DOI: 10.2478/doc-2026-0003

Rania Za'rour

Department of English Language and Literature, The University of Jordan, Amman, Jordan

r.zarour@ju.edu.jo

ORCID ID: 0000-0001-9873-6099

Abdel Rahman Mitib Altakhaineh

Department of English Language and Literature, The University of Jordan, Amman, Jordan

a.altakhaineh@ju.edu.jo

ORCID ID: 0000-0001-7605-2497

False Positive Risk in AI Detection of L2 Academic Writing: A Case Study

Article history:

Received	**16 March 2026**
Revised	**25 April 2026**
Accepted	**26 April 2026**
Available online	**23 June 2026**

Abstract: AI-text detectors are increasingly used in education and scholarly publishing. While these tools are meant to protect academic integrity, they may also create new inequities for multilingual writers. This exploratory case study investigates why legitimate second language (L2) academic prose can be misclassified as AI-generated. We start from the premise that L2 academic writing functions as a distinct register whose recurrent linguistic features create predictability, which is one of the main metrics to detect AI-generated text. Using a single-author corpus repeatedly flagged as machine--generated, we develop a detector-inspired simulation capturing five proxies of high predictability: low perplexity, low burstiness, high formulaic language, consistent formality, and low lexical diversity. We simulated two common detection models: a prevalence model (flagging frequent proxies) and a cluster--based model (flagging co-occurrence of features). Prevalence-based detection flagged 62% of the sample manuscripts, while cluster-based detection flagged 38% with three or more co-occurring proxies. Although dense clustering (four to five proxies) was uncommon (7.8%), our sample was repeatedly flagged by both logics. These findings do not establish equivalence with commercial detection systems, but they suggest that highly conventionalized L2 academic prose may face a heightened risk of false positive AI flagging. We term this phenomenon the "detection paradox".

Keywords: AI detection, academic stratification, L2 writing, multilingual authors, linguistic bias

Introduction

AI-generated text detection is increasingly used in editorial workflows, academic integrity procedures, and writing assessment, shaping desk decisions and author feedback (Kehkashan et al., 2025). AI-generated text detection refers to automated procedures that estimate the likelihood that a text was produced by a large language model rather than a human. Detector outputs are, nevertheless, probabilistic and sensitive to genre, writer background, and text preprocessing (Liang et al., 2023; Weber-Wulff et al., 2023; Wang et al., 2024). The operational adoption of detection tools grants them practical authority, even as concerns about their generalizability and fairness persist in the literature (Abd-Elaal et al., 2022; Giray et al., 2025; Kehkashan et al., 2025). Because institutional workflows amplify the consequences of automated signals, false positives carry higher operational risk (Giray, 2024), particularly for second-language English authors, whose legitimate stylistic patterns can be mistaken for the statistical regularities that detectors interpret as machine-generated (Giray et al., 2025).

While AI detection-related false positives may affect a range of writer populations, this study focuses specifically on second language (L2) academic writers. Prior research suggests that L2 academic writing, even at advanced proficiency levels and across diverse L1 backgrounds, often exhibits discourse and stylistic patterns that distinguish it from prototypical L1 academic prose. These include greater reliance on formulaic sequences (Paquot & Granger, 2012), more explicit cohesive devices (Hinkel, 2001), more conservative lexical choices (Laufer & Waldman, 2011), more repetitive sentence structures (Hinkel, 2003; Lu, 2011), and closer adherence to genre conventions (Hyland, 2016). At the same time, we do not claim that these features are unique to L2 writers; highly conventionalized L1 academic prose may display similar configurations (Shin, 2019). Our argument, rather, is that the probabilistic clustering of such features may be especially common in some forms of L2 academic writing, thereby increasing the likelihood that multilingual authors are misread by current detection heuristics despite producing legitimate human-authored texts.

Critically, these L2 features closely resemble statistical patterns that many AI detection approaches are designed to capture (Opara, 2025). Most detectors

operationalize human writing as relatively variable and AI writing as relatively uniform, using measures tied to predictability, sentence- and paragraph-level variability, and repetition (Gehrmann et al., 2019; Giray, 2024; Kar et al., 2025; Liu et al., 2024; Mitchell et al., 2023; Picazo-Sanchez & Ortiz-Martin, 2024; Showemimo, 2025). Pattern predictability is typically operationalized using two models of detection. The first identifies localized patterns, flagging text when a threshold of co-occurring heuristics is surpassed within a document. The second analyzes document-wide heuristics to generate aggregate percentage scores representing the estimated proportion of AI-generated content.

Nevertheless, these statistical cues are not unique to AI-generated text (Doughman et al., 2025). Academic L2 writing's formulaic expressions and repetitive structures increase local predictability and lower perplexity. Additionally, conservative lexical choices reduce lexical dispersion, while adherence to genre conventions produces uniform syntactic patterns, all of which are treated as hallmarks of AI-generated prose (Muñoz-Ortiz et al., 2024). We argue that these cues are more likely to co-occur in legitimate L2 academic writing, especially in highly conventional sections, which makes multilingual writers structurally more vulnerable to false positives, with predictable downstream effects for educational assessment and publication gatekeeping (Ahmed & Ahmed, 2026; Giray et al., 2025; Kar et al., 2024; Liang et al., 2023).

Empirical results underline the practical consequence of this miscalibration. In some evaluations, more than 60% of NNS essays were incorrectly labeled as AI-generated by widely used detectors, indicating that false positives are not a marginal edge case but a systematic risk for multilingual writers (Giray, 2024; Kirchenbauer et al., 2023; Liang et al., 2023). Further evidence suggests that this risk intensifies when multiple L2 features co-occur, particularly in specialized academic genres where disciplinary conventions further constrain linguistic variation (Flitcroft et al., 2024; Giray et al., 2025; Popkov & Barrett, 2025).

This study is motivated by our experience with the heightened risk of AI detection as L2 academic writers. Despite high proficiency and a strong publication record, multiple manuscripts by the first author were flagged during editorial screening, constituting a recurring pattern and an empirically consequential point of departure for the study. To investigate this pattern in

a controlled and interpretable manner, we frame the present analysis explicitly as a single-author case study. The corpus was intentionally limited to one proficient L2 academic writer for two primary reasons. First, the author used AI tools in a controlled manner while writing the sample manuscripts, opting for minimal grammatical checking. Second, this approach aimed to reduce inter--author stylistic variation, allowing for a thorough examination of the textual features that could have triggered AI flagging in this specific case. Accordingly, the findings are exploratory and analytical rather than broadly generalizable, and they are not intended to support sweeping claims about L2 writers as a whole. Building on evidence that detectors often rely on predictability-oriented statistics (Miralles-González et al., 2025) and recognizing that commercial detectors are proprietary and opaque (Tufts et al., 2025), often providing only aggregate AI percentages, we engineered a transparent AI-detection simulator that operationalizes these heuristics under each hypothesis. This approach is designed to test plausible triggering mechanisms rather than to claim empirical equivalence with commercial systems. We do not benchmark proprietary commercial tools because their opacity precludes inference about why texts are flagged and limits comparison to surface-level similarity in outcomes rather than explaining underlying causes. Since our goal is to investigate mechanisms of false positive risk rather than reproduce black-box outputs, the simulator provides a more interpretable and reproducible basis for analysis. Specifically, we address the following research questions:

1) To what extent do these proxies, individually and in combination, yield elevated suspicion scores within our L2 corpus?
2) How are these simulated suspicion scores distributed across research article sections (e.g., Introduction, Methods, Results, Discussion)?
3) Which proxy configurations characterize the most highly suspicious segments, and how do these configurations correspond to known features of conventionalized academic prose?

