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.

Linguistic Resistance to AI Writing Tools: A Review and Conceptual Framework

Sharip Isaev

isaev.shm@edu.spbstu.ru

Peter the Great St. Petersburg Polytechnic University https://orcid.org/0009-0006-8141-5110

Research Article

Keywords: AI writing tools, linguistic resistance, academic writing, professional identity, detection bias, post-AI education, engineering education

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10348445/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

The rapid adoption of writing tools based on large language models (LLMs) has placed many writers in a contradictory position: institutions encourage AI-assisted productivity while simultaneously deploying detection tools whose error rates fall disproportionately on particular writer populations. This paper is a hybrid contribution: a PRISMA 2020-informed structured literature search and screening procedure (41 included studies, 2020–2025) combined with thematic synthesis and conceptual development. We define linguistic resistance as the deliberate modification of text intended to reduce the probability of its classification as machine-generated, and we derive a three-tier behavioural typology (stylistic affirmation, strategic degradation, model personalisation) from the reviewed evidence. We then formulate the Efficiency Inversion Paradox, the proposition that detection pressure can redistribute writing effort rather than reduce it. We also introduce two exploratory heuristic instruments, the Linguistic Resistance Index (LRI) and the R-Score, offered strictly as candidate operationalisations for future empirical validation rather than as validated measures. The synthesis distinguishes documented effects (detector bias against non-native English writers, LLM-induced vocabulary shift in scientific writing, and the perception of AI-assisted text as professional but less sincere) from plausible mechanisms and speculative extensions. We discuss implications for engineering and polytechnic education, where assessed writing serves both communicative and competency-formation functions, and identify validation of the proposed instruments as the priority for subsequent primary research.

1 Introduction

1.1 Writing Technology as Effort Redistribution

Writing technologies have historically redistributed effort rather than eliminated it. The typewriter removed illegible handwriting at the cost of motor retraining; word processors moved correction effort from the physical to the structural level; grammar checkers reduced surface error while homogenising register. Kranzberg's first law states that technology is neither good nor bad, nor is it neutral [11]. It captures the recurring pattern. Each new tool restructures both the locus of writing effort and the standards against which output is judged.

LLM-based writing tools differ from earlier instruments in one respect that matters for this analysis: they generate semantic content rather than assist with form. Their adoption has been correspondingly fast and broad. Corpus evidence from biomedical publishing indicates the scale: an analysis of more than 15 million PubMed abstracts (2010–2024) found abrupt post-2022 frequency increases in characteristic style words and estimated that at least 13.5% of 2024 abstracts were processed with LLMs, with some subcorpora reaching about 40%. This linguistic footprint exceeds that of the COVID-19 pandemic [2].

1.2 The Paradox: When Detection Increases Effort

Alongside adoption, institutions deployed automated detectors that treat statistical features of text, such as low perplexity, high grammatical regularity, and characteristic lexical choices, as probabilistic markers

of machine authorship. Independent evaluations document substantial reliability problems. Liang et al. [3] tested seven commercial GPT detectors on 91 human-written TOEFL essays and 88 essays by native US eighth-grade writers: the mean false-positive rate was 61.3% for the non-native essays against 5.1% for the native essays, with 19.8% of TOEFL essays flagged unanimously by all seven detectors and 97.8% flagged by at least one. Weber-Wulff et al. [22] evaluated 14 detection tools and concluded that none was sufficiently accurate or reliable for high-stakes use, with accuracy degrading further for manually edited and machine-paraphrased text. Documented cases of scholars falsely accused on the basis of detector output, disproportionately non-native English speakers and writers with distinctive styles, illustrate the practical consequences [14].

Under these conditions a counterintuitive cost structure emerges. A writer whose draft, whether AI-assisted or not, risks being flagged has an incentive to spend additional effort modifying the text: inserting conversational phrasing, removing precise vocabulary, breaking regular syntax. Detection-evasion effort of this kind is demonstrably effective. In the Liang et al. study, prompting an LLM to enrich the linguistic diversity of the TOEFL essays reduced the mean false-positive rate from 61.3% to 11.8% [3]; conversely, Perkins et al. [4] showed that adversarial prompting sharply degraded detector performance: the institutional detector identified only about 54.8% of the AI-generated content, even though 91% of the experimental submissions still received some flag; academic staff, in turn, escalated only 54.5% of those submissions to misconduct procedures. Evasion is therefore feasible, effortful, and largely invisible to the institution, a combination that the academic-integrity literature identifies as corrosive to assessment validity [12].

A further asymmetry compounds the problem. Detectors calibrated on average human writing penalise text that is statistically too predictable (a property of careful non-native writing with restricted lexical variability [3]), while highly polished expert prose can likewise approach the statistical profile of LLM output. Writers at both ends of the proficiency spectrum thus face elevated misclassification risk from opposite stylistic directions, and the rational response in both cases is deliberate textual modification.

1.3 Theoretical Foundations

The analysis draws on five established research traditions. First, Cognitive Load Theory [21] supplies the effort vocabulary: modifications undertaken to evade detection constitute extraneous load, unrelated to the rhetorical task of the text itself. Second, the automation literature has long observed that automation tends to relocate rather than remove human work, generating new monitoring and compensation tasks. These are the “ironies of automation” described by Bainbridge [24], and detection-evasion effort is a textual instance of this pattern. Third, research on AI-mediated communication (AIMC) [13] documents how algorithmic involvement changes both language production and the social perception of messages [7, 8]. Experimental work shows that algorithmic response suggestions shift register toward positive emotional language while suspected AI use lowers interpersonal evaluations [7], and that AI-assisted workplace messages are rated as professional and effective but lower on sincerity and caring, particularly when AI use is disclosed [9]. Fourth, the academic-integrity and detection-policy literature [12, 14, 22, 4] frames the institutional side of the interaction. Fifth, work on professional identity and

authenticity in writing explains why writers experience detection pressure as a threat to identity and not merely as a compliance cost. This tradition includes the documented stigma of “AI shaming” [6]. Human inability to identify AI-generated language reliably, demonstrated across six experiments with 4,600 participants [10], implies that these perceptual and reputational dynamics operate largely independently of actual authorship.

