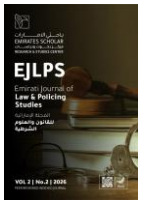


Emirati Journal of Law and Policing Studies
Vol 2 Issue 1 (2026)
Pages 48 -55
Available at
www.emiratesscholar.com

Whose Words Are These? Forensic Linguistics, Authorship Evidence, and the Courts in the Age of Large Language Models

Olivia Wilson, Ethan J. Cooper, Mina Kowalski

Department of Literature, Yale University, New Haven, CT, United States, Department of Computer Science, University of Toronto - Canada, Department of Linguistics, University of Warsaw, Warsaw, Poland

ARTICLE INFO

KEYWORDS

forensic linguistics; authorship attribution; large language models; AI-generated text; stylometry; expert evidence; admissibility

HOW TO CITE

Wilson, O., Cooper, E. J., & Kowalski, M. (2026). Whose Words Are These? Forensic Linguistics, Authorship Evidence, and the Courts in the Age of Large Language Models. Emirati Journal of Law and Policing Studies, 2(1), 48-55.
https://doi.org/10.54878/5s9t9f87

ABSTRACT

For half a century, forensic linguistics has rested on a quietly powerful premise: that people leave themselves in their language. Habits of lexis, syntax, spelling, and discourse accumulate into an idiolect distinctive enough, in favourable conditions, to link a questioned text to its author, and courts in many jurisdictions have admitted analysis built on that premise in cases ranging from threatening letters to disputed confessions. Large language models unsettle the premise at its root. When any literate person can produce fluent text through a model, when a coerced author can be assisted, imitated, or replaced by a machine, and when the tools marketed to detect machine text prove fragile under paraphrase and biased against non-native writers of English, the evidential status of the written word requires re-examination. This paper conducts that re-examination across the three disciplines the problem now spans. From linguistics, we assess which components of idiolect survive machine mediation and which dissolve. From computer science, we review the technical state of AI- text detection, watermarking, and stylometric attribution under adversarial conditions, and formalise the underlying inference problem. From literary and interpretive studies, we ask what authorship itself now means when composition is distributed between human intention and machine articulation, and what that shift does to legal categories, confession, threat, defamation, plagiarism, built on unitary authorship. We argue that forensic authorship analysis is not obsolete but must be rebuilt on likelihood-ratio reporting, validated error rates, and explicit uncertainty about machine mediation, and that courts should treat current AI-text detectors as investigative leads rather than evidence. Recommendations for practitioners, courts, and researchers follow.

Double-blind peer review by independent reviewers.

Received: 14 Jan 2026, Accepted: 19 May 2026, Published: 22 Jun 2026

*Corresponding author: olivia.wilson@yale.edu

1. Introduction

In 2013, a Sunday newspaper asked whether a debut crime novel by one Robert Galbraith might have been written by J. K. Rowling; computational stylistics supplied part of the answer, and the attribution held. The episode became the friendly public face of a serious enterprise: the use of linguistic evidence to answer questions of authorship and meaning in legal settings. Behind it stands a discipline with real casework, threatening communications traced through idiosyncratic spellings, disputed police statements exposed by register analysis, trademark and plagiarism disputes resolved on lexical patterning, and a methodological literature that has laboured, with mixed success, to put numbers on its judgements (Coulthard, Johnson, & Wright, 2017; Solan & Tiersma, 2005; Grant, 2007).

The discipline's foundations were laid in a world with one kind of author. A text might be copied, dictated, edited, or coerced, but somewhere behind it stood a human whose linguistic habits had shaped it, and the analyst's task was to characterise those habits. Large language models end that world. Text generated by a transformer-based model (Vaswani et al., 2017; Brown et al., 2020) is fluent, register-flexible, and carries no idiolect in the human sense; text merely edited or paraphrased by such a model is a blend whose human signal has been attenuated to an unknown degree. Meanwhile, an industry of detection tools promises to distinguish human from machine text, and their known failure modes, fragility under paraphrase (Sadasivan et al., 2023), systematic bias against non-native English writers (Liang et al., 2023), are exactly the failure modes that should alarm a legal system committed to validated, error-quantified expert evidence (Daubert v. Merrell Dow Pharmaceuticals, 1993).