Literature review

L2 academic writing

Empirical evidence demonstrates that L2 academic writing, particularly in English, exhibits characteristic linguistic patterns that distinguish it from L1 academic prose (Subandowo et al., 2025), even at advanced proficiency levels (Zaini & Ollerhead, 2019). These features arise not from deficient or developing L2 competence, but from legitimate L2 writing strategies, cross-linguistic influence (Jarvis & Pavlenko, 2008), and cognitive resource management during L2 composition (Zaini & Ollerhead, 2019). L2 writers strategically draw on these patterns to meet academic genre conventions, achieve clarity and precision in a non-native language, and produce manuscripts that satisfy the linguistic expectations of global scholarly publication and evaluation systems (Flowerdew, 2001; Hyland, 2016). For non-Western, non-Anglophone scholars, mastery of such conventionalized patterns is essential for participation in English-dominant academic discourse communities (Canagarajah, 2002; Lillis & Curry, 2010). Recognizing these features is crucial for understanding how L2 texts may interact with automated detection systems sensitive to lexical repetition and reduced diversity (Mitchell et al., 2023; Gehrmann et al., 2019).

For example, L2 academic writing is marked by heavy use of formulaic sequences and lexical bundles, that is, recurrent multi-word units in academic discourse (Paquot & Granger, 2012). Compared to L1 prose, L2 texts rely on more repetitive scaffolds and a narrower range of lexical combinations (Pan, 2018; Taşçı & Öztürk, 2021), which support cohesion, signal rhetorical moves, and maintain disciplinary register (Kashiha & Chan, 2015). This reliance is a strategic adaptation, as L2 writers draw on high-frequency academic phrases to reduce cognitive load and focus on content (Ädel & Erman, 2012). In technical sections, L2 academic writers further restrict lexis to preserve terminological precision, a tendency especially visible in STEM disciplines (Haq et al., 2023).

More broadly, L2 academic prose often shows simplified lexis and conservative word choice (Laufer & Waldman, 2011), with reduced lexical diversity, preference for high-frequency academic vocabulary, and limited use of low-frequency near

synonyms. Reusing familiar items promotes clarity and minimizes risks of register violations (Jarvis & McNamara, 2012), while repeated exposure to high-utility formulaic sequences and stable collocational patterns, rather than native-speaker stylistic norms, drives fluency in academic language (Casal & Yoon, 2023; Du & Hashimoto, 2023). At the syntactic level, L2 writers similarly lean on recurrent sentence templates and familiar clause-combining routines (Hinkel, 2003; Lu, 2011), recycling entrenched patterns to meet processing demands and maintain accuracy in high-stakes genres (Al Fattah, 2018; Canli & Yağiz, 2024).

Relatedly, information structure and thematic organization in L2 academic prose are strongly shaped by cross-linguistic influence. Because information packaging is language-specific, non-native English-speaking academics often transfer L1 discourse preferences into their L2 writing (Halliday & Matthiessen, 2014; Lambrecht, 1994), producing L1-patterned topicalization, theme placement, and clause-initial adverbials that affect how information is introduced and sequenced (Hinkel, 2002; Shehata & Eldakar, 2018). These patterns align with known mechanisms of second language acquisition and cross-linguistic influence (Murdoch et al., 2021). For example, L1-based thematic progression routines and preferences for clause-initial topics and discourse-organizing adverbials persist in L2 English (Pedersen, 2011; Mu & Zhang, 2018). When such L1-driven routines interact with academic genre norms that already favour predictable move structures and cohesive patterns, the resulting prose can appear highly patterned and formulaic, especially for writers with still-developing grammatical resources (Connor, 1996; Mu & Zhang, 2018). Cohesion is further shaped by a preference for overt discourse signaling, as L2 writers use more explicit cohesive devices like conjunctive adverbials, demonstratives, and lexical repetition, compared with L1 writers (Hinkel, 2001). This pattern reflects pedagogical emphasis and strategic compensation for perceived limits in implicit cohesion.

Patterns of stance and authorial presence also distinguish L2 texts. Non-native authors often underuse or redistribute hedges and self-references, reflecting cross-cultural rhetorical norms rather than simple linguistic deficiency (Ajideh et al., 2025; Ganjavi et al., 2024). These differences are especially salient in argumentative and evaluative sections, producing profiles of epistemic caution, engagement, and author visibility that may diverge from dominant

Anglophone conventions while remaining functionally effective (Mu & Zhang, 2018; Shchemeleva, 2021). Collectively, these information-structuring, cohesive, and stance-taking practices reinforce highly regular and explicitly signposted patterns of discourse organization in L2 academic writing.

The non-native academic writing genre itself exhibits disciplinary and national heterogeneity, with error profiles, bundle usage, and rhetorical patterns clustering by L1 and cultural proximity (Albelihi & Al-Ahdal, 2024). For example, East Asian L2 writers demonstrate patterns distinct from Romance-language or Arabic L2 writers in terms of syntactic complexity, lexical diversity, and rhetorical organization (Mavrou & Chao, 2023; Altakhaineh et al., 2024; Eid & Mutahar, 2025). These L1-specific patterns reflect both typological differences between languages and culturally mediated rhetorical traditions.

Disciplinary context also shapes L2 writing patterns. Different sections of academic research articles follow well-established stylistic conventions, with Methods and Results sections being typically formulaic and terminologically consistent (Ellis et al., 2008). This patterned variation means that highly standardized sections naturally exhibit controlled phrasing and reduced linguistic variance, making them comparatively easier for L2 authors to write, especially in STEM disciplines where shared technical conventions reduce language variance (Haq et al., 2023). In contrast, the Introduction and Discussion sections remain the most challenging writing tasks for L2 writers, as these sections demand more rhetorically complex, persuasive, and culturally mediated writing (Valipour et al., 2017; Bajwa et al., 2020; Junina et al., 2025). Argumentative sections require nuanced stance-taking and authorial voice that are especially challenging for L2 writers (Mu & Zhang, 2018; Shchemeleva, 2021).

L2 academic writing is therefore far from monolithic. Error profiles, bundle usage, and rhetorical routines vary systematically by L1 background, disciplinary field, and article section. Yet across this heterogeneity, tendencies toward regularity, repetition, and constrained variation at multiple levels of language persist. These features, we propose, closely resemble the surface patterns that contemporary AI detection tools target through proxies such as lexical diversity, burstiness, and formulaicity, providing the rationale for the next subsection's focus on AI detection methods and mechanisms.

