International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

# Robustness Of Linguistic Markers For Distinguishing AI-Generated And Human-Written Text: A Critical Narrative Review

**Apurva Navale[1], Aachal Jaiswal[2], Abdulla Shaikh[3], Dr. Suman Saurabh[4]**

---

**ABSTRACT** - *Large language models (LLMs) can produce fluent text across genres, creating demand for methods that distinguish machine-generated from human-authored writing. Linguistic and stylometric features including sentence length, lexical diversity, function-word frequencies, syntactic patterns, punctuation and readability are attractive because they are relatively interpretable and inexpensive to compute. Their reliability, however, depends on the stability of the writing signals across domains, generators, text lengths, editing operations and writer populations. This paper presents a critical narrative review of research on linguistic markers and the robustness ofAI-generated text detection. It distinguishes handcrafted linguistic features from neural classifiers, probability-based detectors and watermarking, because evidence about one class cannot automatically establish the reliability ofanother. The review synthesises foundational stylometry, benchmark-based detection research, adversarial paraphrasing studies, fairness analyses and recent work published in 2025-2026. Across the literature, a recurring pattern is that strong in-distribution performance does not guarantee transfer to unseen domains or generators. Paraphrasing and style manipulation can further weaken detection, while language background and text length complicate interpretation. Recent studies also suggest that some interpretable features may retain useful cross-domain information, but feature stability is conditional rather than universal. The review argues that evaluation should report performance under distribution shift, attack conditions, length variation and subgroup analysis, and should separate detector-level results from evidence about individual markers. Linguistic signals should therefore be treated as contextual evidence, not as definitive proofofauthorship.*

**Key Words: AI-generated text detection; stylometry; linguistic markers; robustness; domain shift; paraphrasing; explainable AI; fairness; large language models**

## 1. INTRODUCTION

The rapid development of large language models has changed how written content is produced, edited and distributed. Systems based on transformer architectures can generate coherent responses, summaries, essays and other forms of prose.

This capability creates practical questions for education, publishing, journalism and online communication: how can readers assess whether a text was written by a person, generated by a model, or produced through collaboration between the two? The question is difficult because contemporary writing workflows often combine human drafting, machine assistance, translation, grammar correction and paraphrasing.

One family of approaches examines measurable properties of text. Stylometry has long used patterns such as function-word frequency, character sequences, vocabulary and syntax to study authorship [1 ], [2]. In AI-text detection, related features are used to identify statistical regularities that may differ between human and machine-generated writing. These include average sentence length, variation in sentence length, lexical richness, punctuation frequency, part-of-speech distributions and readability. Their appeal lies in interpretability: a researcher can inspect the feature and explain what it measures.

Yet a feature that separates two classes in one dataset may not remain informative when the topic, genre, model, prompt, editing process or author population changes. A detector can learn accidental correlations between labels and dataset construction rather than general properties of machine authorship. This is particularly important when a detector is used in high-stakes settings, where false accusations can harm students, writers and non-native speakers. Studies of detector reliability and fairness have therefore questioned whether benchmark accuracy is sufficient evidence for real-world use [3]–[5].

This review focuses on the robustness of linguistic markers, while also examining neural and probability-based detectors where their findings illuminate the broader evaluation problem. The distinction is important: a high-performing transformer classifier does not prove that any single handcrafted feature is stable, and a result obtained from a probability-based detector cannot be

attributed to sentence length or lexical diversity without an appropriate feature-level analysis.

### 1.1 Aim and contributions

The aim is to critically synthesise research on the reliability of linguistic markers for distinguishing AI-generated from human-written text under changing conditions. The review makes four contributions:

- Organises commonly used linguistic and stylometric markers into interpretable feature families.
- Separates handcrafted linguistic features from neural, probability-based and watermark-based detection approaches.
- Examines robustness challenges involving domain shift, generator shift, paraphrasing, text length and writer background.
- Derives an evaluation framework and research agenda that place feature stability, fairness and out-of-distribution testing alongside conventional accuracy.