This paper asks what remains of forensic authorship evidence, and what must change. We write from three disciplines because the problem is genuinely tripartite: linguistic, because idiolect and its markers are empirical claims about language; computational, because both generation and detection are engineering artefacts with measurable behaviour; and interpretive, because authorship is a concept with a history, and the law's version of that concept is now out of date. Our jurisdictional frame is the admissibility standards of common-law systems, with the US Daubert factors as the sharpest formulation, but the argument travels: every legal system that receives written evidence now receives, knowingly or not, machine-mediated text.

Section 2 reviews the discipline as it stood before generative models. Section 3 analyses the disruption: what LLMs do to idiolect, and what detection can and cannot deliver. Section 4 formalises the inference problem and proposes reporting standards. Section 5 examines the interpretive question of distributed authorship and its legal consequences. Section 6 discusses implications for courts and investigators, and Section 7 concludes.

2. Forensic Authorship Analysis Before the Machines

2.1 Idiolect and its markers

The working hypothesis of forensic authorship analysis is that every speaker-writer possesses a distinctive repertoire, an idiolect, observable as preferences among the options language offers: this word rather than that synonym, this clause structure, this punctuation habit, this way of opening a request or closing a threat (Coulthard et al., 2017). The hypothesis is probabilistic, not fingerprint-like. No single marker identifies an author; the case is built from the joint improbability of many shared preferences between questioned and known writings, assessed against how common those preferences are in a relevant population (Grant, 2007). Computational stylometry systematised the idea, demonstrating that frequencies of function words, character n-grams, and syntactic patterns separate authors with useful accuracy under controlled conditions (Juola, 2006; Stamatatos, 2009), while casework linguists emphasised markers closer to the surface: spelling errors, dialect forms, formulaic phrases, text-message abbreviations (Chaski, 2005).

Table 1. organises the marker inventory by linguistic level. The right-hand column anticipates the argument of Section 3: these levels are not equally robust once a language model stands between author and text.

Table 1. Levels of authorship markers in forensic linguistic analysis, and their exposure to machine mediation.

Level	Typical markers	Evidential value (pre-LLM)	Survival under machine mediation
Orthographic	Idiosyncratic spellings, punctuation habits, capitalisation, typos	High when consistent; famously decisive in casework	Low: models normalise spelling and punctuation

Lexical	Word choice among synonyms, regionalisms, taboo terms, formulaic phrases	Moderate–high with adequate data	Low–moderate: model vocabulary supplants personal lexicon
Syntactic	Clause complexity, subordination habits, characteristic constructions	Moderate; needs quantity	Low–moderate: prompts constrain content, model supplies structure
Discourse-pragmatic	Openings/closings, politeness strategies, threat construction, narrative habits	Moderate; genre-sensitive	Moderate: prompt-level intent may leak through
Content/knowledge	Facts only the author could know; timeline consistency	High but not linguistic per se	High: content originates with the human principal

2.2 The courtroom career of linguistic evidence

Legal receptions of this evidence have always been uneasy. In the United States, Daubert requires that expert methods be testable, peer-reviewed, of known error rate, and accepted in a relevant community (Daubert v. Merrell Dow Pharmaceuticals, 1993), and forensic stylistics has been challenged on each prong; qualitative marker analysis in particular has been criticised as unfalsifiable connoisseurship, while quantitative stylometry has faced questions about ecological validity, short texts, mixed genres, unknown candidate pools, that laboratory benchmarks under-represent (Chaski, 2005; Juola, 2006). Solan and Tiersma's (2005) survey of language evidence in criminal process reached a balanced verdict that still holds: linguistic analysis earns admissibility where its claims are calibrated and its limits stated, and forfeits it where analysts overreach. That admonition, written for a human-only world, now reads as the entrance requirement for a far harder examination.

