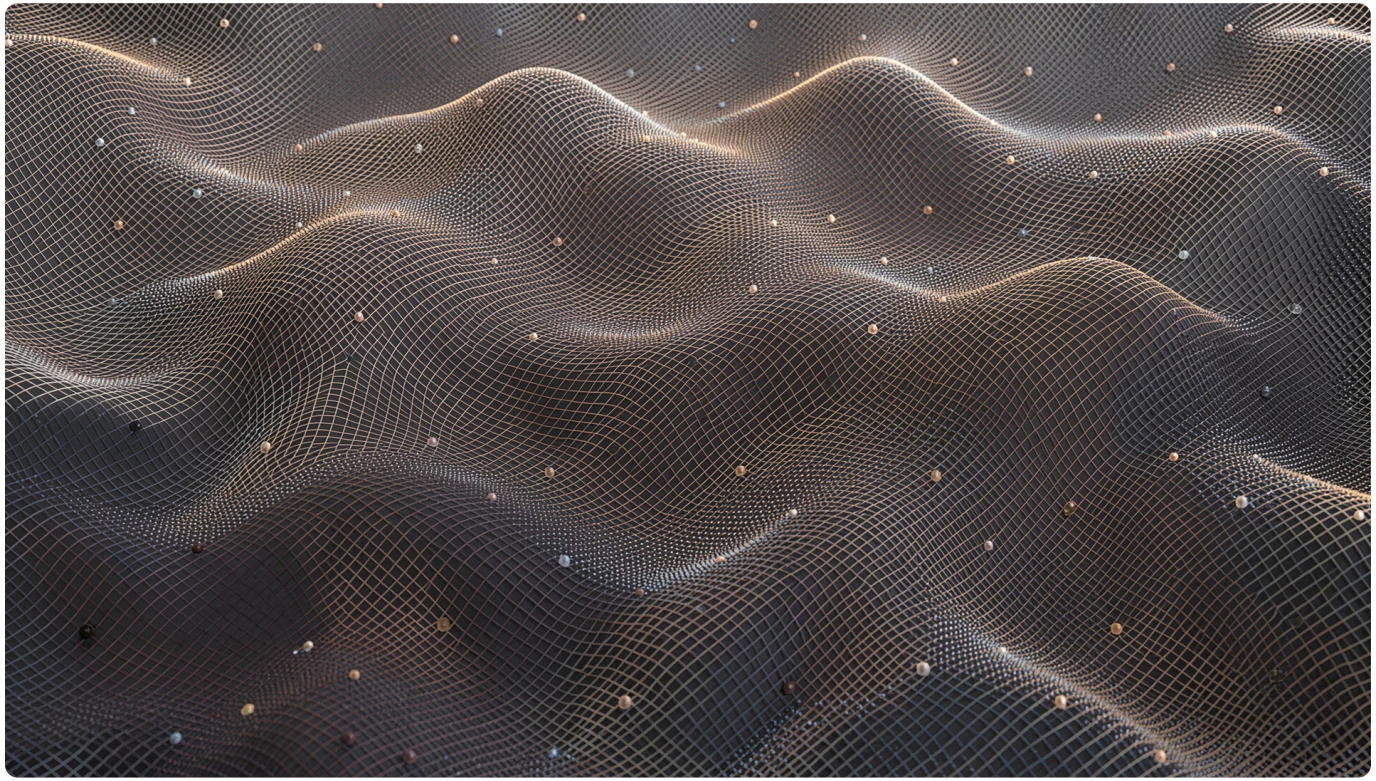
emphahrs@unisa.ac.za

ORCID: https://orcid.org/0000-0002-9917-7089

The Perplexing Problem

You wrote every word yourself — no ChatGPT, no shortcuts. Yet Turnitin flags your work as AI-generated. This isn't a rare glitch. It's a growing, well-documented phenomenon affecting students worldwide, and understanding *why* it happens is the first step to protecting yourself.

How Turnitin's AI Detection Works



Turnitin does **not** detect authorship — it estimates the *probability* that a piece of text resembles AI-generated output. The system relies on statistical heuristics and pattern matching, comparing your writing against known distributions of human and AI text. It produces a likelihood score, not a verdict.

⚠ A high score means "statistically similar to AI," not "definitely written by AI."

The "AI-Like" Fingerprints Detectors Look For

Stylistic Regularity

Consistent sentence structures and predictable transitions like *"Moreover"* or *"Furthermore"* mimic AI rhythm.

