Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Algorithmic Penalty for Non-Native Voices: A Three-Arm Diagnostic Accuracy Audit of Five Commercial AI Text Detectors Against Pre-LLM Scientific Literature — Protocol for the AUDIT-AI Study

Adnan Agha

adnanagha@uaeu.ac.ae

United Arab Emirates University

Eram Anwar

United Arab Emirates University

Research Article

Keywords: AI text detection, academic integrity, algorithmic bias, non-native English, diagnostic accuracy, research integrity, large language models, equity in publishing

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10164745/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Commercial AI text detectors are increasingly used by journals, universities, and grant reviewers to screen scientific writing for AI authorship. A foundational 2023 study reported a 61.3% mean false-positive rate across seven detectors on essays by non-native English speakers, versus near-perfect accuracy on US-authored essays, implying that perplexity-based detectors may penalise constrained linguistic expression rather than detect AI authorship. Whether this bias generalises to long-form peer-reviewed medical literature, and whether the corresponding-author institution's origin predicts misclassification, remains unquantified. This pre-registered, three-arm diagnostic accuracy study (200 texts; STARD-AI) contrasts 50 top-cited cardiometabolic papers from native-English-speaking institutions (Arm A) with 50 from non-native institutions (Arm B), stratified across two epochs (1995–2007, 2008–2020), against 100 Claude Sonnet 4.6 rewrites from a decoupled, air-gapped pipeline (Arm C). Introduction and Discussion sections are each scored twice across five detectors (iThenticate, Originality.AI, GPTZero, Copyleaks, Winston AI), yielding 4,000 observations. The primary outcome is the continuous AI-probability score, analysed with a linear mixed-effects model; the primary contrast, Arm B minus Arm A, is tested one-sided. The design provides approximately 80% power for a medium effect (d=0.5). The study involves no human participants and was determined non-human-subjects research; all code/schema/de-identified outputs will be released under CC-BY-4.0.

Registration: Open Science Framework (DOI to be assigned at deposit).

INTRODUCTION

Background

Commercial AI text detectors have been rapidly absorbed into editorial, academic integrity, and admissions workflows since the public release of ChatGPT in November 2022. Vendors typically advertise false-positive rates (FPRs) of ≤ 1%. The empirical literature, however, has accumulated more than 80 peer-reviewed validation studies documenting FPRs that range from 0% under tightly calibrated thresholds to over 80% under default settings, with variability driven by detector choice, threshold, text genre, length, author first language (L1), and the language model being detected.[1,2,3,4,5] The reach of this technology has been amplified by the rapid uptake of large language models across healthcare education and research[6] and by the embedding of generative AI into the everyday writing tools that authors use.[7]

Prevalence-side urgency is real and growing. A 2026 longitudinal analysis of 7,251 articles in JAMA Network Open (January 2022 through March 2025) reported that the monthly proportion of articles classified as containing AI-generated text rose from 0.0% to 11.3% ($P < .001$), with the highest detected rates in invited commentary (6.7%), global health (8.6%), and medical education (5.6%).[8] As journals continue to release author-disclosure policies and reviewer guidelines, the practical question of which screening tool, applied to which text, on what threshold, has become editorial policy in real time.[9,10,11]

The cornerstone empirical study of detector bias is Liang et al. (Patterns, 2023), who evaluated seven commercial GPT detectors on two contrasting human-written corpora: US eighth-grade student essays and Test of English as a Foreign Language (TOEFL) essays written by non-native English speakers.[1] While detectors classified the US-authored essays with near-perfect accuracy, they incorrectly labelled more than half of the TOEFL essays as “AI-generated,” with a mean false-positive rate of 61.3%.[1] The same study provides a mechanism-consistent intervention: prompting ChatGPT to enhance the vocabulary of the TOEFL essays to “sound more like that of a native speaker” increased the perplexity of those essays and reduced their misclassification rate from 61.3% to 11.6%.[1] The mirror-image experiment also held: simplifying US eighth-grade essays increased the misclassification of those native-author essays as AI-generated.

The mechanistic explanation offered by Liang et al. and extended in follow-on work[12] is that perplexity, a measure of how predictable a sequence of words is under a reference language model, is itself proxy-correlated with linguistic proficiency rather than authorship source. Many commercial detectors operationalise "AI-likeness" as low perplexity (text the language model finds easy to predict) combined with low burstiness (low structural variance in sentence length and complexity), an approach exemplified by zero-shot methods such as DetectGPT that score the probability curvature of text under a reference model.[13] Non-native English writers, particularly those for whom English is a second or third language, are described in the linguistics literature as exhibiting reduced lexical diversity, simpler syntactic structures, and more formulaic constructions.[12,14] These properties make their writing more predictable under reference language models, and therefore lower-perplexity in exactly the direction that perplexity-threshold detectors interpret as "machine-generated." The bias is not malicious; it is structural to the detection methodology itself.