2.3 What the casework tradition teaches

Three recurring lessons from the casework tradition deserve restating because each is about to be stress-tested. The first is that the strongest linguistic evidence has usually been negative or exclusionary: showing that a disputed confession contains register features characteristic of police statement-taking rather than of the suspect's known speech, or that a suicide note's phrasing is inconsistent with the deceased's writings, is generally safer inference than positively identifying one author among an open population (Coulthard et al., 2017; Solan & Tiersma, 2005). Exclusion requires only that the questioned text be unlike the candidate; identification requires that it be unlike everyone else. The second lesson is that data quantity governs everything: reliable stylometric separation needs known writings measured in thousands of words, in comparable genre and register, and much real casework fails this threshold before analysis begins (Juola, 2006; Chaski, 2005). The third is that the discipline's failures have come from overclaiming under adversarial pressure, experts asserting unique identification from marker sets whose population frequencies were never measured. Grant's (2007) programme of quantified, population-calibrated evidence was the corrective, and it is the part of the tradition most worth carrying into the machine era, because it is the part built to survive hostile cross-examination.

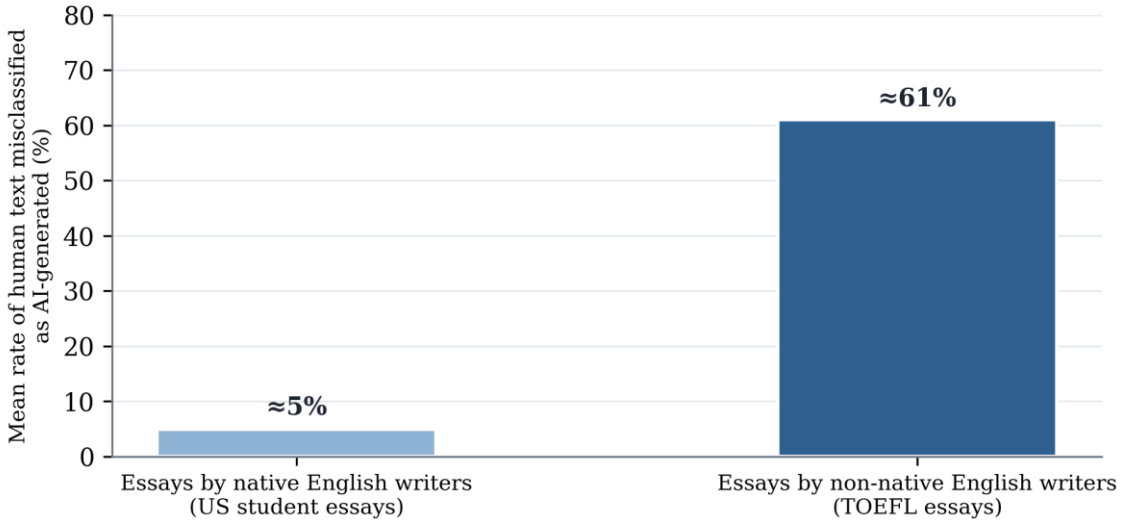
3. The Disruption: Generation, Mediation, Detection

3.1 What language models do to idiolect

A large language model is trained to predict the next token over vast corpora, and in doing so learns a distribution over language that is, by construction, an aggregate (Vaswani et al., 2017; Brown et al., 2020). Text sampled from that distribution exhibits style, but not idiolect: its regularities reflect training data and decoding parameters, not a biography. Bender and colleagues' (2021) characterisation of such systems as stochastic parrots was aimed at questions of meaning and harm, but it states the forensic problem exactly: the text has fluency without a speaker standing behind it in the idiolect-bearing sense. When a human author uses a model to draft, translate, or polish, the resulting document is a composite in which the human's orthographic and lexical habits, the levels forensic casework most relied upon, are the first casualties, while discourse-level intent and content knowledge survive best (Table 1). The evidential centre of gravity of authorship analysis is thus forced upward, from surface habit to content and circumstance, which is precisely where linguistic expertise blurs into ordinary investigation.