1.4 Knowns, Gaps, and Contributions

What is known can be stated compactly. LLM uptake in scholarly and professional writing is large and measurable [2]. Automated detection is unreliable and biased against identifiable writer populations [3, 14, 22]. AI involvement changes how text and its author are perceived [7, 9, 10]. Stigma around AI use is documented [6]. What remains underexplored is the writer’s behavioural response to this configuration of incentives. Individual phenomena such as evasion editing, vocabulary avoidance, and disclosure reluctance have been observed in separate literatures. However, no unified construct connects them, no typology organises their variation, and no measurement proposals exist for the effort they consume.

This paper addresses that gap with three contributions. First, it provides a working definition and a three-tier behavioural typology of linguistic resistance, grounded in a thematic synthesis of 41 studies identified through a PRISMA 2020-informed structured search. Second, it formulates the Efficiency Inversion Paradox and proposes two exploratory heuristic instruments, the Linguistic Resistance Index (LRI) and the R-Score, as candidate operationalisations for future empirical work. The paper does not claim to validate them. Third, it offers an evidence map that explicitly separates documented effects from plausible mechanisms and speculative extensions. The contribution is therefore a conceptual synthesis with operational hypotheses, not the demonstration of a new empirical regularity.

The framing is anchored in engineering and polytechnic higher education for a substantive reason: writing tasks in technical programmes function simultaneously as professional communication products and as competency-formation instruments. If students under detection pressure adopt deliberate quality degradation as a default writing strategy, both functions are undermined at once, an outcome particularly at odds with the profession’s normative emphasis on precision.

2 Methodology

2.1 Review Type and Rationale

We conducted a hybrid review combining (i) a PRISMA 2020-informed structured literature search and screening procedure [19] with (ii) thematic synthesis and conceptual development. We deliberately describe the search as PRISMA-informed rather than as a full systematic review: the procedure follows PRISMA 2020 reporting logic for identification, screening, and inclusion, but screening was performed by a single reviewer and no meta-analytic synthesis is undertaken, so the paper does not claim the evidentiary status of a registered systematic review. The hybrid format was selected because the target phenomenon is recent, sparsely studied as a unified construct, and distributed across heterogeneous

traditions (natural language processing, applied linguistics, education research, human–computer interaction, and the science of science). A quantitative meta-analysis would not be defensible given this heterogeneity; a purely narrative review would not provide the transparency needed to appraise the evidence base.

2.2 Databases and Search Period

Searches were conducted in four bibliographic databases: Web of Science Core Collection, Scopus, PubMed (for vocabulary-shift and biomedical-publishing evidence), and the ACM Digital Library (for HCI and writing-tool studies). The search window was 1 January 2020–31 October 2025; the lower bound corresponds to the public availability of GPT-3-class models, and the final search was executed on 31 October 2025.

2.3 Search Strings

Five domain queries were applied to titles, abstracts, and author keywords (TS field in Web of Science; TITLE-ABS-KEY in Scopus; [tiab] tags in PubMed; default index in ACM DL), with Boolean structure held constant across databases:

1. **Q1**. AI writing tool adoption: ("AI writing tool*" OR "ChatGPT" OR "large language model*" OR "LLM") AND ("academic writing" OR "scientific writing" OR "writing process" OR "student writing");
2. **Q2**. Detection accuracy and bias: ("AI detect*" OR "GPT detect*" OR "AI-generated text detection") AND ("false positive" OR "bias" OR "non-native" OR "ESL" OR "TOEFL");
3. **Q3**. AI-mediated communication: ("AI-mediated communication" OR "AI-assisted writing") AND ("perception*" OR "professional*" OR "sincerity" OR "authenticity");
4. **Q4**. Professional identity and stigma: ("AI shaming" OR "AI stigma" OR "professional identity") AND ("writing" OR "academic" OR "scholar*");
5. **Q5**. LLM-induced language change: ("LLM" OR "ChatGPT" OR "AI") AND ("vocabulary" OR "lexical" OR "stylometr*" OR "linguistic markers")

Database-specific syntax adjustments (field tags, truncation symbols) were applied without altering the conceptual structure of the queries. Reference lists of all included studies and of two recent reviews were hand-searched (backward citation chasing).

2.4 Inclusion and Exclusion Criteria

Table 1
Eligibility criteria applied at full-text assessment.

Dimension	Included	Excluded
Publication type	Peer-reviewed journal articles; full conference papers; methodologically explicit preprints with adequate reporting	Editorials, opinion pieces, blog posts, theses, white papers
Date	1 Jan 2020–31 Oct 2025	Outside the window
Topic	Empirical, conceptual, or systematic analysis of LLM writing tools, AI text detection, AI-mediated communication, or LLM-induced linguistic change	Fully automated generation with no human-writer interaction; tool descriptions without analysis of writing effects
Language	English	Non-English (resource limitation; see Sect. 4.5)
Method reporting	Method, sample, and outcomes reported in sufficient detail for appraisal	Reporting too limited to permit appraisal

2.5 Screening Process and PRISMA Flow

The flow of records is shown in Fig. 1. Searches yielded 487 records (Web of Science = 178; Scopus = 164; PubMed = 72; ACM DL = 73). After removal of 76 duplicates, 411 records were screened on title and abstract, of which 317 were excluded as out of scope. Of 94 full-text articles assessed, 53 were excluded (out of scope, n = 29; ineligible publication type, n = 12; insufficient method reporting, n = 8; no relevant outcome, n = 4), leaving 41 studies in the qualitative synthesis. Eight of these report quantifiable outcomes directly relevant to the constructs of this review and form the evidence matrix in Table 3.