AI detection technologies: Methods and mechanisms

AI-generated text detection has evolved rapidly since the widespread deployment of large language models, developing along partially overlapping chronological and thematic lines (Kehkashan et al., 2025). At their core, these detection tools operate by identifying predictable linguistic patterns, whether at the word level, sentence level, inter-sentence level, or across broader stylistic and genre conventions. We hypothesize that this fundamental mechanism creates a critical problem. Post-hoc detectors cannot reliably distinguish between (1) AI-generated text, which exhibits statistically probable patterns by design, and (2) L2 academic writing, which employs strategically formulaic patterns as a compensatory literacy strategy.

AI detection tools employ two primary methodological approaches, both of which rely on pattern predictability. Clustering-based models identify localized patterns characteristic of AI generation, flagging text when a threshold number of such features co-occur within a document. These clusters typically include high lexical diversity within constrained semantic fields, consistent sentence-level complexity, uniform transition patterns between ideas, and standardized academic register markers. Tools like GPTZero, Originality.AI, and Writer.com exemplify this approach. Prevalence-based models, by contrast, analyze document-wide statistical features to generate aggregate percentage scores representing the estimated proportion of AI-generated content. These models calculate frequencies of occurrence across target texts, measuring rates of formulaic expressions, transitional phrases, hedging devices, and syntactic templates, then produce normalized statistical values. Tools like Turnitin, Copyleaks, and Winston AI employ prevalence-based or hybrid detection mechanisms that combine both clustering and frequency analysis. Nevertheless, commercial vendors of AI detection tools uniformly fail to disclose their detection thresholds, i.e., the specific point at which their algorithms categorize text as *probably* machine-generated versus human-authored. This opacity applies across widely-used tools and prevents independent validation, also obscuring the empirical basis for classification decisions (Kehkashan et al., 2025). Understanding the technical foundations of these methods is essential

for evaluating their performance characteristics, limitations, and potential sources of systematic bias.

Many early detectors analyze token-level probabilities. A language model assigns a probability to each word given its context, and these signals are aggregated to distinguish human from machine-generated text (Gehrmann et al., 2019). Three related heuristics are central: perplexity, burstiness, and token-level likelihood. Perplexity captures how predictable a sequence is overall; lower perplexity indicates more predictable text. Because large language models favour high-probability continuations, AI-generated text tends to have lower perplexity than human writing (Flitcroft et al., 2024). Burstiness indexes variation in predictability across sentences. Human texts often alternate between more and less predictable passages, while AI outputs tend to be more uniform and thus less bursty (Popkov & Barrett, 2025). Token--probability heuristics apply thresholds to per-token likelihoods and their dispersion, flagging texts that consistently select high-probability tokens with low variance. These metrics are frequently combined into a composite detector score (Flitcroft et al., 2024; Popkov & Barrett, 2025). For L2 academic writing, these proxies align closely with the features described in the literature review. Legitimate L2 strategies can therefore depress perplexity, reduce burstiness, and concentrate probability mass on high-frequency tokens in ways that superficially resemble AI generation. Although statistical detectors can perform well under constrained conditions (same-domain text, sufficient length, minimally edited outputs), they are sensitive to domain shift, short inputs, and post-hoc editing (Gehrmann et al., 2019; Flitcroft et al., 2024), creating scope for systematic misclassification of patterned but human-authored L2 prose.

A second family of tools uses discriminative classifiers rather than explicit probability thresholds. In supervised settings, models are trained on labelled human and AI-generated texts and learn to separate them using features at multiple linguistic levels (Guo et al., 2023). In inference, an input is encoded and passed through the classifier, which outputs a probability or label indicating whether it more closely resembles human or AI examples; when training and test distributions align, such systems can achieve strong in-domain accuracy (Uchendu et al., 2020).

Zero-shot and representation-based detectors extend this idea by leveraging pretrained language models (e.g., BERT, RoBERTa) without task-specific fine--tuning. Instead of learning from large, labelled AI versus human corpora, they use embeddings or attention patterns as general-purpose representations of linguistic structure and infer whether a text is more "machine-like" or "human--like" (Ippolito et al., 2020). These approaches reduce the need for dedicated training data and, in principle, can detect outputs from novel generative systems without retraining. However, both supervised and zero-shot detectors inherit constraints from their data and architectures. Performance degrades under domain shift, when texts differ in genre, topic, register, or language variety (Aldeen et al., 2025), and models can often be evaded through paraphrasing or synonym substitution that preserves meaning while altering surface features (Guo et al., 2025). Zero-shot systems, in particular, inherit the biases and discourse norms of their pretrained models, which are largely based on English Wikipedia, academic publications, and web text reflecting Western academic and journalistic conventions (Bender et al., 2021). This creates a specific risk for L2 academic writing. The lexical, syntactic, and discourse-level regularities can become precisely the stylometric and representational features that classifiers associate with "machine-like" text.

Training data and benchmark limitations

The effectiveness of AI detection systems is limited by the datasets used for their development. These datasets largely focus on English and traditional academic styles, reflecting the demographics of major AI research communities (Wang et al., 2024). Consequently, detectors are trained on data dominated by Western, English L1 writers (Gosselin, 2025) and fail to adequately represent World Englishes, L2 academic writing, and non-Western rhetorical styles (Caines et al., 2023; Won et al., 2025). These systems often equate "human" text with Anglophone native-speaker norms, overlooking legitimate L2 strategies as "machine-like" (Liang et al., 2023; Weber-Wulff et al., 2023). Features common in L2 academic writing may be misinterpreted as negative signals in AI training, raising concerns about detector validity for diverse global academic texts.

Current systems focus more on surface-level statistical patterns than on deep semantic markers of authorship. Evidence shows that even minor paraphrasing can significantly lower detector scores, suggesting that classification depends on distributional cues rather than a clear notion of "authorship authenticity" (Liu et al., 2024).

Methodology

Study design

This study uses an experimental model-building approach to reverse-engineer AI detection mechanisms and test how they behave on a small sample of authentic L2 academic writing. The design responds to a central limitation in current practice. Commercial vendors and editorial offices provide no transparency about how their tools classify text as "AI-generated" or which features trigger flags. Without access to proprietary algorithms or detailed feedback on individual decisions, we cannot directly observe why specific passages in our corpus were misclassified. We therefore adopt a theory-driven simulation of the two dominant AI detection paradigms, i.e., clustering models and prevalence models, and ask whether the L2 writing patterns in our data systematically align with the heuristics these paradigms operationalize. Rather than attempting to replicate any single commercial system, we approximate the core proxy features that detectors are known to target and examine how often legitimate L2 prose exhibits those signatures.

To do this, we analyzed a single-author corpus to control for inter-author stylistic variation while maintaining ecological validity. The corpus comprised four research articles (approximately 28,000 words) written by the first author between 2022 and 2024. All four were rejected by peer-reviewed journals on suspicion of AI generation and flagged by commercial detection tools, despite only minimal use of generative AI (restricted to grammatical checking). Two articles address applied statistics topics and two address applied linguistics topics, each following the standard IMRD structure. These articles are used

strictly as data and are not cited as references in this manuscript to preserve anonymity during review.