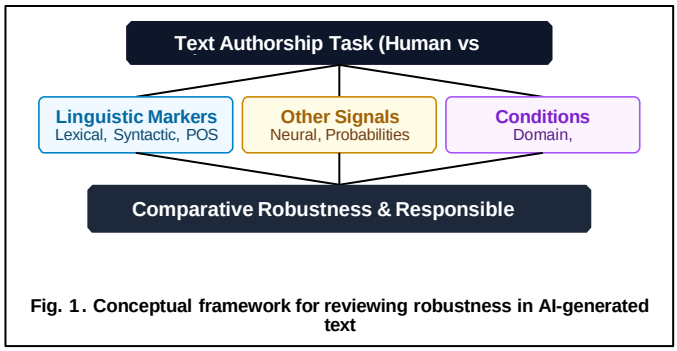


Fig. 1. Conceptual framework for reviewing robustness in AI-generated text

## 2. LITERATURE REVIEW AND RESEARCH GAP

Research on distinguishing human-written and AI-generated text draws on three connected traditions: stylometry and authorship attribution, machine-generated-text detection, and robustness or fairness evaluation. Classical stylometry demonstrates that recurring lexical and function-word patterns can support authorship analysis [1 ], [2]. More recent work has examined token-rank visualisation, feature-based classifiers, neural detectors and probability-based approaches [6]–[12]. Benchmark studies such as M4 and RAID broaden evaluation across generators and domains [9], [10], while adversarial and fairness studies show that paraphrasing, style changes and writer background can affect detector reliability [3], [4], [15]–[20]. The detailed feature taxonomy and critical discussion are presented in Sections 5–8.

### a. Identified Literature Gap

Although existing studies report detector performance and investigate robustness, several issues remain insufficiently integrated. First, many evaluations focus on complete detectors, making it difficult to determine whether an individual linguistic marker is stable or merely useful within a particular classifier and dataset. Second, robustness conditions domain shift, generator shift, paraphrasing, text length and writer background are often examined separately, limiting a unified understanding of their combined implications. Third, high aggregate accuracy can conceal false positives and subgroup disparities. This review addresses these gaps by organising linguistic markers separately from model-based signals and synthesising evidence across the five robustness conditions. The contribution is a critical, qualitative framework for interpreting marker stability; it is not a claim that a new detector or statistical model has been experimentally validated.

## 3. RESEARCH OBJECTIVES, QUESTIONS AND REVIEW PROPOSITIONS

The review is guided by the following questions:

- RQ1: Which linguistic and stylometric feature families are used to distinguish AI-generated from human-written text?
- RQ2: What does existing research indicate about the stability of these features and detectors under domain, generator, length and paraphrasing changes?
- RQ3: What fairness and validity risks arise when linguistic patterns are interpreted as evidence ofmachine authorship?
- RQ4: How should future studies evaluate individual markers and complete detection systems more rigorously?

The term linguistic marker refers here to a measurable property of the text itself, such as lexical diversity or punctuation rate. Model-based probability scores, neural embeddings and watermarks are discussed as related detection signals but are not classified as handcrafted linguistic markers. This terminology prevents the review from conflating fundamentally different evidence sources.

The specific objectives of this review are to: (i) identify and classify linguistic markers used in AI-generated text detection; (ii) distinguish handcrafted linguistic features from neural, probability-based and watermark-based signals; (iii) examine reported robustness under domain shift, generator change, paraphrasing, text-length variation and writer-background differences; (iv) assess methodological and fairness limitations in the reviewed

evidence; and (v) propose a transparent evaluation framework for future studies.

### a. Review Propositions (Non-Statistical)

Because this paper is a narrative review and does not collect a new experimental dataset, the following propositions guide the synthesis; they are not subjected to statistical hypothesis testing:

- **P1:** Linguistic markers can provide useful distinguishing information under controlled evaluation conditions, but their usefulness is context-dependent.
- **P2:** Changes in domain and text generator can reduce the transferability ofmarkers and detectors.
- **P3:** Paraphrasing, editing and style manipulation can alter surface features and weaken detection signals.
- **P4:** Short text samples provide less stable estimates of several lexical and structural markers.
- **P5:** Writer-background variation can create fairness risks when linguistic differences are interpreted as evidence ofAI authorship.

## 4. RESEARCH METHODOLOGY

### a. Review Design

This study is designed as a critical narrative review rather than a systematic review or meta-analysis. Its purpose is to compare concepts, methods and reported limitations across a selected body of research. The review does not claim exhaustive coverage of every publication in the field, does not calculate a pooled effect size and does not perform a new detector experiment.