3.2 The detection industry and its two failure modes

The obvious response, detect the machine text, has produced a family of tools: perplexity-based detectors that flag text a model finds too predictable, trained classifiers, curvature methods such as DetectGPT that probe the local probability landscape (Mitchell et al., 2023), and cryptographic watermarking that biases generation toward a pseudo-random token set detectable later with a key (Kirchenbauer et al., 2023). Their performance under friendly conditions can be respectable. Their performance under adversarial or merely realistic conditions is the forensic issue, and here two failure modes dominate. The first is fragility: light paraphrasing, human or automated, sharply degrades detector accuracy, and theoretical analysis suggests that as model output distributions approach human text distributions, reliable detection may be unattainable in principle (Sadasivan et al., 2023). The second is bias: commercial detectors misclassified a majority of essays by non-native English speakers as machine-generated while classifying native-speaker essays as human with high reliability, because the constrained lexical and syntactic range of second-language writing mimics the low-perplexity signature detectors treat as machine evidence (Liang et al., 2023). Figure 1 displays the disparity.



Average across seven commercial GPT detectors; approximate values as reported by Liang et al. (2023). Detection is also fragile in the other direction: light paraphrasing of AI text sharply degrades detector accuracy (Sadasivan et al., 2023).

Figure 1. False-positive rates of commercial GPT detectors on human-written essays, by writer background (approximate means reported by Liang et al., 2023).

In a policing context the bias finding is not a technical footnote; it is a civil-rights finding. Any investigative or judicial use of detector scores against text written by second-language speakers, asylum applicants' statements, threat communications in multilingual communities, student misconduct cases, imports a documented, directional error against an identifiable class of writers. Under Daubert's known-error-rate prong, a tool whose error rate is not merely high but demographically structured should not survive scrutiny as substantive evidence (Daubert v. Merrell Dow Pharmaceuticals, 1993; Liang et al., 2023).

3.3 Watermarking and provenance

Watermarking inverts the problem: rather than detecting machine text after the fact, the generator marks its own output (Kirchenbauer et al., 2023). Where a watermark is present and the key available, detection is statistically principled, and provenance standards for synthetic media point the same direction. The forensic limits are structural, however. Watermarking is voluntary per model operator; open-weight models need not mark at all and can be used to launder marked text by paraphrase; and absence of a watermark is evidence of nothing. Provenance infrastructure will therefore stratify the evidential landscape rather than restore it: some machine text will be provably machine, while the unmarked remainder, precisely the text produced with intent to deceive, remains contested terrain (Sadasivan et al., 2023).

Table 2. Approaches to distinguishing human from machine text, and their forensic limitations.

Approach	Principle	Strength	Forensic limitation
Perplexity / probability-curvature detection	Machine text is unusually probable under a reference model (e.g., DetectGPT)	No training data needed; model-agnostic variants	Fragile under paraphrase; penalises constrained human styles (Mitchell et al., 2023; Sadasivan et al., 2023)
Trained classifiers	Supervised learning on human vs. machine corpora	Strong in-domain accuracy	Domain shift; documented bias against non-native writers (Liang et al., 2023)
Cryptographic watermarking	Generation biased toward keyed token set	Statistically principled when present	Voluntary; removable by paraphrase; absence proves nothing (Kirchenbauer et al., 2023)
Stylometric comparison	Compare questioned text to candidate authors' known writings	Retains validity for human-vs-human questions	Weakened where machine mediation dilutes idiolect (Stamatatos, 2009)
Provenance metadata	Signed content credentials at creation	Robust where ecosystem adopted	Coverage gaps; strategic non-adoption by bad actors