All screening was conducted by the single author. We state this explicitly as a methodological limitation affecting screening reliability (Sect. 4.5). To partially mitigate the risk, every full-text exclusion was documented with a written rationale, and the screening log is available from the author on request.

2.6 Quality Appraisal

Methodological quality of the included empirical studies was appraised with the Mixed Methods Appraisal Tool (MMAT, 2018 version) [17], applied by the single author within the rubric matching each study design (one point per criterion met, 0–5). Appraisal scores for the eight evidence-matrix studies are reported in Table 3. Commentaries, policy analyses, and perspective papers were retained in the qualitative synthesis where they met the eligibility criteria, but were not appraised with MMAT and do not contribute rows to the evidence matrix.

2.7 Approach to Heterogeneity and Evidence Levels

The included studies span corpus analyses (up to 15 million abstracts), surveys (N up to 887), controlled experiments (N up to 4,600), evaluation studies of detection tools, and conceptual work. This

heterogeneity precludes a defensible pooled effect estimate, and none is computed. Instead, the synthesis distinguishes three evidence levels, applied consistently throughout Sect. 4:

- Documented effects: quantified outcomes replicated in at least two independent empirical studies, or a single large-scale study with strong internal evidence, explicitly marked as single-study.
- Plausible mechanisms: theoretical proposals supported by converging indirect evidence without direct experimental confirmation.
- Speculative extensions: claims that follow logically from documented effects but remain untested, marked as such wherever they appear.

2.8 Status of the Proposed Instruments

The Linguistic Resistance Index (LRI) and the R-Score introduced in Sect. 3.4, and the Conceptual Heuristic Causal Framework (CH-CF) in Sect. 3.2, are conceptual proposals, not validated instruments or fitted models. They are presented as candidate operationalisations of constructs identified by the review, without claims of construct validity, reliability, or measurement invariance; no numerical outputs of these instruments are reported in this paper, and the CH-CF carries no estimated path coefficients. Their function is to make the constructs testable in subsequent primary research.

3 Synthesis

3.1 Linguistic Resistance: Definition and Three-Tier Typology

Thematic synthesis of the 41 included studies yields the following working definition:

Definition 1

(Linguistic Resistance). Linguistic resistance (LR) is any deliberate or semi-deliberate textual practice through which a writer modifies a text in order to reduce the probability that the text is identified as AI-generated by automated detectors or human evaluators. The text in question may be produced wholly or partially with AI assistance, or produced without AI but at risk of AI misclassification. LR is distinct from quality-oriented editing: its defining criterion is that the modification responds to an anticipated authorship judgment rather than to communicative norms alone.

Three behavioural tiers recur across the reviewed studies, corresponding to three relationships between writer and tool: avoidance of machine-typical features, exploitation of detector weaknesses, and integration of resistance into model configuration. Each tier carries a distinct effort direction and identity outcome (Fig. 2). A conceivable fourth tier, complete rejection of AI tools, falls outside the definition. Where no AI is used and no misclassification risk is anticipated, no resistance is required. The

occurrence descriptors in Fig. 2 are qualitative syntheses of where each tier is reported in the literature; the review provides no basis for population-level prevalence estimates, and none are offered.

3.2 Conceptual Heuristic Causal Framework (CH-CF)

The CH-CF organises the constructs identified by the review into direction-of-effect claims, each anchored in cited evidence. It is not a structural equation model and reports no path coefficients. The principal qualitative relationships are the following. (a) Institutional detection enforcement raises anticipated misclassification risk and exposure to AI-related stigma [3, 14, 22, 6]. (b) Elevated risk and stigma push writers toward Tier 2 strategic degradation in high-stakes contexts [3, 4]. (c) Tier 2 modification adds an extraneous cognitive-load component to the writing process, in the sense of Cognitive Load Theory [21]. (d) Repeated Tier 2 modification co-occurs with reported identity dissonance and writing-related stress [6]. (e) Writers with technical literacy and adequate tool access can shift toward Tier 3, absorbing part of the resistance cost into model configuration [5, 15]. Quantitative calibration of these relationships is left to primary research.

3.3 Documented Effects

Detection bias against non-native English writers. Liang et al. [3] report a mean false-positive rate of 61.3% for human-written TOEFL essays against 5.1% for native US eighth-grade essays across seven commercial detectors, attributable to the lower perplexity of non-native writing; enriching the linguistic diversity of the same essays reduced the mean false-positive rate to 11.8% (Fig. 3). Weber-Wulff et al. [22], evaluating a different set of 14 tools, independently document overall accuracy levels they judge inadequate for high-stakes use, with further degradation for edited and paraphrased text. The two studies converge on the conclusion that current detectors cannot be relied upon for individual-level decisions, and the Liang et al. results specifically document the non-native bias; case evidence of resulting false accusations corroborates the practical impact [14]. We treat detector unreliability, including the documented bias against non-native writers, as a documented effect.

Vocabulary shift in LLM-influenced scientific writing. Kobak et al. [2] analyse more than 15 million PubMed abstracts (2010–2024) and document abrupt post-2022 excess frequencies of characteristic style words, estimating that at least 13.5% of 2024 abstracts were LLM-processed, with subcorpora reaching about 40%. Because the corpus and time series derive from a single primary study, we report this as a single-study documented effect with strong internal evidence. A further proposition is that writers will begin avoiding publicised marker words, which would appear as subsequent frequency declines for those markers. This is an inference consistent with the resistance framework but not yet established by peer-reviewed longitudinal evidence. We treat it as a plausible mechanism (Sect. 4.2) and a directly testable prediction.

AI-assisted text is perceived as professional but less sincere. Coman and Cardon [9] (survey of 887 working adults) report that AI-generated workplace messages are consistently rated professional, effective, and direct, while sincerity and caring are rated lower in some disclosed conditions. Hohenstein

et al. [7] show experimentally that algorithmic response suggestions shift language toward positive emotional registers and that communicators suspected of using algorithmic responses are evaluated more negatively. Jakesch, Hancock and Naaman [10] demonstrate across six experiments (N = 4,600) that human identification of AI-generated self-presentations remains near chance, indicating that these evaluative shifts attach to suspected rather than actual authorship. Together these studies document a perceptual divergence: AI involvement raises perceived professionalism while lowering perceived sincerity, mediated by suspicion rather than detection.