Low Burstiness

Evenly paced sentences without natural variation — short, punchy lines mixed with longer ones — read as machine-like.

Low Perplexity

AI chooses statistically "safe," predictable words. Formal academic prose often does too, confusing the detector.

Syntactic Uniformity

Too-regular punctuation patterns and low syntactic variety can register as algorithmically generated.

Use of Generic or Common Phrasing

Phrases like:

- “This essay will discuss…”
- “In today’s society…”
- “It is important to note that…”

These are common in both AI and human writing—but detectors may associate them with AI due to frequency.

Heavy Paraphrasing or Editing Tools

Even if you didn’t use AI to *write*:

- Grammar checkers
- Paraphrasing tools
- Translation tools

can make your text more “AI-like.”

Why Your Human Writing Might Get Flagged

Generic Prose

Standard academic structure clear topic sentences, predictable transitions closely mirrors AI output patterns.

Polished Formality

Well-organized, concise writing reads as machine-like precisely because it is so deliberate and smooth.

Heavy Editing

Extensive revisions iron out the rough edges that signal authentic human authorship to detection models.

The cruel irony: the better you write — the more polished, structured, and precise the more likely a detector is to question whether a human wrote it at all.

Who Is Most at Risk?

English Language Learners

Formal English learned through structured academic methods closely resembles the predictable patterns AI detectors flag.

Neurodivergent Students

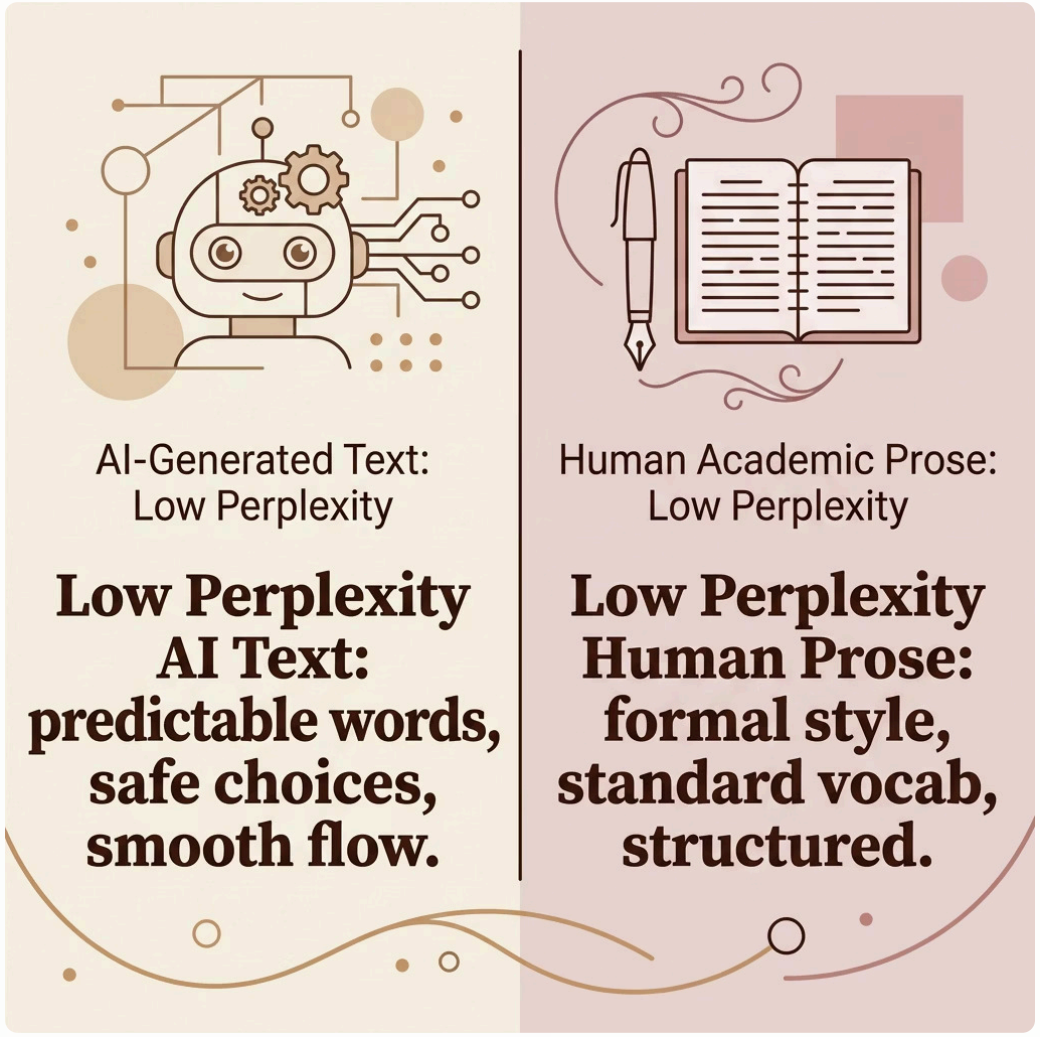
Reliance on repeated phrases and structured frameworks for clarity can inadvertently trigger detection algorithms.