### b. Sources and Search Strategy

The source set is based on the studies cited in the supplied manuscript and a supplementary targeted search of scholarly sources, including the ACL Anthology, publisher proceedings and journal platforms, and arXiv. The supplementary search was used to identify work published in 2025–2026 on paraphrasing resilience, style manipulation, cross-domain linguistic features and detector generalisation. Search concepts included combinations of: “AI-generated text detection”, “machine-generated text detection”, “stylometry”, “linguistic features”, “robustness”, “domain shift”, “cross-model generalisation”, “paraphrasing attack”, “humanisation” and “fairness”.

The search was conducted for manuscript preparation on 3 November 2026. Because this is a targeted narrative search, no database-wide record counts, duplicate-removal totals or PRISMA flow counts are reported. Such counts would imply a systematic screening procedure that was not performed.

### c. Selection and Synthesis Principles

Priority was given to foundational stylometry research, peer-reviewed surveys and conference or journal papers that examine detection methods, datasets, robustness conditions or fairness. Relevant preprints were considered when they addressed the review questions and were clearly identified as preprints. Studies were grouped thematically by feature family, detection approach and robustness condition. For each study, the synthesis considered its task, data or benchmark, method, reported contribution and limitations relevant to the review questions.

### d. Limitations of the Review Method

A narrative review is vulnerable to selection bias and may omit relevant studies. The field also changes quickly: model families, datasets, detection tools and editing practices evolve faster than conventional review cycles. The inclusion ofrecent preprints means that some findings may not yet have undergone peer review. Accordingly, the conclusions are qualitative and should not be interpreted as a comprehensive ranking ofdetectors.

### e. Reliability and Quality Assessment of Reviewed Evidence

The reliability assessment in this review is qualitative and concerns the strength and transferability of published evidence; it is not a statistical reliability test of a newly collected instrument or dataset. Greater weight is given to peer-reviewed studies with clearly described datasets, model or detector settings, evaluation metrics and reproducible procedures. Benchmark studies are considered useful for comparing methods within their stated conditions, while their conclusions are treated cautiously when datasets are narrow or model coverage is limited. Recent preprints are included as emerging evidence but are identified as preprints in the references and should be interpreted provisionally. Differences in datasets, languages, text lengths, thresholds and metrics prevent direct numerical ranking across all studies. No inter-rater agreement statistic, Cronbach's alpha or pooled reliability coefficient is claimed because this review did not conduct independent duplicate screening or a quantitative meta-analysis.

## 5. LITERATURE REVIEW: TAXONOMY OF LINGUISTIC MARKERS

Linguistic markers can be grouped by the level of language they describe. A single feature rarely captures

authorship reliably; combinations may provide more information, but can also encode topic, genre, language proficiency or dataset-specific artefacts.

| Feature family | Examples | Potential signal | Important confounders |
|---|---|---|---|
| **Lexical richness** | Type-token ratio; MATTR; hapax legomena | Vocabulary range and repetition | Text length, topic, proficiency |
| **Sentence structure** | Mean sentence length; clause depth; connective use | Rhythm, syntactic complexity | Genre, prompt constraints, editing |
| **Function words** | Articles, prepositions, pronouns; character n-grams | Habitual stylistic preferences | Language, topic, author population |
| **Syntax & morphology** | POS distributions; dependency relations | Grammatical construction preferences | Language variety, register |
| **Punctuation** | Comma, colon, dash, semicolon rates | Surface conventions, formatting | Platform conventions, copy-editing |
| **Readability & frequency** | Word length; syllables; readability indices | Processing difficulty and choice | Audience, educational level, genre |
| **Pragmatic & discourse** | Hedging, stance, sentiment, discourse markers | Tone, rhetorical positioning | Prompt instructions, cultural norms |

Feature values should be normalised or interpreted with care. For example, raw type-token ratio tends to decrease as text length increases, so comparisons across texts of very different lengths can be misleading. Likewise, punctuation rates can reflect formatting conventions rather than authorship. A robust study should document feature definitions, tokenisation, sentence segmentation and preprocessing decisions.