Prior work

This L1-linked false-positive bias has since replicated across detectors, populations, and study designs, from ESL graduate writing to international students' coursework, with consistent disparities reported for texts from underrepresented groups, and with even experienced linguists unable to reliably distinguish AI-generated from human academic prose.[14,15,16,17,18] In the medical domain specifically, validation studies on controlled pre-large-language-model corpora report detector error rates and reliability that vary widely across tools, thresholds, and specialties, too unstable to serve as standalone evidence in high-stakes adjudication.[19,20,21,22,23,24,25] In response, several universities have disabled commercial AI detection, the US National Institutes of Health (NIH) has banned generative AI in grant peer review, and the US Department of Education has framed disparate detector error rates as a civil-rights concern; a structured account of this evidence base, with the specialty-by-specialty false-positive figures and the full institutional and policy citations, is provided in online supplemental file 1.[26,27,28,29,30,31]

The unaddressed gap

Despite this empirical and policy ferment, no peer-reviewed study to date has tested whether the geographic origin of the corresponding-author institution, the practical proxy used by journal editors and reviewers, predicts detector misclassification on long-form scientific writing in a single specialty. Existing work tests student essays, abstracts of fewer than 300 words, or mixed-genre corpora that conflate length, register, and topic. Existing work also relies overwhelmingly on the Liang TOEFL framing of L1 versus L2 within an educational context, rather than the editorial-screening context that governs grant peer review and journal submissions. This is the gap the AUDIT-AI Study addresses, reframing detector misclassification not as a technical performance metric but as a question of accountability and fairness in the tools that increasingly govern who is trusted to have authored their own scientific work.

Objectives and hypotheses

The AUDIT-AI Study has five locked, pre-registered hypotheses.

H1 (primary, directional). The Mean AI-probability scores will be significantly higher for cardiometabolic papers from non-native institution authors (Arm B) than for native institution authors (Arm A) across the detector ensemble, with an absolute difference of at least 15 percentage points.

H2. No detector achieves sensitivity of at least 0.85 with a specificity of at least 0.85 simultaneously across both human arms.

H3. Inter-detector agreement at the 50% binary classification threshold was poor (Fleiss' $\kappa < 0.40$).

H4. The L1 disparity is larger for the Discussion than for the Introduction sections, reflecting the more stylistically expressive narrative of the Discussion versus the more structured background-and-rationale form of the Introduction.

H5. Author-institution geographic origin remains a significant predictor of AI-probability score after controlling for journal impact factor, publication year/epoch, word count, and Flesch-Kincaid grade level.

The five hypotheses are locked at the OSF pre-registration deposit. Any subsequent changes are documented as registered amendments.

METHODS

Study design

This is a three-arm cross-sectional diagnostic accuracy and discrimination study (See Figure for study design). Reporting will follow the Standards for Reporting of Diagnostic Accuracy Studies for AI (STARD-AI 2025) guidelines (Supplementary File 1). The study is considered as Non-Human Subjects Research (NHSR) after review (Supplemental File 2).

Corpus construction (n = 200 texts)

Arm A (native-institution human authors, n = 50). Cardiometabolic peer-reviewed papers indexed in MEDLINE under the Medical Subject Headings (MeSH) terms "Diabetes Mellitus", "Obesity", "Non-alcoholic Fatty Liver Disease", and "Cardiovascular Diseases", published between 1995 and 2020 inclusive, in English, with both first and corresponding-author institutional affiliations in the United States, United Kingdom, Canada, Australia, New Zealand, or Ireland. Top-cited papers (Citation Classics methodology) are selected from Scopus and Web of Science within each of two epochs (1995–2007 and 2008–2020), with 25 papers per epoch.

Arm B (non-native institution human authors, n = 50). The same eligibility window, MeSH terms, and language requirement, but with first and corresponding author institutional affiliations in the Middle East and North Africa, East Asia, South Asia, Latin America, or non-English-speaking Europe. Same epoch stratification (25 papers per epoch).

Arm C (Claude Sonnet 4.6 rewrites, n = 100). One AI rewrite of each Arm A and Arm B paper, generated through a decoupled three-stage pipeline (described below). The Claude LLM used is an object of study and only a corpus-generation instrument, not a writing aid for the manuscript text.

Exclusion criteria for both the human arms: papers with mixed first/corresponding-author affiliations spanning native and non-native categories; multinational consortium papers without identifiable lead institution; and non-English language publications. Papers with a declared professional English-editing service in the acknowledgements are retained as a sensitivity-analysis stratifier rather than excluded.

Geographic classification. Two independent raters classify every paper's geographic stratum using first and corresponding-author institutional affiliation. Disagreements are resolved by a third blinded reviewer. Inter-rater agreement is reported as Cohen's κ; target at least 0.80.

Sections analysed (hybrid design). Each human paper contributes two narrative sections (Introduction and Discussion) for detector scoring. Abstract, Methods, and Results sections are extracted as metadata but are never submitted to detectors. This hybrid design is justified on three grounds: (1) Methods and Results are stylistically dominated by low-perplexity technical boilerplate that elevates baseline detector FPR equally for native and non-native human writing, compressing the L1 signal toward null; (2) in real-world adversarial AI use in scientific publishing, AI is overwhelmingly deployed for narrative synthesis (Introduction and Discussion), not raw data reporting; and (3) abstracts of 250–350 words sit at or below detector reliability thresholds (iThenticate requires at least 150 words; Originality recommends at least 300; GPTZero is unreliable below approximately 200 words).