3.4 Exploratory Heuristic Instruments

We propose two instruments for future empirical validation and restate that they are not validated and yield no meaningful point estimates in their present form. The equations are reported for their conceptual content.

Detection enforcement intensity λ. Institutional detection pressure is proposed as a composite $\lambda \in [0, 1]$ over three observable indicators. These are the existence of a formal AI-detection policy (binary), the proportion of assessed work subject to detection screening (continuous), and the severity of stated consequences (ordinal, rescaled to [0, 1]). The composite is provisionally defined as the arithmetic mean of the three; equal weighting is a starting assumption to be revised by measurement research, not an empirical claim.

Linguistic Resistance Index (LRI). The LRI is proposed as a measure of excess writing effort attributable to resistance modification, relative to a direct-composition baseline:

$$LRI = (T_{AR} - T_A) / T_A + \lambda \cdot \Delta P(w) \quad (1)$$

where $T_{\square_r}$ is total production time for AI-assisted text including resistance modification; $T\square$ is the estimated direct-composition time for a text of equivalent scope; $\Delta P(w)$ is a perplexity-shift term for text w, positive when complexity is deliberately reduced toward detector-safe registers; and λ is the enforcement intensity defined above. $LRI > 0$ marks the conceptual Efficiency Inversion Zone, in which assisted production plus resistance costs more than direct composition. We do not report a calibrated LRI value: the available primary studies do not jointly provide the time and perplexity data required. Validation requires controlled writing experiments with time tracking, which we identify as the priority for future work.

R-Score. The R-Score is proposed as a text-level indicator of aggregate resistance signal, based on divergence from the LLM-era vocabulary baseline established by Kobak et al. [2]:

$$R(w) = \Sigma_j [f_0(w_j) - f(w_j)] \cdot o(w_j) \quad (2)$$

where j indexes lexical items in text w; $f_0(w\square)$ is the pre-2022 baseline frequency of item $w\square$ in the reference corpus [2]; $f(w\square)$ is its observed frequency in the target text; and $o(w\square) \in \{+1, -1\}$ is an orientation term (+ 1 for LLM-favoured items being avoided, − 1 for LLM-disfavoured items being

amplified). The R-Score inherits all limitations of vocabulary-based stylometry (sensitivity to disciplinary lexicon, topic drift, and ongoing baseline shift) and, like the LRI, requires primary data with annotated resistance intent for validation.

Table 2 summarises the constructs, their proposed operational indicators, expected data sources, and validation status.

Table 2
Construct map for the proposed instruments.

Construct	Operational indicator (proposed)	Expected data source	Validation status
Detection enforcement intensity (λ)	Policy existence; screening coverage; sanction severity (equal-weight composite)	Institutional policy documents; administrative records	Unvalidated; weighting provisional
Resistance effort (LRI)	Relative excess production time plus weighted perplexity shift, Eq. (1)	Controlled writing experiments with time tracking and text logging	Unvalidated; no calibrated values reported
Resistance signal in text (R-Score)	Oriented divergence from LLM-era vocabulary baseline, Eq. (2)	Corpora with annotated resistance intent; baseline from [2]	Unvalidated; baseline-dependent
Behavioural tier (T1–T3)	Coded writer strategy from process data, interviews, or edit logs	Diary and interview studies; keystroke or revision logs	Typology grounded in synthesis; coding scheme untested

3.5 Evidence Matrix

Table 3 lists the eight studies that report quantifiable outcomes directly relevant to the constructs of this review. Outcomes are reproduced as reported in each primary study, in the metric used by that study. No standardised effect sizes are computed or imputed by the present author: the designs, samples, and outcome constructs are too heterogeneous for a common metric, and converting reported percentages or accuracy figures into standardised differences would suggest a comparability the data do not possess.

Table 3
Evidence matrix: studies reporting quantifiable outcomes.

Study	Sample corpus	Design	Key outcome (as reported)	MMAT
Liang et al. 2023 [3]	91 TOEFL + 88 US 8th-grade essays	Detector evaluation	Mean FPR 61.3% (non-native) vs 5.1% (native); 19.8% unanimous flags; enrichment reduced FPR to 11.8%	5/5
Kobak et al. 2025 [2]	> 15M PubMed abstracts (2010–2024)	Corpus analysis	≥ 13.5% of 2024 abstracts LLM-processed (lower bound); up to ≈ 40% in subcorpora	4/5
Weber-Wulff et al. 2023 [22]	14 detection tools; multi-case set	Detector evaluation	No tool achieved accuracy judged adequate for high-stakes use; accuracy degraded for edited/paraphrased text	4/5
Perkins et al. 2024 [4]	22 experimental submissions; 15 staff markers	Field experiment	Detector flagged 91% of submissions but identified ≈ 54.8% of AI content under adversarial prompting; 54.5% escalated by staff	4/5
Jakesch et al. 2023 [10]	N = 4,600 (six experiments)	Online experiments	Human identification of AI-generated self-presentations near chance (≈ 50–52% accuracy)	5/5
Coman & Cardon 2024 [9]	N = 887 working adults	Survey experiment	AI messages rated professional/effective; sincerity and caring rated lower in some disclosed conditions	4/5
Hohenstein et al. 2023 [7]	Two randomised experiments	Experiments	Smart replies increased positive emotional language; suspected AI use lowered interpersonal evaluations	4/5
Jakesch et al. 2023 [15]	N = 1,506	Online experiment	Co-writing with an opinionated model shifted the opinions participants expressed and reported	5/5

All figures are as reported in the cited primary study; no pooled or converted effect sizes are computed. MMAT appraisal (0–5) by the single author; designs outside MMAT rubrics are not appraised.