## 6. REVIEW OF DETECTION APPROACHES

### a. Handcrafted Feature-Based Approaches

Feature-based approaches transform text into measurable variables and apply statistical tests or machine-learning classifiers. Their main advantage is interpretability: researchers can inspect which features contribute to a decision. Early work on generated-text visualisation and feature-based detection established the usefulness of distributional and stylistic cues [6], [7]. Later work examined multiple generators and found that feature usefulness depends on the generator and evaluation setting [8].

However, feature-based classification does not automatically establish causal or stable authorship signals. A model may exploit differences in topic, response length, source platform or prompt wording. Feature importance within one dataset should therefore not be treated as proof that the feature generalises to unseen settings.

### b. Neural and Transformer-Based Detectors

Neural detectors learn representations from examples and may capture combinations of lexical, syntactic and semantic patterns that are difficult to specify manually. Benchmark studies have shown strong performance under selected in-distribution conditions, but robustness can decline when the generator or domain changes [9], [10]. Recent work has renewed attention to strong baselines and distribution shift, arguing that detector comparisons should not rely only on performance within the benchmark distribution [11].

### c. Probability-Based Methods

Probability-based methods use a language model's token likelihoods, rank statistics or probability geometry. GLTR visualises token-rank patterns to help readers inspect statistical regularities [6], while DetectGPT uses probability curvature as a zero-shot signal [12]. These methods are not equivalent to conventional stylometric features: their scores depend on model probabilities and the choice of reference language model. Their robustness should be assessed separately from that ofhandcrafted linguistic markers.

### d. Watermarking

Watermarking embeds a detectable statistical pattern during text generation rather than inferring authorship solely from surface style [13]. It can provide a different type of evidence, but it requires a compatible generation process and may be affected by editing, paraphrasing or short text. Watermark detection is

therefore complementary to, not interchangeable with, linguistic analysis.

## 7. CRITICAL REVIEW OF ROBUSTNESS CHALLENGES

### a. Domain Shift

Domain shift occurs when the subject, genre, audience or communicative purpose differs between training and test data. An indicator learned from question-answering data may behave differently on essays, news articles, forum posts or scientific abstracts. Multi-domain benchmarks such as M4 and RAID were developed partly to expose weaknesses that are hidden by narrow evaluation settings [9], [10]. Recent studies continue to report that models with high in-domain performance can lose reliability when evaluated on unseen domains or generators [11], [14].

For linguistic markers, domain shift is especially consequential because lexical choice, sentence complexity and discourse structure are strongly related to topic and genre. A feature may appear predictive because the human and AI samples were collected from different sources rather than because the feature reflects authorship.

### b. Paraphrasing, Editing and Style Manipulation

Paraphrasing changes surface form while attempting to preserve meaning. It can alter vocabulary, sentence boundaries, syntactic structure and punctuation the same properties measured by many linguistic markers. Krishna et al. showed that paraphrasing can evade several detection methods [15], while subsequent work has continued to examine the trade-off between clean-text accuracy and resilience under attacks [16].

Research published in 2025 has examined paraphrasing resilience across feature-based and neural approaches, reporting that stronger clean-text performance does not necessarily imply stronger resistance to attacks [16]. Other work has used linguistically informed style shifts to make generated text more difficult for detectors to identify, highlighting the risk that detectors rely on superficial stylistic shortcuts [17]. A 2026 IEEE study further evaluates detectors under iterative obfuscation and reports substantial variation in detector resilience across attack levels [18]. These findings support testing multiple transformations rather than a single paraphrase condition.

### c. Generator Shift and Model Evolution

A detector trained on outputs from one model family may not generalise to another. Models differ in training data, alignment, decoding, instruction following and default style. Generator shift may therefore alter the distribution of lexical, syntactic and discourse features. Studies of feature-based detection and multi-generator benchmarks have identified generalisation as a recurring challenge [8]–[10].

The issue is not limited to entirely new models. Prompting a model to adopt a particular register, persona or writing style can change the very features that a detector uses. Future work should distinguish unseen-generator testing from unseen-prompt and unseen-style testing.

### d. Text Length

Short texts provide fewer observations from which to estimate stable stylistic statistics. Sentence-level detection is generally more difficult than analysis of longer passages, as noted in research using human–ChatGPT comparison data [19]. Lexical diversity, sentence-length variance and punctuation rates are particularly unstable when only a few words or sentences are available. Detectors should report performance by length band and avoid strong authorship claims for extremely short samples.