Arm C generation pipeline (locked at v1.3)

The Arm C pipeline is a three-stage decoupled architecture implemented as Python scripts (extract_packet.py, generate_arm_c.py, qa_gate.py) and a JSON Schema v1.2. All source code, schema, and audit logs are deposited on the Open Science Framework (OSF) before data collection.

Stage 1, Extraction

Each source paper PDF is parsed by LlamaParse (vision-language-model OCR via the LlamaCloud API). LlamaParse is chosen over deterministic parsers, such as pdfplumber, because complex medical-trial tables, with merged headers, sub-categorised rows, and compound cells such as “12.4 (±2.1)” or “n = 450 (22%)”, require visual context-awareness for reliable row and column integrity. The extracted Markdown is assembled into a JSON v1.2 packet containing: (1) the abstract with the Conclusion sentence stripped to prevent stylometric leakage; (2) main paper tables verbatim, capped at 4,000 tokens; (3) figure captions sterilised via Claude Haiku 4.5 to remove evaluative adjectives (such as “striking,” “robust,” “marked”), a narrow rule-based transformation, not generative writing; (4) supplementary appendix tables whitelist-filtered (baseline characteristics, primary and secondary outcomes, safety tables only), capped at 8,000 tokens; and (5) the reference list random-sampled to 30 entries with a fixed seed for reproducibility. The total prompt budget is constrained to at most 21,500 tokens. The use of LlamaParse for extraction and Claude Sonnet 4.6 for generation establishes a methodological air-gap: a reviewer cannot argue that Claude is regurgitating pre-2021 papers from its training data because the structured packet that drives Stage 2 generation was produced by a different vendor’s vision pipeline.

Stage 2, Generation

Claude Sonnet 4.6 (Anthropic API, model string claude-sonnet-4-6, temperature 0.7, top_p 0.9, max_tokens 3000 per section) generates two narrative sections per paper: Introduction and Discussion. The system prompt enforces a stylometric override (interleaved short and long sentences for maximum burstiness; banned LLM stock transitions such as “Furthermore”, “Moreover”, “In conclusion”; structural asymmetry; varied vocabulary for maximum perplexity) and a grounding constraint (cite only from the provided reference list; do not invent statistics, citations, or clinical claims). Target word counts are jittered ±5% per paper around design-typical means and are tracked in metadata.

Stage 3, Quality assurance gate

Each generated section is compared against its target word count ±15% tolerance band using a deterministic prose-only word counter that strips inline citations, statistical reporting blocks, section headers, and figure or table

references before counting. Out-of-range sections trigger up to three regeneration attempts with corrective directives. All attempts are logged with response identifiers in a structured audit log.

Detector ensemble (locked)

Five widely deployed commercial AI text detectors form the locked ensemble (Table 1). Each of the 400 sections (200 papers × 2 sections) is scored twice on different days within a single 14-day scoring window, yielding 4,000 total detection observations (200 texts × 2 sections × 5 detectors × 2 retests). Detector versions and dates of API access are frozen at the start of the scoring window and recorded for every submission. Repository submission is disabled for iThenticate to prevent corpus contamination of the UAEU institutional similarity database.

#	Detector	Access route	API / Web
1	Turnitin / iThenticate	UAEU institutional licence	Web upload (AI Writing Detection module)
2	Originality.AI	Direct paid API (monthly tier)	REST API
3	GPTZero	Direct paid API	REST API
4	Copyleaks	Direct paid API	REST API
5	Winston AI	Direct paid API	REST API

Table 1: Locked five-detector commercial AI text detection ensemble

Outcomes

Primary outcome. Continuous AI-probability score (0–100%) per text × detector × section × retest.

Secondary outcomes. Binary classification at the 50% AI-probability threshold (with sensitivity analyses at 25% and 75%); receiver operating characteristic (ROC) curve and AUC per detector for distinguishing Arm C (AI) from Arms A + B (human); Fleiss' κ inter-detector agreement at binary thresholds; intraclass correlation coefficient (ICC (2, k)) test–retest reliability per detector at the continuous level; Section × Arm interaction (the section moderator); and a word-weighted document-level AI percentage, derived as the sum of flagged words across Introduction and Discussion divided by the sum of total words across the two sections, multiplied by 100. The study is registered on open science framework (OSF) platform (Supplementary File 3), along with workbooks for iThenticate and Master Workbook for analytical tracking (Supplementary Files 4 and 5)

Sample size and power

Anchor estimates were derived from PubMed-verified medical domain detector studies. Native-author false-positive rates cluster at approximately 9% (SD 6%) across detectors (Habibzadeh approximately 10%; Pan 5.8% to 18.6% across GPTZero, Originality, and Copyleaks; Bohler 8.6%; Howard approximately 0% after retrospective threshold tuning).[19,20,22,24] No medical-domain estimate exists for non-native-institution authors, the gap this study addresses, so the non-native rate is treated as an assumed target informed by the educational-setting literature (Liang 61.3%; Lege 69.9% on ESL writing).[1,14]