4 Discussion

The discussion applies the three-level evidence hierarchy of Sect. 2.7, namely documented effects, plausible mechanisms, and speculative extensions, to prevent slippage between empirical findings and conceptual extrapolation.

4.1 Documented Effects: What the Evidence Currently Supports

Three findings are sufficiently established to be treated as documented within this review. (i) Current AI detectors are unreliable for individual-level decisions and carry a measured false-positive bias against non-native English writers [3, 14, 22]. (ii) LLM-influenced scientific writing exhibits a large, measurable lexical shift [2] (single-study, large-scale). (iii) AI involvement in text is associated with a perceptual divergence, namely higher perceived professionalism but lower perceived sincerity, that attaches to suspected rather than actual authorship [7, 9, 10]. These findings jointly support a minimal proposition: institutional deployment of detection tools creates measurable downstream costs for identifiable writer populations.

4.2 Plausible Mechanisms: The Efficiency Inversion Paradox

The Efficiency Inversion Paradox (EIP) is the proposition that, beyond some enforcement threshold, total writing effort under AI assistance exceeds direct-composition effort, because resistance modification adds a workflow phase with no counterpart in unassisted writing. The EIP combines two cost components: temporal displacement and extraneous cognitive load [21]. It is best treated as a plausible mechanism. Indirect support comes from the demonstrated effectiveness and non-triviality of evasion editing [3, 4], from professional-editor evidence that aligning LLM output with human stylistic norms requires substantial, skilled revision [5], and from the documented stigma that motivates concealment [6]. Direct confirmation does not yet exist. Such confirmation would require a controlled comparison of total task time and load measures across detection regimes. The LRI (Sect. 3.4) is designed to support exactly such experiments.

Two adjacent mechanisms have the same status. First, the link from AI-related stigma to Tier 2 degradation is consistent with the qualitative literature [6] but has not been isolated experimentally. Second, an arms-race dynamic follows from theoretical work on the limits of reliable detection [23] and from the homogenisation pressures documented in human–AI co-writing [15, 5]. This dynamic involves detector retraining, drift in human writing norms, and vocabulary avoidance feeding back into one another, and its long-run behaviour is unobserved. The prediction that publicised marker words (e.g., those identified in [2]) will decline in frequency as writers learn to avoid them is directly testable against future corpus releases.

4.3 Speculative Extensions: Where the Evidence Does Not Yet Reach

Several propositions of the framework exceed the current evidence and are marked accordingly. The equity claim attached to Tier 3 is consistent with general analyses of generative-AI inequality [1] but has not been demonstrated for resistance behaviour specifically. That claim holds that model

personalisation, while offering the most favourable effort profile, is disproportionately accessible to already advantaged writers. The distinction between perceived and actual authenticity, and its differential weight across disciplines, is conceptually motivated but unstudied. Projected effects of interventions are plausibility arguments, not impact estimates. One example is the expectation that normalising disclosure would reduce Tier 2 incidence. Recent work suggests disclosure itself carries perception costs [9, 18], which complicates simple policy predictions and reinforces the need for evaluation rather than assumption. A related speculative extension concerns Tier 1 stylistic affirmation at scale: if many writers converge on similar AI-suggested phrasings despite individual customisation, the aggregate effect could resemble the reduction in collective content diversity reported for unconstrained generative-AI assistance [25]. This has not been examined for resistance-modified text specifically.

4.4 Implications for Engineering and Polytechnic Education

The implications most directly supported by the evidence concern detection policy. International engineering programmes enrol large shares of students writing in English as an additional language, precisely the population for which false-positive rates are documented to be highest [3, 14]. Institutional deployment of detectors should therefore be evaluated against population-specific false-positive rates rather than vendor-reported aggregate accuracy, and the documented bias is in itself sufficient justification for a mandatory human-review and appeals channel wherever detector output can trigger sanctions. Emerging policy scholarship converges on the same conclusion: detection-centred regimes erode trust without resolving the underlying assessment problem, and process-oriented and disclosure-based approaches are the principal alternatives under discussion [12, 16].

A second implication concerns register pedagogy. If polished, precise prose is increasingly read as machine-like in informal evaluative contexts while remaining obligatory in formal technical genres [7, 9], students face a register-management problem that current technical-communication curricula do not address explicitly. The magnitude of this effect in engineering settings specifically is unknown; targeted primary research would be valuable.

4.5 Limitations

Five limitations require explicit acknowledgement. First, screening and quality appraisal were conducted by a single reviewer; documented exclusion rationales mitigate but do not remove the associated reliability risk, and the search protocol was not prospectively registered. Second, the search was restricted to English-language publications, excluding potentially relevant work on writer experience and detection bias in non-Anglophone academic contexts. This is a substantive gap, since the detection-bias findings themselves concern non-native writers, and early evidence from other languages suggests the bias structure may differ across linguistic settings. Third, the LRI and R-Score are unvalidated; the equations describe conceptual structure and must not be used as measurement instruments in their present form. Fourth, the included studies concentrate on higher education and biomedical publishing; resistance dynamics in other professional sectors may differ. Fifth, no pooled quantitative summary is

computed, by design. The synthesis is correspondingly qualitative, and its conclusions carry the evidentiary weight of a structured narrative synthesis, not of a meta-analysis.

5 Conclusions

This review proposes that, in high-enforcement institutional contexts, LLM writing tools have not eliminated writing effort but may redistribute it into a resistance-modification phase. The proposition is consistent with three documented findings and with a cognitive-load mechanism formulated here as the Efficiency Inversion Paradox. The three findings are detector unreliability with measured bias against non-native English writers, large-scale LLM-induced vocabulary shift, and the perceptual divergence between professionalism and sincerity. The three-tier behavioural typology and the exploratory instruments (LRI, R-Score) are offered as conceptual scaffolding for primary research, not as validated measures.

For engineering and polytechnic education, the best-supported implications concern detection policy for non-native English writers and the emerging register-management problem in technical communication. Broader institutional responses (disclosure normalisation, process-based assessment, and AI-literacy instruction) remain plausible but unevaluated as causal interventions.