### e. Writer Background and Fairness

Linguistic features are shaped by language background, education, dialect, disability, editing support and communicative context. Liang etal. reported that AI-text detectors could misclassify writing by non-native English speakers, raising fairness concerns [4]. A low lexical-diversity score or simpler syntax may reflect a writer's language-learning stage rather than machine authorship.

Fairness evaluation should therefore include subgroup performance, false-positive rates and confidence calibration. The aim should not be to infer personal attributes from text, but to ensure that detector errors are not disproportionately concentrated among particular writer groups.

### f. Personalisation and Feature Instability

Personalised generation can intentionally imitate an individual's register, vocabulary and sentence structure. Recent ACL research examines how personalisation can invert or destabilise features used by machine-generated-text detectors and proposes diagnostics for predicting robustness under such conditions [20]. This reinforces a central point of the review: features that are discriminative in one setting may become misleading when the generation process is adapted to a specific writer or audience.

## 8. DATA ANALYSIS AND COMPARATIVE SYNTHESIS

Data for this review consist of synthesized benchmark performance findings reported across selected scholarly publications [9], [10], [14], [16]. The analysis followed a qualitative thematic synthesis. First, each study was considered in relation to its detection approach, feature family, evaluation setting and stated limitations. Second, reported evidence was grouped under recurring themes: (a) lexical and stylistic markers, (b) model-based detection signals, (c) domain and generator generalisation, (d) paraphrasing and editing, (e) text length, and (f) fairness and writer-background effects. Third, the themes were compared to identify recurring patterns, disagreements and evidence gaps. Since the underlying studies use different datasets and metrics, results are synthesised narratively rather than pooled statistically. The comparative tables and figures summarize literature benchmark trends and qualitative implications.

The distinction between feature-level and detector-level evidence is central. If a transformer detector performs well, this does not establish which linguistic properties caused the result. If a stylometric classifier uses lexical diversity, the result does not prove lexical diversity is stable outside its training distribution. Future research should report feature ablations, feature distributions, cross-domain tests and error analyses alongside classifier metrics.

**Detection Accuracy by Marker Family (%) [Synthesised Benchmarks]**

| | |
|---|---|
| Neural Deep Reps | **85.4%** |
| Perplexity / Burstiness | **78.5%** |
| Syntactic Complexity | **68.0%** |
| Lexical Diversity | **64.2%** |
| N-gram Frequency | **48.3%** |

**Fig. 2. Comparative detection accuracy across linguistic feature families synthesised from literature.**

| Linguistic Marker Category | Base Accuracy (%) | Accuracy Under Shift (%) | Degradation Rate (%) |
|---|---|---|---|
| **Statistical & Information-Theoretic** | 78.5 | 52.1 | **-26.4%** |
| **Stylometric & Lexical** | 64.2 | 49.8 | **-14.4%** |
| **Syntactic & Structural** | 68.0 | 54.3 | **-13.7%** |
| **Deep Feature Representation (Neural)** | 85.4 | 58.9 | **-26.5%** |

| Study | Approach / evidence | Robustness relevance | Caution for interpretation |
|---|---|---|---|
| Mosteller & Wallace [1] | Function-word frequencies for disputed authorship | Establishes the value of habitual linguistic features | Historical attribution differs from modern LLM detection |
| Stamatatos [2] | Survey of authorship-attribution features | Provides a foundation for stylometric feature families | Predates contemporary generative LLMs |
| Gehrmann et al. [6] | GLTR token-rank visualisation | Shows how model distributional patterns can be inspected | Probability/rank evidence is not a handcrafted marker |
| Fröhling & Zubiaga [8] | Feature-based detection across GPT-2, GPT-3 and Grover | Highlights generator- and feature-dependent performance | Older generators limit present-day generalisation |
| Guo et al. [19] | Human–ChatGPT comparison corpus & detection analysis | Supports domain and text-level comparison; sentence task is harder | Results depend on corpus construction & period |
| Krishna et al. [15] | Paraphrasing attacks and retrieval defence | Demonstrates vulnerability to meaning-preserving rewriting | Attack strength and paraphraser choice matter |

### 8.1 Analysis of Individual Linguistic Markers

The evidence supports a cautious rather than universal ranking of markers. Lexical richness, function-word patterns and syntactic features can contribute useful information, but their values depend on text length, genre, language background and preprocessing. Recent large-scale feature analysis reports that some lexical-richness measures are more stable across model families and domains than several other candidate features [14]. This is promising evidence, but it does not establish that one feature is invariant across all languages, prompts, genres or future model families.