This design offers several methodological advantages. First, it is controlled for AI-aided writing. Moreover, single authorship helps isolate L2-typical features from idiosyncratic personal style. Using real rejection cases, rather than artificially constructed texts, grounds the analysis in actual editorial outcomes and provides ecological evidence for the detection paradox. The inclusion of both technical and more discursive articles allows us to explore whether detection patterns are section- and discipline-specific or generalize across domains. The corpus yields 334 analyzable segments, providing sufficient granularity for feature-level statistical analysis. Finally, because the same author and texts were previously flagged by commercial tools, we can assess whether our reverse-engineered detector reproduces similar flagging patterns, offering a construct-valid approximation of how current detection architectures respond to L2 academic writing.

The AI detection simulator

Rather than running black-box commercial tools directly, we implement an AI detection simulator grounded in the literature on detection mechanisms and L2 writing characteristics. This simulator is a proof-of-concept tool; our aim is diagnostic, i.e., to test whether L2 writing strategies produce the specific statistical signatures that contemporary detectors interpret as evidence of machine generation. Instead of first coding the corpus for generic L2 features (e.g., formulaic sequences, transition markers, hedging devices) and then passing it through opaque detectors that return only aggregate scores, we code the corpus directly for AI detection heuristics that correspond to those L2 patterns. This allows us to examine, at the level of individual segments, whether the same proxies associated with AI output, i.e., high predictability, low variability, formulaicity, constrained lexis, are also systematically present in authentic L2 academic prose. The five features used in the simulator are deliberately simplified proxies that attempt to capture broad tendencies rather than precise linguistic or cognitive constructs.

The simulator operationalizes two competing detection logics. Clustering, where segments are flagged when multiple heuristics co-occur, and prevalence, where segments are flagged when individual heuristics exceed a threshold. In both cases, we focus on five proxy features identified in the literature as characteristic of both AI-generated text and L2 academic writing:

1) Low perplexity: predictable word sequences, measured via n-gram probabilities.
2) Low burstiness: uniform sentence-length distribution, captured by the coefficient of variation.
3) High formulaic language: reliance on recurrent multi-word expressions, measured by n-gram frequency.
4) Consistent formality: minimal variation in register, operationalized as low variance in formality scores.
5) Low lexical diversity: restricted vocabulary range, measured via type–token ratio.

All five features are implemented using computational measures rather than manual coding (see Appendix A for technical details). This choice has three main motivations. First, computational metrics provide objective, replicable quantification of linguistic patterns, reducing subjective coder bias. Second, they enable efficient analysis of a corpus size that would be impractical to annotate manually. Third, and most importantly, they simulate the statistical signatures that AI detection algorithms themselves target: we use comparable language models (e.g., GPT-2) and metrics (e.g., perplexity, burstiness) to approximate how detectors "see" the text.

Window-based analysis framework

To analyze how AI-detection-prone features are distributed across the papers, we used a window-based analysis. Instead of treating each article or section as a single unit, window-based analysis divides the text into many smaller, overlapping pieces (windows) of fixed length. This makes it possible to

see where particular patterns cluster in the text, rather than assuming that linguistic features are uniform across an entire article. In this study, each article was split into overlapping 200-word windows. Each 200-word window was treated as an independent unit of analysis. For every window, we calculated the five AI-detection proxies described above and noted whether each proxy crossed its threshold. We then assigned each window a suspicious score from 0 to 5, defined as the number of proxies that were flagged in that window. A score of 0 means that none of the AI-like features were present; a score of 5 means that all five features occurred together in the same 200-word span.

A window was classified as suspicious when it showed two or more flags. The two-flag threshold was chosen because a single flagged feature can simply reflect an isolated stylistic choice, whereas the co-occurrence of multiple features is closer to the pattern that AI detectors are designed to identify. Importantly, the term *suspicious* does not imply wrongdoing or that the text is actually AI-generated. It simply indicates that the window displays the statistical profile that detection tools are likely to interpret as machine-like. We report results at three levels of suspicion:

- 2+ flags (moderate suspicion): at least two of the five features present; this is the baseline threshold at which an AI detector might begin to treat a passage as questionable.
- 3+ flags (high suspicion): at least three features present; this indicates a stronger convergence of AI-like patterns.
- 4+ flags (very high suspicion): four or all five features present; this represents an extreme concentration of AI-like signals that is relatively rare in typical human writing but can still occur in highly conventionalized L2 academic prose.

For each suspicious window, we also recorded which specific features were flagged so that we could analyze common combinations (for example, whether low perplexity tends to co-occur with high formulaicity). These co-occurrence patterns are reported in the Results section. To illustrate how the flagging system

works, consider Window 35, an excerpt from the discussion section of one of the linguistics papers ("Linguistics 2"):

> "generalizable finding in the data. Positive face (PF) strategies, which focus on building solidarity and positive social relations (like appreciation, gratitude, motivation, and starting conversations), also show a positive effect on acceptability (+0.37). However, the p-value of 0.081 means this effect is not statistically significant at the conventional 0.05 level…"

This passage reads like typical discussion-section prose: it reports statistical results in formal academic language and then interprets them. In our simulator, this window received a suspicious score of 4 out of 5. It was flagged as:

- Low perplexity (F1): the wording is highly predictable because it uses standard phrases such as "positive effect on acceptability" and "statistically significant at the conventional 0.05 level".
- Low burstiness (F2): sentence structure and length are relatively uniform, producing a smooth, regular rhythm.
- High formulaic language (F3): the passage relies on common academic word combinations and technical expressions.
- Low lexical diversity (F5): the vocabulary is somewhat restricted, with repeated use of key terms related to statistical reporting.

By contrast, formality variance (F4) was not flagged in this window because the text moves slightly between technical reporting and more explanatory language, creating some register variation. This is genre-appropriate for interpretive social science writing and shows that even windows with high suspicious scores can still contain subtle stylistic shifts. (Full numeric details for this and additional examples are provided in Appendix B). This window-based setup directly mirrors the two detection logics modelled in our simulator. At the prevalence level, we ask how often individual proxies cross their thresholds across all windows, whereas at the clustering level, we treat high-suspicion

windows (2+, 3+, or 4+ flags) as segments where multiple proxies converge in the way commercial detectors are designed to target.

Ethical considerations

To align with academic standards and to protect the integrity of ongoing editorial processes, the full titles of the four manuscripts are not disclosed. Two manuscripts are already published, and two remain under consideration; withholding titles reduces re-identification risk, preserves the anonymity expected in peer review contexts, and prevents conflation of evaluative judgments with the present methodological investigation. This confidentiality does not affect reproducibility because all analyses operate on de-identified text segments, and the release package includes hashed document identifiers, section labels, and complete feature and outcome matrices.