Because the unit of inference for the primary Arm B minus Arm A contrast is the paper (50 per arm), with each paper scored 20 times (5 detectors × 2 sections × 2 retests) so that per-paper mean scores carry little residual measurement error, power for the primary linear mixed-effects contrast is well approximated by a two-sample comparison of 50 papers per arm. Because H1 is directional and locked before data collection, the primary contrast is tested one-sided at $\alpha = 0.05$. At n = 50 per arm, this provides approximately 80% power for a medium standardised effect (Cohen's d = 0.5) and greater than 99% power for the large effect (d = 0.8 to 1.0) implied by the foundational first-language literature; power for a small effect (d = 0.3) is approximately 44%, and the minimum effect detectable at 80% power is $d \approx 0.50$. Under a more conservative two-sided test, the corresponding figures are approximately 70% (d = 0.5) and greater than 97% (d = 0.8), and 80% power at d = 0.5 would require 64 papers per arm. For the secondary binary false-positive-rate comparison (two-proportion test, $\alpha = 0.05$, two-sided) against a verified native-author rate of approximately 10%, n = 50 per arm provides approximately 73% and 88% power to detect non-native rates of 30% and 35% respectively. The pre-registered contingency to inflate to n = 60 per arm is retained, triggered by a high pilot pooled standard deviation.

Pilot phase

A five-paper-per-arm pilot (15 texts × 5 detectors × 2 sections × 2 retests = 300 observations) is conducted before full data collection. Pre-registered contingency: if the observed pooled standard deviation exceeds 25, the sample size is inflated to n = 60 per group; if the Arm B minus Arm A delta exceeds 20 percentage points in the pilot, the planned n = 50 per group is retained. The pilot decision is registered on OSF regardless of the outcome.

Statistical analysis plan (locked)

Primary model. Linear mixed-effects regression: AI_score ~ Arm + Section + Epoch + IF_tertile + WordCount_z + (1 | paper_id) + (1 | detector), where Arm is a three-level factor (native human / non-native human / AI), Section is a two-level factor (Introduction / Discussion), and Epoch is a two-level factor (1995–2007 / 2008–2020). The single pre-specified primary contrast for H1 is Arm B minus Arm A; because H1 is directional, it is tested one-sided ($\alpha = 0.05$) against the alternative that Arm B exceeds Arm A, with an effect in the opposite direction reported as non-confirmatory rather than as supporting H1. As a single pre-planned contrast, it carries no multiplicity adjustment; all other pairwise comparisons, including those involving the AI arm, are secondary, two-sided, and Tukey-adjusted. Estimation uses restricted maximum likelihood with Kenward-Roger degrees of freedom; Satterthwaite is used as a sensitivity check.

Secondary models. Model 2 (H4 section moderator) adds an Arm × Section interaction term. Model 3 (H5 adjusted geographic effect) is fitted within Arms A and B only, with controls for impact factor tertile, epoch, word count, Flesch-Kincaid grade level, and declared English-editing-service flag. A document-level outcome model mirrors the primary model with the section removed and one record per paper × detector × retest combination.

Other pre-specified analyses. Two-proportion z-test with Wilson 95% confidence intervals for the binary FPR comparison; DeLong's test for differences between detector AUCs; Fleiss' κ across the five detectors at each binary threshold; bootstrap (10,000 iterations) for non-parametric confidence intervals on FPR; a concordance sub-study (n = 20) in which 10 Arm A and 10 Arm B papers are additionally submitted to all five detectors as a single combined two-section document (Introduction + Discussion concatenated) and compared with the derived word-weighted document AI percentage using Bland-Altman analysis, Lin's concordance correlation coefficient, and ICC(2,1); and a sentence-level exploratory analysis (Originality, GPTZero, Copyleaks only) of the proportion of words within each section in sentences scored at or above 85% AI probability (descriptive only, not used in primary inference).

Pre-specified subgroups and sensitivities. Regions within Arm B (MENA, East Asia, South Asia, Latin America, non-English Europe); section type; epoch; impact factor tertile; declared English-editing service; and threshold variations (25%, 50%, 75%).

Quality control and stopping rules

- Inter-rater geographic classification reliability: Cohen's κ of at least 0.80; if the pilot returns κ < 0.70, a third independent rater is added before scaling.
- Test–retest detector reliability: ICC(2,k) of at least 0.85 per detector. An ICC < 0.70 for any detector triggers either a third scoring day or pre-registered exclusion of that detector from the primary analysis.
- QA-gate convergence: Any section that fails to converge to the ±15% target after three regenerations is accepted with the closest output and flagged. A non-convergence rate of more than 5% triggers a registered amendment.
- Vendor-account integrity: If any detector vendor terminates the paid account during scoring, partial data are reported transparently; no mid-study substitution is made without a registered amendment.