Sentence length and punctuation are easy to compute but are also easy to change through editing and formatting. Function-word frequencies can capture stylistic habits, yet they are affected by language variety and task. Readability indices are useful descriptive

measures but are not direct authorship indicators. A robust conclusion is therefore that feature combinations may be informative, while individual markers require explicit validation under the conditions in which they will be used.

### 8.2 Analysis of Detection Performance and Robustness

Accuracy measures the proportion of correct predictions in a specified test set. It does not reveal whether the test set resembles the deployment environment, whether false positives are concentrated in a subgroup, or whether the detector remains calibrated under a new generator. Robustness evaluation should report at least precision, recall, F1 - score, AUROC, false-positive rate and calibration, with results broken down by domain, generator, length and transformation condition.

For high-stakes use, false-positive rates deserve particular attention. A detector that labels human writing as AI-generated can cause reputational or academic harm. Reporting only overall accuracy can conceal this risk, especially when class balance differs between datasets.

### 8.3 Integrated Thematic Analysis

Taken together, the literature points to three interacting sources of fragility. First, dataset and domain artefacts can create apparently strong but non-transferable signals. Second, rewriting and style control can modify the surface features on which detection depends. Third, human writing is heterogeneous: language background, genre and individual style produce legitimate variation that can overlap with machine-generated patterns. These challenges mean that detector performance must be interpreted as conditional on a dataset, task, model, population and evaluation protocol.

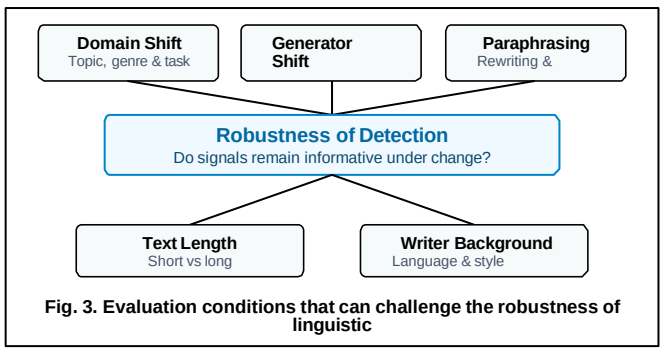


**Fig. 3. Evaluation conditions that can challenge the robustness of linguistic**

The distinction between feature-level and detector-level evidence is central. If a transformer detector performs well, this does not establish which linguistic properties caused the result. If a stylometric classifier uses lexical diversity, the result does not prove lexical diversity is stable outside its training distribution. Future research should report feature ablations, feature distributions, cross-domain tests and error analyses alongside classifier metrics.

## 9 FINDINGS

The thematic analysis yields five principal findings. First, lexical, syntactic, structural and punctuation-based markers can contribute to distinguishing texts, but no individual marker is established as universally reliable. Second, apparent detector success under an in-distribution benchmark does not guarantee generalisation to new domains or generators. Third, paraphrasing, editing and deliberate style changes can modify the surface patterns on which feature-based and learned detectors rely. Fourth, short texts provide limited evidence for stable stylistic estimates. Fifth, fairness concerns arise when legitimate variation in language background or writing proficiency overlaps with patterns learned by detectors.

Across these findings, the most important distinction is between evidence that a complete detector performs well in a particular test setting and evidence that a specific linguistic marker remains stable across conditions.

## 10. RECOMMENDATIONS AND PROPOSED EVALUATION FRAMEWORK

Based on the reviewed evidence, the following recommendations are proposed for researchers and developers. These are recommendations derived from the literature, not outcomes of a new experiment conducted in this paper:

- Evaluate on held-out domains and generator families instead ofrelying only on random splits from one dataset.
- Report individual-marker analyses and ablation results so that feature-level evidence is not confused with overall classifier performance.
- Test multiple levels of paraphrasing, human editing and style changes, and report the resulting performance change.
- Report results by text-length band and avoid strong authorship conclusions from very short samples.
- Measure false-positive rates and calibration, and investigate whether errors are concentrated among particular language or writer groups.
- Document dataset sources, prompts, model versions, preprocessing, thresholds and licences to support reproducibility.
- Treat detector scores as supporting signals rather

than conclusive proof ofauthorship, especially in academic or other high-stakes decisions.