Results

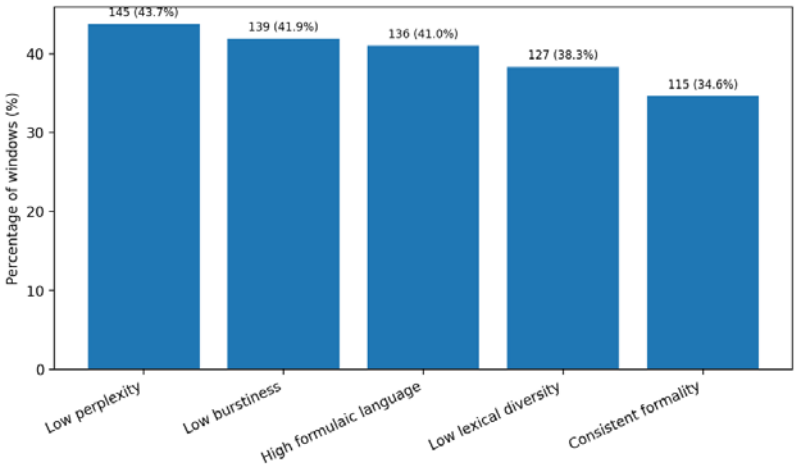
The four articles in the corpus contain 25,627 tokens and 3,350 distinct word types. Together, they comprise 404 paragraphs and 1,690 sentences, with an average sentence length of 15 words. None of the papers is an outlier in terms of length, sentence length, or paragraph count, indicating that they are structurally typical research articles rather than unusually short or long texts. After applying the 200-word sliding windows described in the methodology, the corpus yielded 332 analyzable windows: Technical Paper 1 (72 windows), Technical Paper 2 (107 windows), Linguistics Paper 1 (102 windows), and Linguistics Paper 2 (51 windows). Across this corpus, the AI detection simulator flagged 202 windows (60.8%) as suspicious at the baseline threshold (two or more flags), meaning they displayed at least some convergence of the five AI-like features. This broad pattern roughly mirrors the feedback from editorial offices and commercial tools.

Competing hypotheses of AI detection

Prevalence of individual features (prevalence model)

We first examined how often each of the five AI-detection proxies appeared across the 332 windows, treating each feature in isolation. The features occur at comparable rates, with no single proxy dominating the others. Low perplexity (predictable word sequences) appeared in 145 windows (43.7%), low burstiness (uniform sentence patterns) in 139 windows (41.9%), high formulaic language (recurrent multiword expressions) appeared in 136 windows (41.0%), low lexical diversity (restricted vocabulary range) appeared in 127 windows (38.3%), and consistent formality (little register variation) appeared in 115 windows (34.6%). Figure 1 illustrates the distribution of the identified proxies in the corpus.

Figure 1. Distribution of AI proxies in the sample



Source: Author's own elaboration.

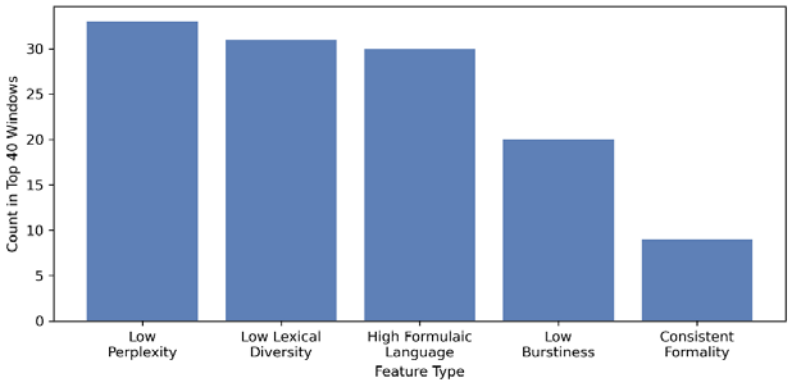
These results suggest that the corpus does not strongly resemble AI output when evaluated at the level of individual windows. As a prevalence model, this suggests that if detectors flagged the texts as AI-generated, they did so even though high concentrations of suspicious features are rare. Instead, the more plausible mechanism is that detectors treat any single feature crossing

a threshold as sufficient to raise suspicion, even if no other AI-like properties are present. If prevalence-based detectors treat any one of these features as a warning signal, then texts become vulnerable to flagging even when they exhibit only mild degrees of AI-like patterning.

Composition of high-scoring windows (clustering model)

We next turned to the clustering model, which focused on windows where multiple features co-occurred. Here, we examined the 40 most suspicious windows, defined as those with the highest cumulative scores, to see which combinations of features drive suspicion, and a combination of three proxies stands out. Low perplexity is present in 33 of 40 windows (82.5%), low lexical diversity appears in 31 windows (77.5%), and high formulaic language appears in 30 windows (75.0%). In contrast, low burstiness appears in 20 windows (50.0%), and consistent formality appears in 16 windows (40.0%). Figure 2 illustrates this pattern.

Figure 2. Clustering of AI proxies in the 40 most suspicious windows



Source: Author's own elaboration.

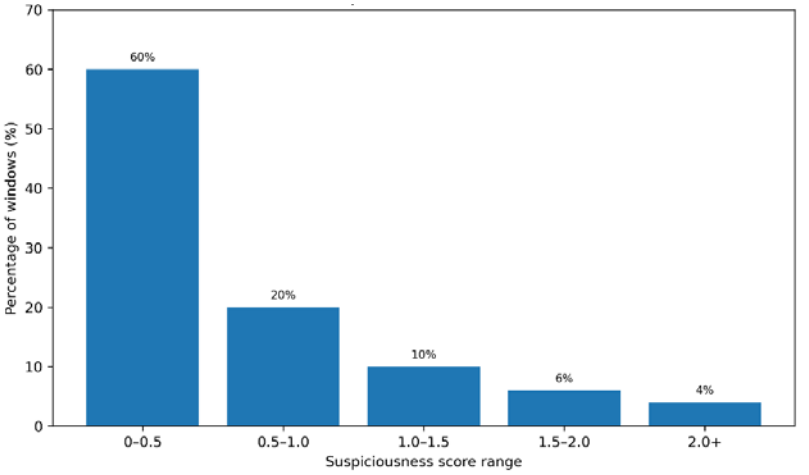
This pattern indicates that, when multiple features cluster together, three signals are doing most of the work. Predictable phrasing, heavy reuse of formulaic expressions, and restricted vocabulary are precisely the properties that prior research identifies both as core L2 academic strategies (e.g.,

formulaic bundles, conservative lexis, stable reporting frames) and as proxies used by AI detectors (low perplexity, high formulaicity, low lexical diversity). Crucially, the high-scoring windows do not occur at random in this data set. They are concentrated in sections that report statistical results, passages that describe standardized methodological procedures, and spans that present quantitative findings or parameter lists. In other words, the windows that look most AI-like to our simulator are also the most conventionalized parts of academic writing, where disciplines expect authors to use highly standardized wording. The simulator, therefore, flags segments that perform precisely the communicative functions that the IMRD structure requires, rather than passages that are stylistically unusual or semantically thin.

In our analysis, each article was divided into overlapping 200-word windows, and each window received a suspiciousness score from 0 to 5, reflecting how many of the five AI-detection proxies crossed their thresholds in that span. Scores of 0 or 1 are common and largely unremarkable, since a single flagged feature can arise from a local stylistic choice. Higher scores indicate the co--occurrence of multiple AI-like patterns in the same window, which is closer to the profile targeted by detection tools. Windows with two or more flags were treated as suspicious, with 2+ flags marking moderate suspicion, 3+ flags high suspicion, and 4+ flags very high suspicion.