Data management and reproducibility

All artefacts are deposited under a Creative Commons Attribution 4.0 (CC-BY 4.0) licence at the time of publication of the results manuscript: JSON Schema v1.2 (locked at pre-registration); pipeline code (extract_packet.py, generate_arm_c.py, qa_gate.py, prose_word_count.py); all 100 data packets in data_packets/*.json with human_qa_passed=true; all 200 generated section Markdown files plus per-section metadata audit logs; de-identified detector output scores (paper-level identifiers retained as P001–P100, author names suppressed); sterilisation logs and QA-gate decision logs; and a reproducible analysis pipeline (R + Quarto + Git). Cryptographic hashes (SHA-256) of all source PDFs are deposited at pre-registration; the PDFs themselves are not redistributed due to publisher copyright.

Anticipated findings

AUDIT-AI is designed to produce the first specialty-locked, pre-LLM, three-arm test of whether the geographic origin of the corresponding-author institution predicts commercial AI text detector misclassification in long-form cardiometabolic research literature. If the primary hypothesis is supported, the study will provide direct, peer-reviewable evidence that journal and grant-review reliance on commercial detector scores carries a measurable equity cost for biomedical authors based at non-native-institution sites, a population that is substantial and growing globally. If the primary hypothesis is not supported, the study will produce the first negative finding in a structured, pre-registered, medical-domain test of the L1 disparity, with implications for how the Liang TOEFL-essay result generalises beyond educational settings into peer-reviewed scientific publishing.

For **H1**, based on the literature anchors, we anticipate a mean Arm B minus Arm A AI-probability score difference of approximately 12 to 22 percentage points across the detector ensemble, statistically significant at α = 0.05 (Tukey-adjusted) with greater than 90% power if the true effect size d exceeds approximately 0.50. For **H2**, we anticipate that no single detector will achieve both sensitivity of at least 0.85 and specificity of at least 0.85 simultaneously across both human arms; we expect Originality.AI and Copyleaks to achieve the lowest FPRs on Arm A (consistent with prior reports of low false-positive rates on native-author medical and computing texts[2,20,22,32]) and the most adversarial tools to produce elevated false-positive rates on Arm B, consistent with the low reliability and wide inter-tool variability

documented in prior medical-domain evaluations.[21,23] For **H3**, we anticipate Fleiss' $\kappa < 0.40$ across the five detectors at the 50% binary threshold. For **H4**, we anticipate that the Arm B minus Arm A disparity will be larger in the Discussion sections than in the Introduction sections, consistent with the mechanism that L1-linked stylometric uniformity is more detectable in expressive narratives than in formulaic background prose. For **H5**, we anticipate that geographic institutional origin will remain a significant predictor after adjustment for impact factor, epoch, word count, and Flesch-Kincaid grade level. We will report all five hypotheses regardless of the outcome and clearly distinguish pre-registered from exploratory analyses in the results manuscript.

Relation to prior work and novelty

AUDIT-AI extends the existing literature in five concrete ways. **(1) Long-form pre-LLM medical text.** Existing replications of the Liang TOEFL finding[1] have been conducted predominantly on student essays of 250–500 words or short abstracts of 200–300 words; AUDIT-AI tests the detector behaviour on the long-form (typically 600–1,500 words per section) narrative prose that constitutes the actual editorial-screening target, a different perplexity sampling regime than short-text tests. **(2) Geographic-institutional proxy.** Existing studies have predominantly used self-reported L1 status, English-language-learner status, or TOEFL-essay framing; AUDIT-AI uses the editor's practical proxy of first and corresponding-author institutional affiliation, which corresponds more directly to the signal a journal editor sees on submission. **(3) Cardiometabolic specialty lock.** Existing medical-domain validation studies[19,20,21,22,23,24,25] have spanned multiple specialties, conflating genre, vocabulary register, and topic distribution with the L1 disparity; AUDIT-AI restricts to a single specialty (diabetes, obesity, MASLD, cardiovascular disease) with well-characterised landmark trials. **(4) Methodological air-gap for the AI arm.** Existing AI-comparator studies have typically used the same vendor for extraction and generation, exposing them to the training-data-contamination critique; AUDIT-AI decouples the two. **(5) Pre-registered, locked design.** All hypotheses, sample sizes, statistical model, threshold choices, and pre-specified subgroups are locked at OSF deposit before any data collection, addressing the threshold-cherry-picking critique documented in retrospective tuning studies.[22]

Implications and future directions

Independent of the primary outcome, AUDIT-AI will produce the cleanest pre-registered, specialty-locked test of commercial detector reliability and fairness on pre-LLM medical literature to date, informing three accountability decisions

Journal policy: whether journals should continue to deploy commercial detectors as a soft screening tool when the corresponding author institution is non-native, and whether detector-based flagging should trigger additional editorial scrutiny only, rather than being treated as evidence in misconduct adjudication.

Grant peer review: The NIH July 2023 ban on generative AI in peer review explicitly cited accuracy concerns;[29] AUDIT-AI's results will inform whether similar concerns apply to detector-based screening of grant applications.

Academic integrity adjudication: Several universities (Vanderbilt, NYU, University of British Columbia, University of Cape Town) have already disabled Turnitin AI detection in coursework systems,[26,27,28,30] but the biomedical research equivalent, research integrity committee reliance on detector scores in misconduct investigations, has not yet been systematically addressed.