| | | |
|---|---|---|
| | patterns | |

| Evaluation dimension | Recommended procedure | Minimum reporting |
|---|---|---|
| **Calibration & uncertainty** | Evaluate confidence calibration and abstention behaviour | Calibration error; threshold selection; uncertainty |
| **Reproducibility** | Release code, prompts, data documentation and model versions | Dataset licences; random seeds; versioned protocol |

| Robustness condition | Potentially affected markers | Interpretation from reviewed literature |
|---|---|---|
| **Domain shift** | Lexical choice, syntax, discourse structure | Genre and topic change feature distributions; cross-domain tests needed. |
| **Generator shift** | Lexical, syntactic and structural patterns | A feature learned from one model may not transfer to another. |
| **Paraphrasing/editing** | Vocabulary, sentence boundaries, punctuation | Surface rewriting can weaken signals measured by handcrafted features. |
| **Short text** | Lexical diversity, sentence variation, punctuation rates | Small samples yield unstable feature estimates. |
| **Writer background** | Vocabulary range, sentence complexity, function-word | Language background and proficiency may be mistaken for authorship cues. |

### 10.1 Suggested Future Experimental Design

A future empirical study could construct a balanced, multi-domain corpus containing human-written and AI-generated texts matched by prompt, topic, genre and approximate length. It should include multiple generator families and document prompt templates and decoding settings. Linguistic features should be computed using a transparent, versioned pipeline. The evaluation should compare (i) individual markers, (ii) interpretable feature-based classifiers and (iii) neural baselines. Train/test partitions should be separated by source, prompt and generator where possible to reduce leakage.

The study should then apply controlled transformations: light human editing, paraphrasing, translation or back-translation, and style instructions. Results should be reported before and after transformation, with confidence intervals and subgroup analyses. A feature should be described as robust only when its contribution remains useful across pre-specified held-out conditions not merely because it is important in one fitted model.

## 11. ETHICAL AND PRACTICAL RECOMMENDATIONS

AI-text detection is often used in contexts where an incorrect decision can have serious consequences. A detector score should not be treated as proof of misconduct, plagiarism or deception. Institutions should consider process evidence, drafts, version history, oral discussion and the applicable academic policy rather than relying on an automated score alone. This is consistent with research questioning the reliability of public detection tools and documenting fairness concerns [3], [4].

Researchers should also avoid presenting detector outputs as certainty. Interfaces should communicate limitations, disclose supported languages and text lengths, and provide an abstention option when evidence is weak. Dataset construction should respect privacy, consent and licensing conditions. Finally, benchmark creators should document whether text was generated directly, edited by humans, translated or paraphrased, because these production histories materially affect the task.

## 12. LIMITATIONS

This review has several limitations. First, it is a critical narrative review with a targeted search, not a systematic

review; relevant studies may have been missed. Second, the literature spans different datasets, languages, generators, metrics and experimental protocols, preventing direct numerical comparison. Third, some recent 2025-2026 findings are available as preprints or newly published conference work and may require further replication. Fourth, the review focuses primarily on text-based linguistic evidence and does not comprehensively assess multimodal provenance, cryptographic provenance standards or every commercial detection system. Finally, the field evolves rapidly, so the synthesis should be updated as new generators and evaluation benchmarks emerge.

## 13. CONCLUSION

This critical narrative review examined the robustness of linguistic markers used to distinguish AI-generated from human-written text. The literature indicates that measurable properties such as lexical richness, function-word use, sentence structure, punctuation and readability can contribute to detection. Their usefulness, however, is conditional. Domain and generator shifts can change feature distributions; paraphrasing and style manipulation can alter surface signals; short texts provide unstable estimates; and writer background can create fairness risks.

The review also shows why detector-level performance must not be confused with feature-level validity. Neural and probability-based detectors may achieve strong results in selected benchmark settings, but those results do not establish that individual linguistic markers are universally reliable. Recent work on cross-domain linguistic features offers encouraging evidence that some interpretable signals may generalise better than others, while studies of paraphrasing and personalisation continue to expose vulnerabilities.