Future work should pursue four objectives. The first is experimental validation of the LRI with time tracking across detection regimes. The second is testing the marker-avoidance prediction against successive corpus releases. The third is cross-linguistic extension of the typology beyond English. The fourth is longitudinal evaluation of detection-policy effects on writer behaviour, including the equity profile of Tier 3 adoption.

Declarations

Funding: No funding was received to assist with the preparation of this manuscript.

Competing interests: The author has no relevant financial or non-financial interests to disclose.

Ethics approval: Not applicable. This article is a review of previously published studies and did not involve any new research with human participants or animals conducted by the author.

Consent to participate: Not applicable. No human participants were involved in this study.

Consent for publication: Not applicable. This study contains no data or images relating to any individual person.

Data availability: This study is a review of published literature; all data analysed are contained in the works cited in the reference list. The screening log and the documented full-text exclusion rationales are available from the author on reasonable request.

Code availability: Not applicable. No custom code or software application was generated or used in this study.

Clinical trial number: Not applicable.

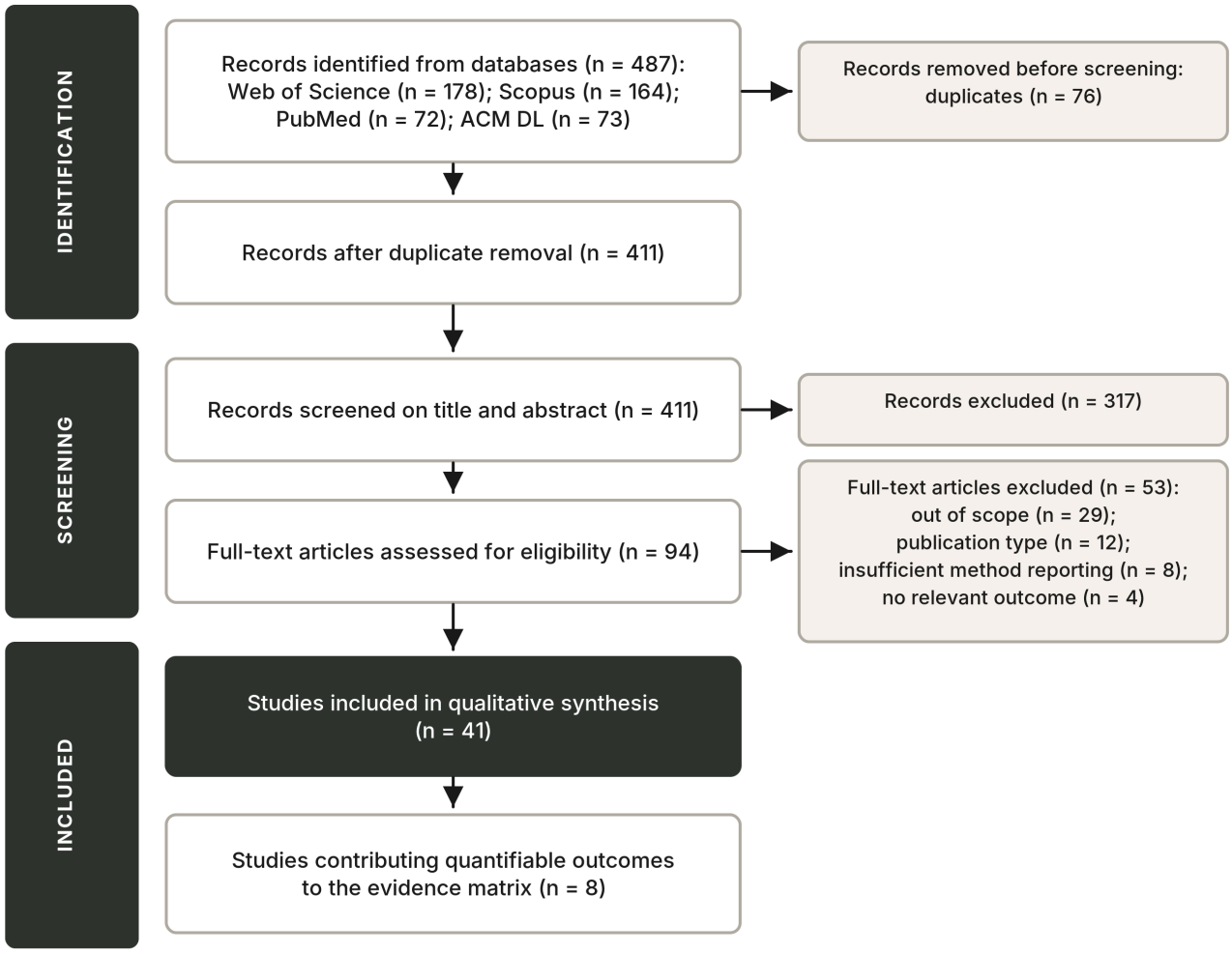
Authors' contributions: Author 1 conceived and designed the study, conducted the literature search and screening, performed the quality appraisal and thematic synthesis, developed the conceptual framework, and wrote and approved the final manuscript.