A planned Phase 2 follow-up study (registered at OSF as part of this pre-registration) will use the 100 human-source paper titles and 100 AI-generated alternate paper titles from this study's Arm C pipeline as the corpus for a separate publishable enquiry: a Likert-scale rater study examining whether non-endocrinologist clinical readers can distinguish, and whether they prefer, AI-generated versus human-written titles for cardiometabolic landmark papers. The Phase 2 study has distinct hypotheses, methodology, and ethics review pathway; the framework is pre-registered here to lock the corpus for the eventual follow-up.

Abbreviations

AI
artificial intelligence
AUC
area under the receiver operating characteristic curve
CC-BY
Creative Commons Attribution
ELL
English language learner
ESL
English as a second language
FPR
false-positive rate
ICC
intraclass correlation coefficient
ICMJE
International Committee of Medical Journal Editors
L1
first language
LLM
large language model
MASLD
metabolic dysfunction-associated steatotic liver disease
MeSH
Medical Subject Headings
NHSR
Non-Human Subjects Research
NIH
National Institutes of Health
OSF
Open Science Framework
ROC
receiver operating characteristic
STARD-AI
Standards for Reporting of Diagnostic Accuracy Studies for AI
TOEFL
Test of English as a Foreign Language
UAEU

United Arab Emirates University.

Declarations

ETHICS AND DISSEMINATION

Ethics Approval and Consent to Participate

The study uses only published, peer-reviewed papers and commercial detector outputs. There are no human participants, no animal subjects, no patient records, no biospecimens, and no identifiable private information, therefore no consent for participation or consent for publication required. The study has been determined to constitute Non-Human Subjects Research (NHSR) on May 5, 2026. The Human Research Ethics Committee Chair has confirmed that the project falls outside the committee's jurisdiction. The NHSR determination memo is provided as online supplemental file 2.

AI use disclosure

In line with ICMJE and World Association of Medical Editors guidance,[33] AI tools are used at the following points and will be disclosed in the results manuscript: (1) Claude Sonnet 4.6 was used in protocol drafting and refinement, with the principal investigator retaining final editorial control on all content; (2) Claude Sonnet 4.6 is the locked Arm C generator (deterministic, version-pinned); and (3) Claude Haiku 4.5 is the locked extraction-stage caption sterilisation tool. No AI tool participates in statistical analysis or hypothesis testing, and no AI tool meets ICMJE authorship criteria; none is listed as an author.

Dissemination

The findings will be disseminated through a peer-reviewed results manuscript and by depositing all code, schema, audit logs, and de-identified detector outputs on the Open Science Framework under a CC-BY 4.0 licence. The results will be reported in full, regardless of the direction of the findings, with pre-registered analyses clearly distinguished from exploratory analyses.

FUNDING

The authors did not receive any funding to conduct this study.

COMPETING INTERESTS

The principal investigator declares no financial relationships with any of the detector vendors (Turnitin/iThenticate, Originality.AI, GPTZero, Copyleaks, Winston AI), Anthropic (Claude Sonnet 4.6 and Haiku 4.5), OpenAI, or LlamaIndex/LlamaCloud. All commercial tools are accessed under standard paid or institutional terms.

DATA AVAILABILITY STATEMENT

All data, code, and audit logs produced by this study will be deposited on the Open Science Framework at the time of publication of the results manuscript under a CC-BY 4.0 licence. The cryptographic hashes of the 100 source PDFs are deposited at pre-registration; the PDFs themselves are not redistributed because of publisher copyright.