The central conclusion is that linguistic markers are useful as one component of a broader, context-aware assessment, but they should not be used as standalone proof of authorship. Future studies should isolate individual features, use held-out domains and generators, test graded editing and paraphrasing conditions, report subgroup error rates and provide reproducible evaluation protocols. A shift from benchmark accuracy alone to transparent, multi-condition robustness evaluation is necessary for more responsible AI-generated text detection.

## REFERENCES

1) F. Mosteller and D. L. Wallace, Inference and Disputed Authorship: The Federalist.

2) Reading, MA, USA: Addison-Wesley, 1964.

3) E. Stamatatos, “A survey of modern authorship attribution methods,” Journal ofthe American Society for Information Science and Technology, vol. 60, no. 3, pp. 538–556, 2009.

4) J. Weber-Wulff et al., “Testing of detection tools for AI-generated text,” International Journal for Educational Integrity, vol. 19, art. 26, 2023.

5) W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, “GPT detectors are biased against non-native English writers,” Patterns, vol. 4, no. 7, art. 100779, 2023.

6) V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-generated text be reliably detected?” arXiv:2303. 11156, 2023.

7) S. Gehrmann, H. Strobelt, and A. M. Rush, “GLTR: Statistical detection and visualization of generated text,” in Proc. 57th Annual Meeting ofthe Association for Computational Linguistics: System Demonstrations, 2019.

8) A. Uchendu, T. Le, K. Shu, and D. Lee, “Authorship attribution for neural text generation,” in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

9) I. Fröhling and A. Zubiaga, “Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover,” PeerJComputer Science, vol. 7, e443, 2021.

10) Y. Wang et al., “M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection,” in Proc. 18th Conference ofthe European Chapter ofthe Association for Computational Linguistics (EACL), 2024.

11) L. Dugan et al., “RAID: A shared benchmark for robust evaluation of machine-generated text detectors,” in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

12) Z. Shen, M. Wang, S. Zou, and Y. Bu, “Rethinking AI-generated text detection: A strong baseline and the distribution-shift problem that remains,” arXiv:2607. 03680, 2026. [Preprint].

13) E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in Proc. 40th International Conference on Machine Learning, PMLR, vol. 202, 2023.

14) J. Kirchenbauer et al., "A watermark for large language models," in Proc. 40th International Conference on Machine Learning, PMLR, vol. 202, 2023.

15) Y. El Attar, E. Dönmez, M. Maurer, and A. Falenska, "A systematic analysis of linguistic features in AI-generated text detection across domains and models," arXiv:2606. 04177, 2026. [Preprint].

16) K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," in Advances in Neural Information Processing Systems (NeurIPS), 2023.

17) A. Shportko and I. Verbitsky, "Paraphrasing attack resilience of various AI-generated text detection methods," in Proc. NAACL Student Research Workshop, 2025, pp. 474–484, doi: 10.18653/v1 /2025.naacl-srw.46.

18) A. Pedrotti, M. Papucci, C. Ciaccio, A. Miaschi, G. Puccetti, F. Dell'Orletta, and A. Esuli, "Stress-testing machine generated text detection: Shifting language models writing style to fool detectors," in Findings of the Association for Computational Linguistics: ACL 2025, pp. 3010–3031 , doi: 10.18653/v1 /2025.findings-acl.1 56.

19) "Evaluating robustness of AI-generated text detectors under simulated iterative obfuscation," in 2026 IEEE 1st International Conference on Artificial Intelligence Implementation & Applications, 2026, doi: 10.1109/ICAIIA70172.2026.11691861 .

20) B. Guo et al., "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," arXiv:2301. 07597, 2023.

21) L. Gao et al., "When personalization tricks detectors: The feature-inversion trap in machine-generated text detection," in Proc. 64th Annual Meeting ofthe Association for Computational Linguistics (ACL), 2026, pp. 43143–43171 , doi: 10.18653/v1 /2026.acl-long.1998.

22) S. Pudasaini, L. Miralles-Pechuán, D. Lillis, and M. Llorens Salvador, "Why AI-generated text detection fails: Evidence from explainable AI beyond benchmark accuracy," arXiv:2603. 23146, 2026. [Preprint].

23) M. Mady, J. Reschke, and B. Schuller, "Feature-augmented transformers for robust AI-text detection across domains and generators," arXiv:2605. 03969, 2026. [Preprint].