4. Reframing the Inference: Likelihood Ratios Under Machine Mediation

Forensic inference about authorship is best expressed, as Grant (2007) and the broader forensic-science consensus urge, as a likelihood ratio. For evidence E (the observed linguistic features of a questioned text) and competing hypotheses H_1 (the suspect wrote it) and H_2 (someone else wrote it), the analyst reports

$$LR = P(E \mid H_1) / P(E \mid H_2) \quad (1)$$

and the court, not the analyst, supplies prior odds. The LLM era does not overturn this framework; it adds a hypothesis. The partition of alternatives must now include H_3: the text was generated or substantially mediated by a language model, possibly at the suspect's instigation, possibly at another's. Ignoring H_3 inflates likelihood ratios whenever the observed features are ones a model would also produce, which, for fluent standard prose, is now most features. The practical consequence is that likelihood ratios computed on surface style must shrink toward 1 for machine-plausible texts, and analysts should say so explicitly.

Detection tools occupy a second equation. A detector's evidential value is its own likelihood ratio, P(score | machine) / P(score | human), and the perplexity statistic at the heart of most detectors,

$$PPL(w_1 \ldots w_N) = \exp\left(-(1/N) \sum_i \log p(w_i \mid w_{<i}) \right) \quad (2)$$

is computed against a reference model whose choice materially changes the score. The population baselines P(score | human) documented by Liang et al. (2023) differ by writer demographic, so a single global error rate misstates the evidence for any particular writer. A detector score offered in court without a demographic-conditional error analysis is, in likelihood-ratio terms, a number without a denominator. This is not a counsel of despair but of discipline: the same framework that governs DNA and fingerprint reporting can govern text evidence, provided the validation studies are done, and provided the analyst reports uncertainty about H_3 rather than absorbing it silently.

5. Distributed Authorship and the Law's Author-Function

Literary theory spent the twentieth century dismantling the romantic author, the solitary origin of a text's meaning, and replacing it with something more distributed: authorship as a function assigned by institutions, a role that organises attribution, authority, and liability rather than a natural fact about creation. The law, for good reasons, declined to follow; criminal and civil doctrine alike need a person to hold responsible, and the signature, the confession, the threatening letter are all legal instruments premised on unitary authorship. Machine mediation forces the theoretical question into practical dockets. Who is the author of a threat composed by a model at a human's prompting: the prompter, whose intent it embodies, or nobody, if intent cannot be proved from the text alone? Is a model-polished confession still in the suspect's own words, the traditional touchstone of voluntariness analysis, and what happens to register-based challenges of the kind forensic linguists mounted against police-authored statements when every register is

machine-imitable (Coulthard et al., 2017; Solan & Tiersma, 2005)?

Our answer, developed across the disciplines represented here, is that the law's target was never the text's surface but the intending principal behind it, and machine mediation, rightly understood, changes the evidence for intention rather than the concept. A threat is the prompter's act as surely as a typed letter was the typist-author's; what has changed is that the textual trace connecting act to actor has thinned. The investigative response is to relocate evidence of authorship from style to circumstance: prompt histories, account records, device forensics, the content-knowledge markers of Table 1 that machines cannot supply. The cultural-production context points the same way; analyses of contemporary media work describe creation as already a human-machine workflow in which the audience has become the creator and AI tools are ordinary instruments of composition (Alnajjar, 2024). Forensic doctrine should absorb that reality rather than litigate against it: machine mediation of text is now the base rate, not the anomaly.

A final interpretive complication is multilingual. Authorship analysis has always been hardest across languages, translated texts shed idiolect by design, and machine mediation now makes every second-language writer's situation ambiguous in a new way. A non-native English speaker who drafts in their first language and passes the draft through a model for translation or polish produces a text whose English surface belongs to the machine while its structure and content remain the author's; a detector will flag it (Liang et al., 2023), a stylometric comparison against their English writings will fail to match it, and both results are artefacts. The linguistically sound response is to move comparison into the language of composition where it can be identified, and to treat cross-language authorship claims with a scepticism proportional to how much machinery stands between author and artefact. For police services operating in multilingual societies, this is not an exotic scenario; it is the ordinary case, and training materials for investigators handling text evidence should say so explicitly.