References

1. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns (N Y). 2023;4(7):100779. doi:10.1016/j.patter.2023.100779
2. Walters WH. The effectiveness of software designed to detect AI-generated writing: a comparison of 16 AI text detectors. Open Inf Sci. 2023;7(1):20220158. doi:10.1515/opis-2022-0158
3. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, Šigut P, Waddington L. Testing of detection tools for AI-generated text. Int J Educ Integr. 2023;19:26. doi:10.1007/s40979-023-00146-z
4. Pratama AR. The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. PeerJ Comput Sci. 2025;11:e2953. doi:10.7717/peerj-cs.2953
5. Pudasaini S, Miralles-Pechuán L, Lillis D, Llorens Salvador M. Survey on AI-generated plagiarism detection: the impact of large language models on academic integrity. J Acad Ethics. 2024. doi: 10.1007/s10805-024-09576-x
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887. doi:10.3390/healthcare11060887
7. Bordalejo B, Pafumi D, Onuh F, Roberts B, Vu T, Marpaung E. “Scarlet Cloak and the Forest Adventure”: a preliminary study of the impact of AI on commonly used writing tools. Int J Educ Technol High Educ. 2025;22:5. doi:10.1186/s41239-025-00505-5
8. Wolfrath N, Patel S, Flitcroft M, Banerjee A, Somai M, Crotty BH, Kothari AN. Rising prevalence of detected AI-generated text in medical literature: longitudinal analysis in open access articles. arXiv [preprint]. 2026. doi:10.48550/arXiv.2603.19316
9. Leung TI, de Azevedo Cardoso T, Mavragani A, Eysenbach G. Best practices for using AI tools as an author, peer reviewer, or editor. J Med Internet Res. 2023;25:e51584. doi:10.2196/51584
10. Elsevier. Generative AI policies for journals. 2024. https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals
11. Taylor & Francis. AI policy. 2024. https://taylorandfrancis.com/our-policies/ai-policy/
12. Schwartz O. AI-detectors biased against non-native English writers. Stanford Institute for Human-Centered AI. July 2023. https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers
13. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. DetectGPT: zero-shot machine-generated text detection using probability curvature. International Conference on Machine Learning. 2023. doi: 10.48550/arXiv.2301.11305
14. Lege R. Auditing the fairness of AI-detection tools: a comparative study of ESL, published, and AI-generated texts and their misclassification risks. Int J Teach Learn Educ. 2025;4(5):42–58. doi:10.22161/ijtle.4.5.5
15. Stowe K, Afanaseva S, Raimundo R, Sun Y, Patil K. Identifying bias in machine-generated text detection. arXiv [preprint]. 2025. doi:10.48550/arXiv.2512.09292
16. Basu P, Zhang Y, Raheja V. BAID: a benchmark for bias assessment of AI detectors. arXiv [preprint]. 2025. doi:10.48550/arXiv.2512.11505
17. Mathewson TG. AI detection tools falsely accuse international students of cheating. The Markup. 14 August 2023. https://themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating

18. Casal JE, Kessler M. Can linguists distinguish between ChatGPT/AI and human writing? A study of research ethics and academic publishing. Res Methods Appl Linguist. 2023;2(3):100068. doi:10.1016/j.rmal.2023.100068
19. Habibzadeh F. GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. J Korean Med Sci. 2023;38(38):e319. doi:10.3346/jkms.2023.38.e319
20. Pan ET, Florian-Rodriguez M. Human vs machine: identifying ChatGPT-generated abstracts in Gynecology and Urogynecology. Am J Obstet Gynecol. 2024;231(2):276.e1-276.e10. doi:10.1016/j.ajog.2024.04.045
21. Cooperman SR, Olaniyan A, Brandão RA. AI discernment in foot and ankle surgery research: a survey investigation. Foot Ankle Surg. 2024;31(3):214–219. doi:10.1016/j.fas.2024.10.001
22. Howard FM, Li A, Riffon MF, Garrett-Mayer E, Pearson AT. Characterizing the increase in artificial intelligence content detection in oncology scientific abstracts from 2021 to 2023. JCO Clin Cancer Inform. 2024;8:e2400077. doi:10.1200/CCI.24.00077
23. Cheng A, Lin Y, Reedy G, Joseph C, Wirkowski S, Mallette V, Nagesh V, Krieser D, Calhoun A. Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. Adv Simul (Lond). 2025;10(1):66. doi:10.1186/s41077-025-00396-6
24. Bohler F, Tran T, Burmeister JR, Hadid K, Joshi S, Gao S, Chaiyasate K. Detectable AI-generated text in the Journal of Craniofacial Surgery: a comparison of 2014 and 2024 publications. J Craniofac Surg. 2025;37(3–4):841–844. doi:10.1097/SCS.0000000000012366
25. Erol G, Ergen A, Gülsen Erol B, Kaya Ergen S, Bora TS, Çölgeçen AD, et al. Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. Acta Neurochir (Wien). 2025;167(1):214. doi:10.1007/s00701-025-06622-4
26. Coley CC. Guidance on AI detection and why we're disabling Turnitin's AI detector. Vanderbilt University Brightspace. 16 August 2023. https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector
27. New York University Office of the Provost. Disabling the AI tool in Turnitin. 7 September 2023. https://www.nyu.edu/content/dam/nyu/provost/documents/Disabling%20the%20AI%20Tool%20in%20Turnitin.pdf
28. Learning Technology Hub, University of British Columbia. UBC affirms decision to not enable Turnitin AI detection. 28 August 2023. https://lthub.ubc.ca/2023/08/28/ubc-affirms-decision-to-not-enable-turnitin-ai-detection/
29. National Institutes of Health. The use of generative artificial intelligence technologies is prohibited for the NIH peer review process. Notice NOT-OD-23-149. July 2023. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html
30. University of Cape Town Communications. UCT scraps flawed AI detectors. 24 July 2025. https://news.uct.ac.za/article/-2025-07-24-uct-scraps-flawed-ai-detectors
31. United States Department of Education, Office for Civil Rights. Equal educational opportunity for English learners. 2023.
32. Liu JQ, Lo CL, Chong KC, Tsang CC, Ho CY. Detecting artificial intelligence-generated content in rehabilitation research: a cross-sectional study. Int J Educ Integr. 2024;20:8. doi: 10.1007/s40979-024-00155-6
33. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. 2025. https://www.icmje.org/recommendations/
34. Shamsi A, Wang T, Amraei M, Raju NV. Evaluating AI text detection tools for distinguishing human-written from AI-generated abstracts in Persian-language journals of library and information science. Acta Inform Pragensia. 2026. doi:10.18267/j.aip.293
35. Sabet CJ, Bajaj SS, Stanford FC, Celi LA. Equity in scientific publishing: can artificial intelligence transform the peer review process? Mayo Clin Proc Digit Health. 2023;1(4):596–600. doi:10.1016/j.mcpdig.2023.10.002

