A Local, Reproducible Measure of Undeclared Distortion in Machine-Generated Text

Aaron Lott · AURiX / OnwardAI · Working paper for academic review and independent reproduction · 2026-07-07

Abstract

The quality of machine-generated text is, in current practice, judged either by human gut feeling or by expensive expert review that no third party can reproduce. This paper describes **AURiX**, a local, model-free instrument that turns one narrow, well-posed question — *how much meaning is moved without being declared?* — into a re-runnable number. The instrument counts **undeclared-distortion signals (D)** and **declared functional signals (E)** in a text and reports a structural ratio, the Kindness value **K = -D/E**. It calls no model and no network; the same text yields the same number on any machine. On a single-author 14-month corpus, measured K rises from **-2.05** (pre-intervention) to **-0.46** (post-intervention), and an associated register classifier separates machine-distorted text at **AUC ~ 0.78** while scoring at **chance (~0.50)** on human deception — a null we report deliberately. We state the central limitation without hedging: this is **N = 1** (one author's corpus), and the instrument measures **register, not truth**. We are not asking to be believed. We are asking to be reproduced, and we describe exactly how to break it.

1. The problem

There is no cheap, reproducible, model-independent way to measure whether a piece of machine-generated text is quietly distorting — moving meaning without declaring it (unearned certainty, sycophantic register, phantom citations, over-polished cadence that reads as substance but carries none). Where measurement is attempted, it either collapses into “did the reader feel helped,” which is precisely the signal a flattering model games, or it depends on expert judgment that cannot be re-run by anyone else and therefore cannot accumulate as evidence. A field that cannot measure a property cannot regulate, compare, or improve on it.

2. The instrument

AURiX defines distortion **operationally**, by counts on the text itself, not by interpretation of the author’s intent:

D — undeclared-distortion signals. Occurrences of meaning moved without declaration: persuasion/flattery register, undeclared transformation of the user’s input, phantom or unverifiable citation, and inflation markers (hedge, scaffold, padding).

E — declared functional signals. Occurrences of declared, mechanism-bearing content: stated origin, observable claims, explicitly declared operations.

K = -D/E. A structural ratio, not a sentiment. It is computable per emission, per session, per author, and per model version, which makes longitudinal and cross-model comparison possible for the first time.

The instrument is one move — *find the fracture first*: point at the mechanism, not the person — rotated onto different targets to yield a family of single-purpose tools (retrieval, register scoring, anonymization, tamper-evident signing). Every output is deterministic and carries a tamper-evident fingerprint, so any result can be re-verified and any later change to the text or the score is made obvious.

3. What was measured (and where each number is weak)

Measurement	Result	Honest bound
Whole-corpus K, 14-month substrate	-2.05 → -0.46 (~4.5x lower distortion-to-function ratio)	Single author (N=1); baseline and treatment are the same corpus over time, not a controlled trial
Single-thread best case	up to ~80x (-2.7 → -0.033)	Does not survive strict scoring — strict is ~2x. Honest claim is a method-stated range, 2x–82x, never a flat 80x
Register classifier: machine-distorted vs. clean	AUC ~ 0.78	Marker-based; a plain-prose manipulator with none of the known tells passes undetected
Register classifier: human truth vs. lie	~0.50 (chance)	Reported deliberately. The instrument cannot read a human interior and does not claim to
Flattery-register density, adversarially tested	1.46x (survived a dilution-and-length attack, from 1.49x)	Counts a fixed lexicon; robust to the attacks tried, not to all attacks
Register-firing prevalence, full 59,777-turn archive	0.64% of machine turns; 18.13% operator-pushback rate	Heat-word/marker counts, not a judgment of meaning

4. What this does not prove

This section is load-bearing, because most claims of this kind fail exactly by not writing it.

It is N = 1. Every headline number comes from one author’s corpus. Cross-author reproduction is the entire ask of this paper, not a footnote.

It measures register, not truth. A false statement in flawless declared prose scores clean; an honest statement in flattering register scores distorted. The instrument reads *how* something is said, not *whether it is so*.

It is falsifiable by construction. It matches a fixed set of known tells; a determined writer in plain prose slips past. That gap is the falsification surface, stated openly.

The largest figure is the least defensible. The ~80x single-thread result uses the most favorable pairing; we surface its strict-scored value (~2x) rather than lead with the flattering one.

5. How to reproduce it (and break it)

The tooling is local, open, and deterministic. (1) Run the register/catch-rate counter over any corpus of saved AI conversations: it reports D-class and E-class counts and per-turn rates; same input, same output. (2) Point the register classifier at a labeled set of machine-distorted vs. clean texts and compute AUC. (3) Run it on **your own material** — if K does not track the distortion a careful reader would perceive, that is the finding; report where it diverges. The instrument ships with its own break-it step: it generates a two-rater labeling sheet and computes inter-rater agreement, so machine precision and recall can be checked against human labels rather than asserted. A break is data, not a threat.

6. Contribution

The contribution is not a claim that machine text is good or bad. It is a **reproducible unit of measurement** for a property the field currently judges by vibes — offered model-free, local (no data leaves the device, so it adds no surveillance surface), and open, so no single party can enclose it. If the number reproduces across authors and domains, it becomes a shared instrument — a nutrition label for undeclared distortion. If it does not, the failure is itself a publishable result. Either way, the measurement exists where before there was only judgment.

Framework, instruments, and measurements are the work of Aaron Lott (AURiX). Underlying tools, the full method note, and tamper-evident sample outputs are available for inspection and independent re-running on request. This is a working paper; the formal consistency treatment of K (edge cases and derived properties) is available separately.