6. Discussion: What Courts and Investigators Should Do

Figure 2. assembles the argument as a pipeline with marked points of disruption, and Table 3 translates it into admissibility terms. Three operational conclusions follow.

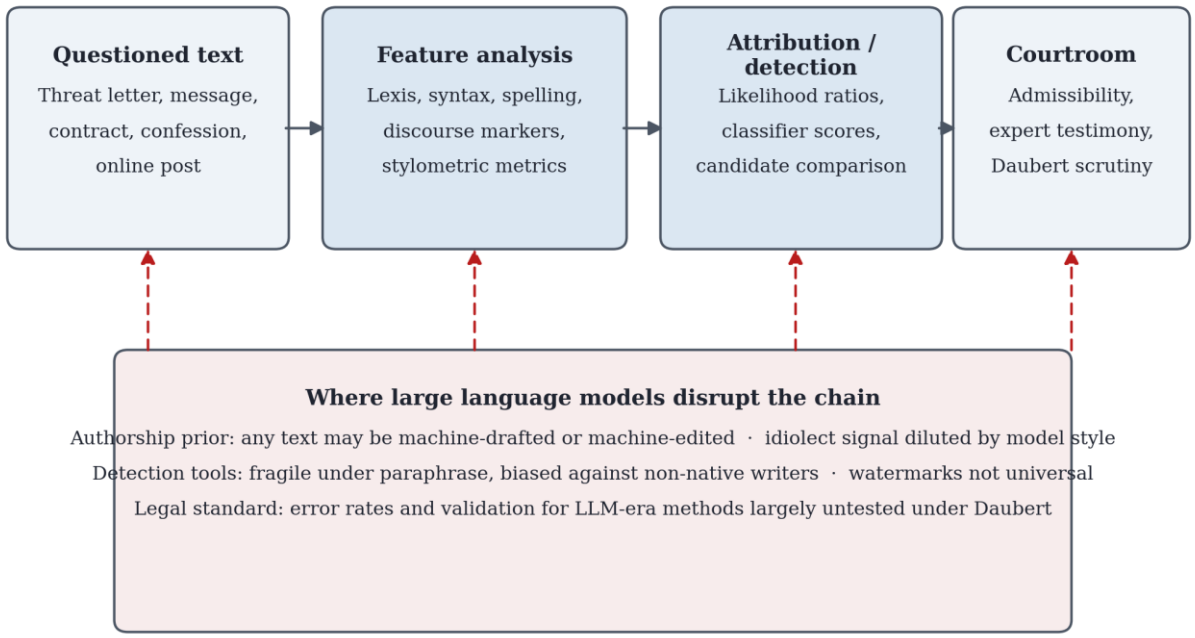


Figure 2. The forensic authorship pipeline and the points at which large language models disrupt it.

First, courts should distinguish sharply between authorship comparison and machine detection. Human-versus-human attribution on adequate known writings retains a validated methodological base (Juola, 2006; Stamatatos, 2009; Grant, 2007), and its likelihood-ratio reporting can be extended to carry machine-mediation uncertainty explicitly. Current machine-detection tools, by contrast, fail the known-error-rate prong for substantive evidential use: their error rates are unstable across domains and demographically structured (Liang et al., 2023; Sadasivan et al., 2023). Treating detector output as an investigative lead, a reason to seek prompt logs or

device evidence, respects both its usefulness and its unreliability.

Second, practice standards need renovating now, not after an appellate embarrassment. Proficiency testing for forensic linguists should include machine-mediated materials; laboratory validation of stylometric methods should report performance on model-assisted texts; and expert reports should state, as a matter of form, what was assumed about machine involvement and how conclusions change if that assumption fails. The forensic-science lesson of recent decades, that overstated certainty eventually detonates in re-examined convictions, applies with special force to a technology moving this fast (Solan & Tiersma, 2005; Daubert v. Merrell Dow Pharmaceuticals, 1993).