However, the resulting score distribution is strongly right-skewed rather than bimodal. Figure 3 illustrates that most windows fall at the low end of the scale, and only a small subset accumulates three or more flags. This pattern indicates that conventional, highly formulaic passages form a long tail of high--scoring windows rather than a separate "AI" cluster, complicating any attempt to draw a hard categorical boundary between human and machine-like text.

Figure 3. Distribution of suspicious scores across windows



Source: Author's own elaboration.

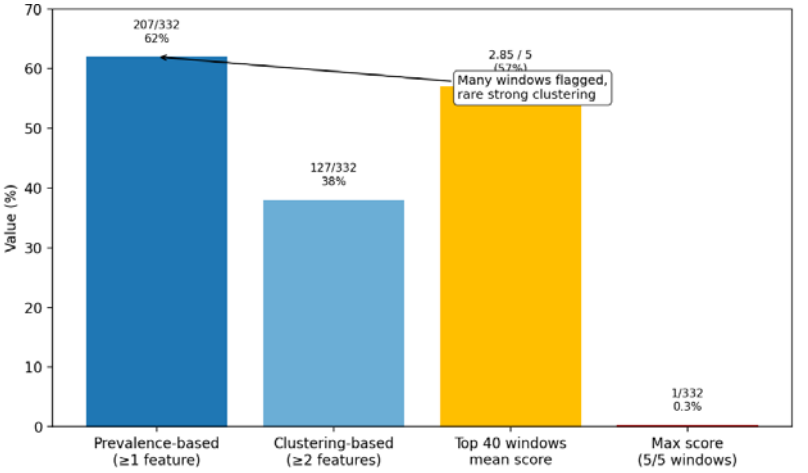
The relatively modest presence of consistent formality among high-scoring windows also suggests that register stability alone does not drive suspicion. Rather, it is the combination of predictable language, formulaic structure, and clustered terminology that produces the statistical profile most likely to be targeted by detection tools. Importantly, these combinations arise because the author follows disciplinary reporting norms, not because the text was generated by a language model.

The detection paradox

Under a prevalence-based model, which flags any window where at least one proxy exceeds its threshold, 62% of windows (207/332) would be treated as at risk. When we require at least two features in the same 200-word span, the clustering-based model still flags 38% of windows (127/332). Even within the 40 highest-scoring windows, the mean suspicious score is only 2.85 out of 5, and just one window in the entire corpus reaches the maximum of five co--occurring features. As visualized in Figure 4, this creates a detection paradox. Dense clustering of AI-like features is rare and mostly confined to genre--required segments, yet both architectures, especially the prevalence-based

model, would mark a substantial proportion of the corpus as suspicious. These findings sit against a striking practical backdrop, i.e., the texts in this corpus had already been rejected by multiple journals because they were "heavily AI--supported". Yet our analyses show that dense clustering of AI-like features is relatively rare, and when it does occur, it is typically confined to genre-required segments such as procedural descriptions, boilerplate method language, or parameter lists.

Figure 4. The detection paradox-flagging rates and score extremes



Source: Author's own elaboration.

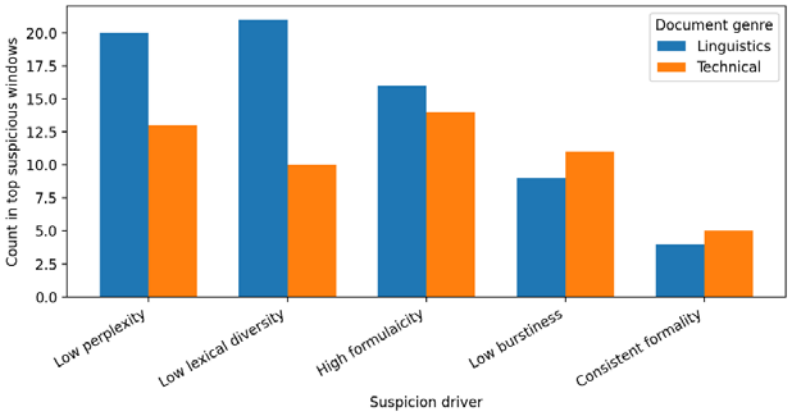
Commercial detectors and our simulator would flag a large share of these manuscripts as suspicious, particularly under prevalence-based logic, not because they contain long stretches of overtly machine-like prose, but because they are suffused with dispersed, individually common features that cross low thresholds. For multilingual scholars writing in English, this creates a structural vulnerability. L2 academic writers are precisely the group most likely to lean on conventional phrasing, fixed formulae, and other high-frequency academic routines to manage linguistic risk and conform to Anglophone norms. In doing so, they reproduce the very statistical signatures that current detectors interpret as AI authorship.

The analysis shows that the five features used in the AI detection simulator do not behave randomly in the corpus but follow a clear hierarchy in how often they appear and how strongly they contribute to suspicious scores. formulaic reuse, low perplexity, and low burstiness emerge as the main drivers of flagging. They appear in roughly 40–43% of all windows and concentrate in parts of the texts where academic conventions most tightly constrain wording, i.e., procedural descriptions, parameter listings, definitional and citation scaffolds, and standardized result statements. Low lexical diversity plays a secondary role, while consistent formality contributes more modestly.

Results by document genre

When we break the data down by document type, technical and linguistic papers show the same ordering of main suspicion drivers. Formulaic reuse, low perplexity, and low burstiness, followed by low lexical diversity and, finally, consistent formality. However, the relative emphasis of these drivers differs by genre. In the linguistics papers, suspicious windows show a slightly stronger contribution from formulaic reuse and low burstiness. This is consistent with sections that describe samples, annotation schemes, or coding procedures using templated phrases, and provide definitional and expository passages that recycle standard explanatory frames. In the technical papers, the same three drivers are evident, but consistent formality contributes less because the register is highly constrained from the outset. Suspicious windows here are dominated by standardized reporting frames (e.g., parameter estimations, model specifications), and parameter-dense prose where domain-specific terms repeat within narrow spans. Both patterns are hallmarks of legitimate discipline-specific writing, not symptoms of automated generation. Figure 5 visualizes the relative contribution of each proxy across the top 40 most suspicious windows by genre.

Figure 5. The most suspicious windows by genre



Source: Author's own elaboration.

Results by section

Finally, we examined where suspicious windows cluster within the IMRD structure. Although academic writing is conventional across all sections, suspicious flagging is not evenly distributed. In this corpus, Methods sections, and to a lesser extent Results, most frequently cross the suspicious thresholds. These sections are where low perplexity and low burstiness co-occur with high formulaic reuse, especially in procedural descriptions, stepwise enumerations, and parameter or variable listings. By contrast, introductions show moderate formulaic reuse and local dips in burstiness around definitional scaffolds and citation-heavy background passages, but these features rarely combine with enough intensity to trigger high suspicion scores.

Additionally, results in technical papers display elevated formulaicity and predictability in line with standardized statistical reporting, while linguistics results are more varied but still rely on recurring reporting frames. Discussions contain relatively few suspicious windows; when they do appear, they are more strongly associated with low perplexity and low burstiness than with heavy bundle reuse, reflecting tight but less templated interpretive prose. Abstracts and other short segments sometimes show low perplexity and low burstiness

owing to dense information packaging, but typically lack the sustained formulaic coverage seen in Methods. These sectional patterns align with the genre-typical functions of each part of the article. Procedural, definitional, and standardized reporting segments are predictably conventionalized and therefore more likely to resemble AI proxies, while interpretive segments are more varied and thus less likely to cross detector thresholds.