Racial Disparities

Turnitin's AI detection:

- Is **not fully reliable**
- Can **disproportionately affect certain groups**
- May **indirectly create racial/ethnic disparities** via language and style patterns.

The Perplexity Problem



AI detection tools measure how **predictable** each word choice is — a metric called *perplexity*. AI models naturally gravitate toward statistically likely words, producing low-perplexity text.

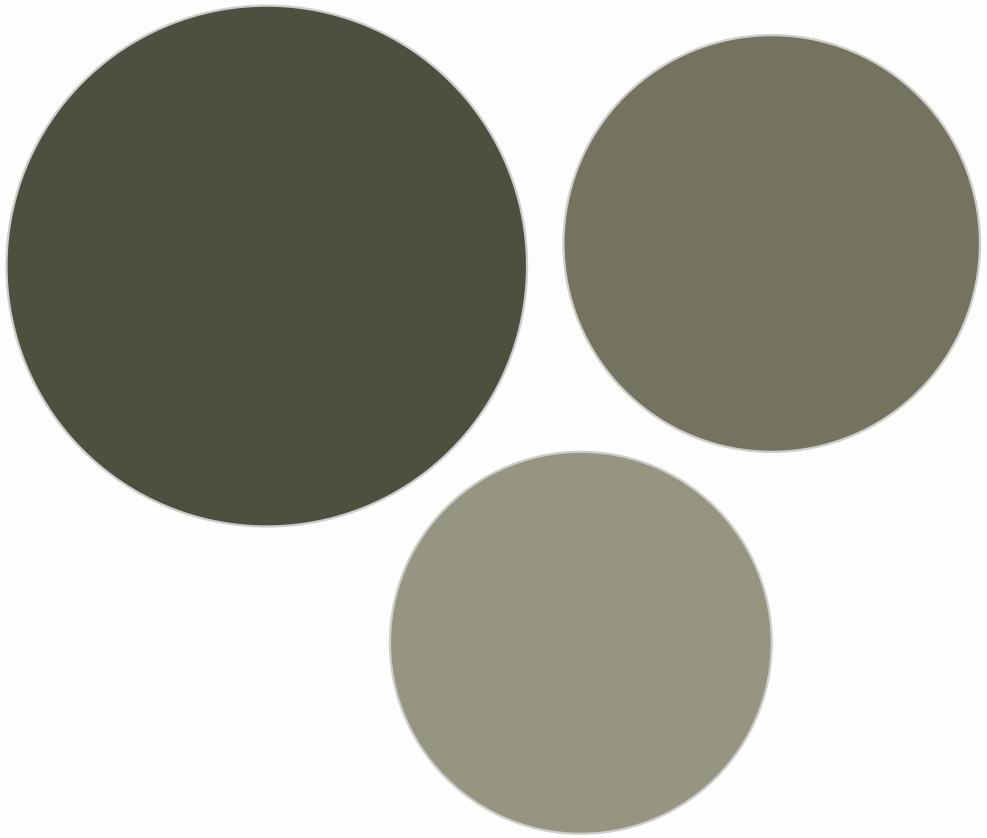
The problem? Students trained in formal academic writing do the exact same thing. Standard vocabulary, disciplinary conventions, and structured argumentation all reduce perplexity — making genuine human work indistinguishable from AI output to the algorithm.

Short Texts & The Fragmentation Effect

Why Short Samples Fail

Detection accuracy depends on volume. Short passages, abstracts, and introductions lack the stylistic breadth needed for reliable analysis. With fewer data points, the model fills gaps with assumptions nand those assumptions often skew toward AI.

⚠ Chunk-level scoring can fragment a longer document, penalizing sections that are locally concise even when the full piece is clearly human.