36. Bossi Andreasen R. Can we trust artificial intelligence in scientific publishing? J Nucl Med. 2025;66(12):1880–1881. doi:10.2967/jnumed.125.271360
37. Open Science Framework. AUDIT-AI pre-registration deposit. 2026 [DOI to be assigned by OSF].
38. Open Science Framework. AUDIT-AI pipeline repository (schema, code, audit logs). 2026 [DOI to be assigned by OSF].

Figures

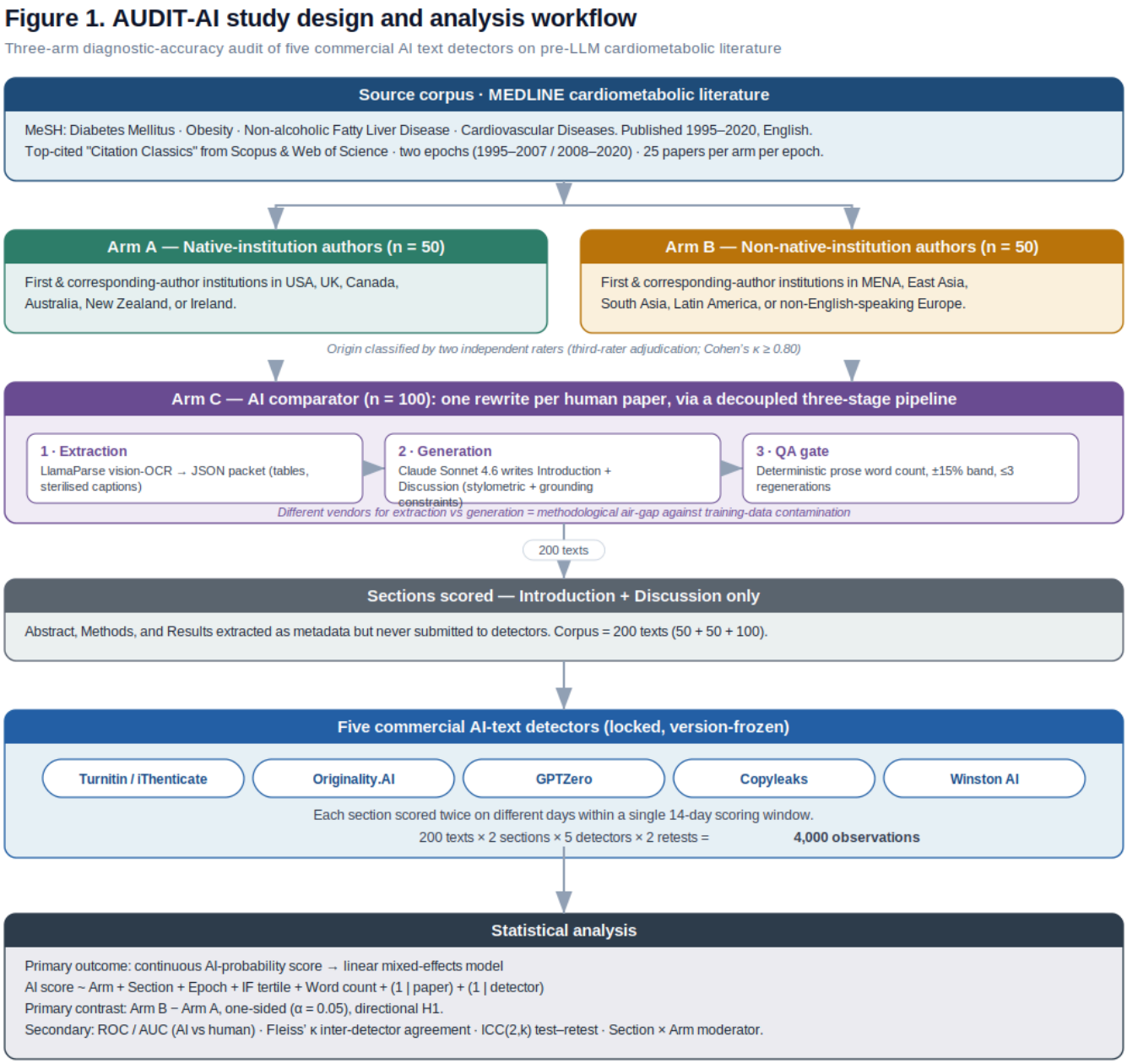


Figure 1

See image above for figure legend

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AUDITAISupplementaryFile1.docx
- AUDITAISupplementaryFile2.pdf
- AUDITAISupplementaryFile3.docx
- AUDITAISupplementaryFile4.xlsx
- AUDITAISupplementaryFile5.xlsx