References

1. Capraro V, Lentsch A, Acemoglu D et al (2024) The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus 3(6):pgae191. https://doi.org/10.1093/pnasnexus/pgae191
2. Kobak D, González-Márquez R, Horvát E-Á, Lause J (2025) Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Sci Adv 11(27):eadt3813. https://doi.org/10.1126/sciadv.adt3813
3. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J (2023) GPT detectors are biased against non-native English writers. Patterns 4(7):100779. https://doi.org/10.1016/j.patter.2023.100779
4. Perkins M, Roe J, Postma D, McGaughran J, Hickerson D (2024) Detection of GPT-4 generated text in higher education: combining academic judgement and software to identify generative AI tool misuse. J Acad Ethics 22(1):89–113. https://doi.org/10.1007/s10805-023-09492-6
5. Chakrabarty T, Laban P, Wu C-S (2025) Can AI writing be salvaged? Mitigating idiosyncrasies and improving human–AI alignment in the writing process through edits. In: Proc. CHI Conf. Human Factors in Computing Systems (CHI '25). ACM, New York https://doi.org/10.1145/3706598.3713559
6. Giray L (2024) AI shaming: the silent stigma among academic writers and researchers. Ann Biomed Eng 52(9):2319–2324. https://doi.org/10.1007/s10439-024-03582-1
7. Hohenstein J, Kizilcec RF, DiFranzo D, Aghajari Z, Mieczkowski H, Levy K, Naaman M, Hancock J, Jung MF (2023) Artificial intelligence in communication impacts language and social relationships. Sci Rep 13:5487. https://doi.org/10.1038/s41598-023-30938-9
8. Fu Y, Foell S, Xu X, Hiniker A (2024) From text to self: users' perception of AIMC tools on interpersonal communication and self. In: Proc. CHI Conf. Human Factors in Computing Systems (CHI '24). ACM, New York https://doi.org/10.1145/3613904.3641955
9. Coman AW, Cardon P (2024) Perceptions of professionalism and authenticity in AI-assisted writing. Bus Prof Commun Q 87(2). https://doi.org/10.1177/23294906241233224
10. Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. Proc. Natl. Acad. Sci. U.S.A. 120(11), e2208839120 https://doi.org/10.1073/pnas.2208839120
11. Kranzberg M (1986) Technology and history: Kranzberg's laws. Technol Cult 27(3):544–560. https://doi.org/10.2307/3105385

12. Cotton DRE, Cotton PA, Shipway JR (2024) Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int 61(2):228–239. https://doi.org/10.1080/14703297.2023.2190148
13. Hancock JT, Naaman M, Levy K (2020) AI-mediated communication: definition, research agenda, and ethical considerations. J Comput -Mediat Commun 25(1):89–100. https://doi.org/10.1093/jcmc/zmz022
14. Giray L (2024) The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. Ser Libr 85(5–6):181–189. https://doi.org/10.1080/0361526X.2024.2433256
15. Jakesch M, Bhat A, Buschek D, Zalmanson L, Naaman M (2023) Co-writing with opinionated language models affects users' views. In: Proc. CHI Conf. Human Factors in Computing Systems (CHI '23). ACM, New York https://doi.org/10.1145/3544548.3581196
16. Ranjbaran Madiseh F, Giray L, Sevnarayan K (2025) Beyond policing: AI writing detection tools, trust, academic integrity, and their implications for college writing. Internet Ref Serv Q 29(1):1–34. https://doi.org/10.1080/10875301.2024.2437174
17. Hong QN, Pluye P, Fàbregues S et al (2018) Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright #1148552, Canadian Intellectual Property Office
18. Cardon PW, Coman AW (2025) Professionalism and trustworthiness in AI-assisted workplace writing: the benefits and drawbacks of writing with AI. Int J Bus Commun. https://doi.org/10.1177/23294884251350599
19. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
20. Giray L (2023) Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng 51(12):2629–2633. https://doi.org/10.1007/s10439-023-03272-4
21. Sweller J, van Merriënboer JJG, Paas F (2019) Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev 31(2):261–292. https://doi.org/10.1007/s10648-019-09465-5
22. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, Šigut P, Waddington L (2023) Testing of detection tools for AI-generated text. Int J Educ Integr 19:26. https://doi.org/10.1007/s40979-023-00146-z
23. Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S (2023) Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156. https://doi.org/10.48550/arXiv.2303.11156
24. Bainbridge L (1983) Ironies of automation. Automatica 19(6):775–779. https://doi.org/10.1016/0005-1098(83)90046-8
25. Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci Adv 10:eadn5290. https://doi.org/10.1126/sciadv.adn5290