Beyond the Score: Context and Judgment

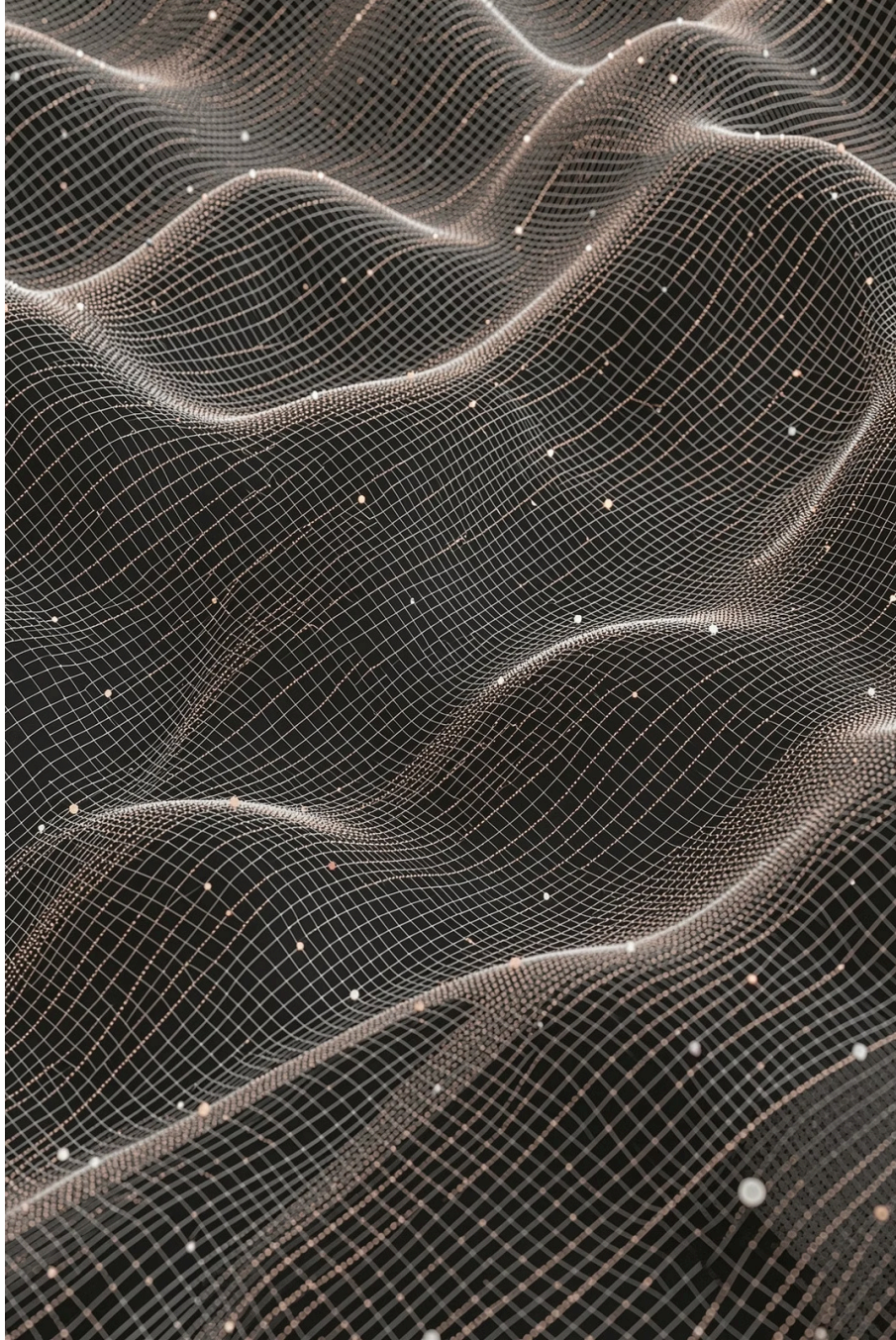
A Turnitin AI score is a **probabilistic estimate** — not a verdict, not proof, and not infallible. Treating it as definitive can have serious consequences for students who did nothing wrong.

Use it as one signal

Educators should weigh AI scores alongside writing history, draft submissions, and in-person knowledge of the student.

Conversation over accusation

A direct, supportive discussion about the writing process is far more reliable and far more fair than any algorithm.



Protecting Your Work

Know the Limits

Understand that no AI detector is definitive. False positives are documented and acknowledged by Turnitin itself.

Develop Your Voice

Incorporate personal insights, varied sentence rhythm, and anecdotes that are uniquely yours and hard to replicate.

Document Your Process

Save drafts, notes, and outlines. A clear writing trail is your strongest defense if a flag occurs.

Advocate for Fairness

Push for institutional policies that treat AI detection as one tool among many — never the sole basis for academic consequences.