Third, the research agenda is tractable and urgent. It includes demographic-conditional validation of any detector proposed for legal use; study of hybrid texts, human-drafted, machine-polished, and the reverse, which dominate real practice; multilingual extension, since the bias evidence to date centres on English (Liang et al., 2023); and adversarial evaluation of watermark robustness in realistic laundering chains (Kirchenbauer et al., 2023; Sadasivan et al., 2023). Interdisciplinary venues matter here: the evidential questions cut across law, linguistics, and computing in the way this journal's forensic-science coverage, from novel biological detection methods (Rawwash, 2025) to the present problem, illustrates, and progress will come from teams, not disciplines.

Table 3. Daubert factors applied to authorship comparison and to AI-text detection, current state.

Daubert factor	Stylometric authorship comparison	AI-text detection tools
Testability	Yes; benchmarked tasks and protocols exist (Juola, 2006)	Yes in principle; adversarial conditions rarely tested by vendors
Peer review & publication	Established literature (Stamatatos, 2009; Grant, 2007)	Rapid preprint literature; commercial tools often unpublished
Known error rate	Reportable under controlled conditions; degrades with short/mixed texts	Unstable across domains; demographically biased (Liang et al., 2023)
Standards & controls	Emerging LR-based reporting standards	No accepted forensic standards; vendor thresholds opaque
General acceptance	Qualified acceptance within forensic linguistics	No acceptance in forensic community for evidential use

6.1 Limitations

Three limitations frame our conclusions. The technical landscape is moving faster than the citation cycle; the detection results we rely on describe the systems of the early LLM period, and both attack and defence will have evolved by the time of reading, though the structural arguments, aggregation versus idiolect, voluntariness of watermarking, demographic baselines, are not version-dependent. Our legal frame is common-law admissibility doctrine, and civil-law systems' free evaluation of evidence will metabolise these problems differently. And the empirical bias literature is thus far concentrated on English; the multilingual case, where most of the world's policing happens, is exactly the gap we urge others to fill.

7. Conclusion

Forensic linguistics built its authority on a fact about human beings, that we cannot help writing as ourselves, and that fact has not changed. What has changed is the population of writers: our texts now share the world with machine text and machine-mediated text, indistinguishable at the surface where the discipline did much of its work. The temptation is to meet the new uncertainty with new tools whose confidence outruns their validation; the record of detector bias against non-native writers shows where that road leads, and who pays its costs. The sounder path is older: state hypotheses, including the machine hypothesis, report likelihood ratios with honest denominators, validate on realistic materials, and let circumstance corroborate style rather than style masquerade as proof. The question in our title, whose words are these, will be asked of more texts, in more courtrooms, in more languages, every year from now on. The disciplines that share responsibility for answering it have the tools to answer honestly. They should not promise more.

Declarations

Conflict of interest. The authors declare no conflict of interest.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements. The authors thank colleagues at their respective departments for comments on an earlier draft.

References

Alnajjar, M. (2024). Shifting trends shaping content creation in media and newsrooms. Emirati Journal of Digital Art & Media, 2(1). Emirates Scholar Center for Research and Studies.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence, 4(1).

Coulthard, M., Johnson, A., & Wright, D. (2017). An introduction to forensic linguistics: Language in evidence (2nd ed.). Routledge.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. International Journal of Speech, Language and the Law, 14(1), 1–25.

Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233–334.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 17061–17084.

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

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 24950–24962.

Rawwash, A. A. (2025). The role of honeybees in forensic evidence detection: Review of their application in environmental and criminal investigations. Emirati Journal of Law and Policing Studies, 1(1), 4–16. Emirates Scholar Center for Research and Studies. https://doi.org/10.54878/wafkw950

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156.

Solan, L. M., & Tiersma, P. M. (2005). Speaking of crime: The language of criminal justice. University of Chicago Press.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556. https://doi.org/10.1002/asi.21001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.