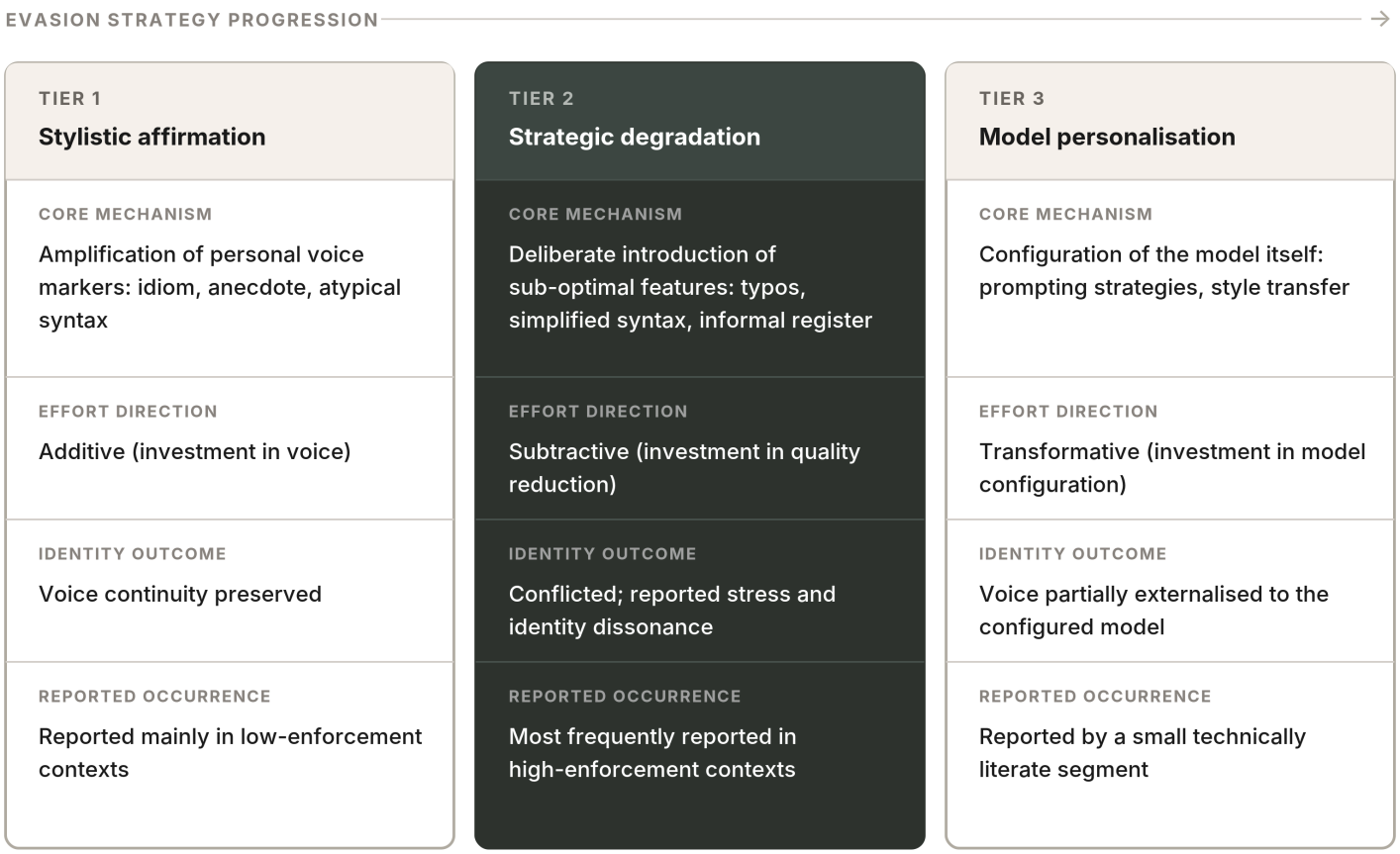
Figures



Search window: 1 January 2020 – 31 October 2025. Final search: 31 October 2025.
Title/abstract and full-text screening conducted by a single reviewer (see Sect. 4.5).

Figure 1

Flow of records through identification, screening, and inclusion, reported following PRISMA 2020 conventions. Single-reviewer screening is acknowledged as a limitation (Sect. 4.5).



Occurrence descriptors are qualitative syntheses of the reviewed studies, not validated population estimates.

Figure 2

Three-tier behavioural typology of linguistic resistance derived from the thematic synthesis. Occurrence descriptors are qualitative; they are not validated population estimates.

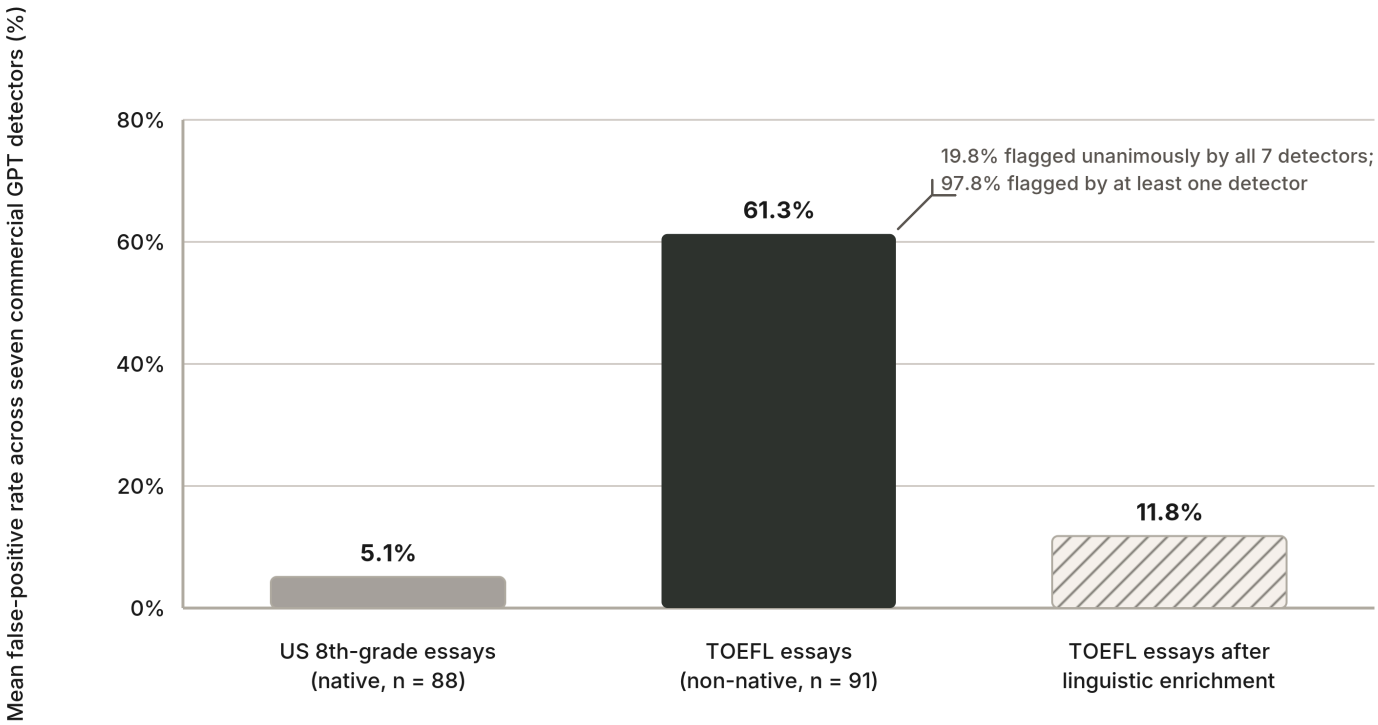


Figure 3

False-positive rates reported by Liang et al. [3]: means across seven commercial GPT detectors for native US eighth-grade essays (n = 88), human-written TOEFL essays (n = 91), and the same TOEFL essays after LLM-assisted linguistic enrichment. Annotated percentages give the share of TOEFL essays flagged unanimously and by at least one detector.