Discussion

Alignment with L2 academic writing and AI detection research

The distribution of features in this corpus is highly consistent with existing work on academic English, L2 writing, and World Englishes (Alfehaid & Alkhatib, 2023; Du & Hashimoto, 2023; Ellis et al., 2008; Li, 2024; Öztürk & Taşçı, 2023) Corpus and genre studies have long documented the heavy use of lexical bundles and formulaic sequences in Methods, Results, and definitional parts of Introductions, as well as in citation-dense stretches of text (Bajwa et al., 2020; Ellis et al., 2008; Öztürk & Taşçı, 2023). Genre analyses also predict highly regular staging in Introductions and standardized procedures in Methods and Results (Bajwa et al., 2020; Haq et al., 2023), which corresponds to lower burstiness and lower perplexity in those parts of the texts (Odri & Yoon, 2023; Picazo-Sanchez & Ortiz-Martin, 2024).

The findings also reinforce research on L2 academic writing that describes formulaic sequences as legitimate scaffolds for cohesion, clarity, and fluency, especially for multilingual writers (Öztürk & Taşçı, 2023). L2 authors are known to rely heavily on recurrent expressions, reporting frames, and definitional templates as they manage the dual demands of disciplinary content and an additional language. This reliance is neither an error nor a shortcut; it reflects successful socialization into the phraseology and routines of academic English (Showemimo, 2025; Giray, 2024). The concentration of suspicious scores in these routines supports the argument advanced in the literature review that L2 compensatory strategies and AI-detection proxies substantially overlap (Liang et al., 2023).

At the same time, the results align with empirical studies of AI detection tools, which show that current systems focus on the distributional properties of text (such as perplexity and burstiness) rather than on provenance, intent, or content knowledge (Aldeen et al., 2025). Prior evaluations have also noted that detectors' performance can vary by genre and section, with highly structured academic prose. For example, Popkov & Barrett (2025) found that ZeroGPT mis-labelled ≈62% of authentic behavioral-health articles as AI-generated, with the greatest errors in method-heavy sections. Similarly, Turnitin's AI-detector reported ≈4% false-positive rates for human-written sentences, but whole-document false positives rose sharply for formally structured essays (Showemimo, 2025). In L2 corpora, the problem intensifies. Studies of TOEFL essays (non-native English) show false-positive rates above 60% across seven popular GPT-detectors, because the limited lexical variability of learners' writing mimics the low-entropy signatures of machine-generated text (Liang et al., 2023).

Theoretical and practical implications

The findings highlight the detection paradox. Our simulator reveals that while individual AI-like features are common, strong clustering of multiple features is rare. Nonetheless, many windows can still be flagged under prevalence--based and clustering-based criteria due to low proxy thresholds used by detectors. This means that "more predictable" is mistakenly equated with "more likely AI-generated", contradicting insights from genre studies and corpus linguistics about academic prose. This dynamic creates a clear double bind for L2 writers. The sections where they are under the greatest pressure to follow disciplinary templates (Methods and Results) are also where our simulator finds the highest suspicious scores. Adhering closely to accepted reporting conventions increases the risk of being misclassified as using AI, but deviating from those conventions can undermine perceived credibility, clarity, and publishability. In this sense, the features driving high suspicious scores (low perplexity, low lexical diversity, and high formulaicity) are better understood as markers of successful participation in academic discourse communities than as signs of machine authorship.

This has several practical implications. For teaching and supervision, the findings underline the need for careful messaging around AI and academic writing. English for Academic Purposes (EAP) courses and supervisors should continue to promote appropriate conventional expressions and reporting frames, especially for multilingual writers, while also raising awareness of how detectors work and where they are likely to misfire. Non-native writers should not be discouraged from using formulaic language that signals disciplinary competence simply because it may resemble AI proxies.

For editorial policy and institutional practice, the results support treating AI detector outputs as screening tools rather than standalone evidence of misconduct. Editorial offices, universities, and funding bodies should be cautious about using raw detector scores in high-stakes decisions, particularly for L2 authors or submissions from diverse linguistic backgrounds. Where detectors raise concerns, follow-up should involve contextual judgement that considers genre, section, and writer background rather than defaulting to automatic sanction.

For AI detection research and development, the study adds empirical support to calls for more genre- and language-sensitive models. If false positives in this L2 corpus largely arise from normal academic conventions, then future detectors need to distinguish more carefully between predictable, formulaic text that is expected for a given genre and genuinely suspicious patterns. This could involve developing section-specific baselines that recognize the inherently higher conventionality of Methods and Results compared to Discussions, and incorporating information about text type, disciplinary genre, and possibly writer background so that expectations about predictability and formulaicity are calibrated rather than uniform, creating more equitable detection.

Conclusion

This proof-of-concept study combines a small, motivated sample of L2-authored research manuscripts with a transparent, detector-inspired simulator to explore one possible mechanism behind AI detection false positives. Within

this constrained setting, the most suspicious simulated segments tended to cluster in highly conventionalized methods and results prose characterized by low perplexity, low burstiness, and high formulaic reuse; precisely the kinds of patterns that are also encouraged by disciplinary writing norms and by legitimate L2 strategies for managing linguistic and cognitive load.

The study makes three main contributions. Theoretically, it refines the detection paradox by demonstrating how detectors can conflate linguistic predictability with machine authorship. Methodologically, it introduces a simulation-based, window-level approach that maps widely discussed detection heuristics onto authentic L2 research articles without relying on opaque commercial tools. Practically, the findings underscore the need for pedagogical guidance, editorial policies, and detector designs that recognize highly conventional academic language, especially in L2 writing, as a normal feature of scholarly communication rather than an automatic marker of AI misuse. The results suggest that protecting the integrity of scholarship will require not only better-calibrated detection architectures, but also institutional practices that allow multilingual authors to contextualize detector outputs and contest automated flags.

This work has limitations. The single-author design enables controlled analysis of detection mechanisms but constrains generalizability across L2 populations. The simulation approach, while transparent, does not directly test commercial detectors or compare false positive rates across tools. Future research should scale this framework to multi-author, multi-institution corpora spanning diverse L1 backgrounds, conduct prospective evaluations of section--weighted detection algorithms, and assess the impact of editorial interventions on false positive rates and perceived fairness in peer review.

References

Abd-Elaal, E.-S., Gamage, S. J. E., & Mills, J. E. (2022). Assisting academics to identify computer-generated writing. *European Journal of Engineering Education*, 47(5), 725–745. DOI: 10.1080/03043797.2022.2046709.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. DOI: 10.1016/j.esp.2011.08.004.

Ajideh, P., Zohrabi, M., & Ohbatalab, R. (2025). Stance markers in academic writing: Native vs. non-native (Iranian) authorship in hard and soft sciences research articles. *Critical Literary Studies*, *7*(2). DOI: 10.22034/cls.2025.63772.

Ahmed, J., & Ahmed, N. (2026). When Human Writing is Marked AI written: Investigating Turnitin's Misclassification of Students' Writing. *Research Journal for Social Affairs*, 4(1), 209–215. DOI: 10.71317/RJSA.004.01.0702.

Aldeen, S. D., Abbas, T., & Abbas, A. R. (2025). Review of Detecting Text generated by ChatGPT Using Machine and Deep-Learning Models: A Tools and Methods Analysis. *Diyala Journal of Engineering Sciences*, *18*(1), 34–54. DOI: 10.24237/djes.2025.18102.

Alfehaid, A., & Alkhatib, N. (2024). (Non)-Conformity to Native English Norms in Postgraduate Students' Writing in UK Universities: Perspectives of Native and Non-Native Students and Academic Staff. *International Journal of Arabic--English Studies (IJAES)*, *24*(1), 1–20. DOI: 10.33806/ijaes.v24i1.562.

Al Fattah, N. (2018). Cognitive load theory in the context of second language academic writing. *Higher Education Pedagogies*, *3*(1), 385–402. DOI: 10.1080/23752696.2018.1513812.

Albelihi, H. H. M., & Al-Ahdal, A. (2024). Overcoming error fossilization in academic writing: Strategies for Saudi EFL learners to move beyond the plateau. *Asian-Pacific Journal of Second and Foreign Language Education*, 9(75). DOI: 10.1186/s40862-024-00303-y.

Altakhaineh, A. R. M., Younes, A. S., & Allawama, A. (2024). A corpus--driven study of gratitude in English acknowledgements by Arabic-speaking

MA students: constructing L2 academic writer identity. *Cogent Arts & Humanities*, *11*(1). DOI: 10.1080/23311983.2024.2346361.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

Bajwa, N. H., König, C. J., & Kunze, T. (2020). Evidence-based understanding of introductions of research articles. *Scientometrics*, *124*, 195–217. DOI: 10.1007/s11192-020-03475-9.

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O. et al. (2023). On the application of large language models for language teaching and assessment technology. *arXiv:2307.08393*. DOI: 10.48550/arXiv.2307.08393.

Canli, Z., & Yağız, O. (2024). A contrastive investigation into the non-native speakers of English academicians' academic writing cognitions and challenges in the first and second languages. *Arab World English Journal*, *15*(1), 117–131. DOI: 10.24093/awej/vol15no1.8.

Casal, J. E., & Yoon, J. (2023). Frame-based formulaic features in L2 writing pedagogy: Variants, functions, and student writer perceptions in academic writing. *English for Specific Purposes*, *71*, 102–114. DOI: 10.1016/j.esp.2023.03.004.

Connor, U. (1996). *Contrastive rhetoric*. Cambridge University Press.

Doughman, J., Afzal, O. M., Toyin, H. O., Shehata, S., Nakov, P., & Talat, Z. (2025, January). Exploring the limitations of detecting machine-generated text. In: *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 4274–4281). DOI: 10.48550/arXiv.2406.11073.

Du, Z., & Hashimoto, K. (2023). TCNAEC: Advancing sentence-level revision evaluation through diverse non-native academic English insights. *IEEE Access*, *11*, 144939–144952. DOI: 10.1109/ACCESS.2023.3342862.

Eid, F. M. S., & Mutahar, M. S. A. (2025). Bridging rhetorical differences: Arabic textual metaphors in academic writing and translation. *European Journal of Arts, Humanities and Social Sciences*, *2*(2), 183–201. DOI: 10.59324/ejahss.2025.2(2).21.

Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. DOI: 10.1002/j.1545-7249.2008.tb00137.x.

Flitcroft, M. A., Sheriff, S. N., Wolfrath, N., Maddula, R., McConnell, L., Xing, Y., & Kothari, A. N. (2024). Performance of artificial intelligence content detectors using human and artificial intelligence-generated scientific writing. *Annals of Surgical Oncology*, *31*(10), 6387–6393. DOI: 10.1245/s10434-024-15549-6.

Flowerdew, J. (2001). Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly*, *35*(1), 121–150. DOI: 10.2307/3587862.

Ganjavi, A., & Conner, M. B. A. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: A bibliometric analysis. *BMJ*, 386, e077192. DOI: 10.1136/bmj-2023-077192.

Canagarajah, A. S. (2002). *A geopolitics of academic writing* (Vol. 163). University of Pittsburgh Press.

Gao, C. A., & Howard, F. N. S. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, *6*(1), 75. DOI: 10.1038/s41746-023-00819-6.

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 111–116). Association for Computational Linguistics. DOI: 10.18653/v1/P19--3019.

Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *The Serials Librarian*, *85*(5–6), 181–189. DOI: 10.1080/0361526X.2024.2433256.

Giray, L., Santos, K., & Manalang, F. (2025). Beyond policing: AI writing detection tools, trust, academic integrity, and their implications for college. *Internet Reference Services Quarterly*, *29*(1), 83–116. DOI: 10.1080/10875301.2024.2437174.

Gosselin, R. D. (2025). AI detectors are poor Western blot classifiers: A study of accuracy and predictive values. *PeerJ*, *13*, e18988. DOI: 10.7717/peerj.18988.

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology*, *50*(5), 2572–2593.DOI: 10.1111/bjet.12864.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv:*2301.0759. DOI: 10.48550/arXiv.2301.07597.

Guo, H., Cheng, S., Zhang, K., Shen, G., & Zhang, X. (2025). CodeMirage: A multi-lingual benchmark for detecting AI-generated and paraphrased source code from production-level LLMs. *arXiv:2506.11059*. DOI: 10.48550/arXiv.2506.11059.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge. DOI: 10.4324/9780203431269.

Haq, Z. U., Naeem, H., Naeem, A., Iqbal, F., & Zaeem, D. (2023). Comparing human and artificial intelligence in writing for health journals: An exploratory study. *medRxiv*. DOI: 10.1101/2023.02.22.23286322.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning*, *12*(2), 111–132.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Routledge.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, *37*(2), 275–301. DOI: 10.2307/3588505.

Hyland, K. (2016). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, *31*, 58–69. DOI: 10.1016/j.jslw.2016.01.005.

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1808–1822). Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.164.

Jarvis, S., & McNamara, D. S. (2012). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity, and conceptual knowledge. In: S. Jarvis & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 106–126). Channel View Publications. DOI: 10.21832/9781847696991-005.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Junina, A. K., Strauss, P. L., Wood, J. K., & Grant, L. (2025). A mixed-method inquiry into Arabic-speaking students' experiences with English academic writing at the undergraduate level. *Asia Pacific Journal of Education*, 45(1), 260–277. DOI: 10.1080/02188791.2022.2101986.

Kar, S. K., Bansal, T., Modi, S., & Singh, A. (2025). How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI--detector Tools. *Indian Journal of Psychological Medicine*, 47(3), 275–278. DOI: 10.1177/02537176241247934.

Kashiha, H., & Chan, S. H. (2015). A little bit about: Differences in native and non-native speakers' use of formulaic language. *Australian Journal of Linguistics*, 35(4), 297–310. DOI: 10.1080/07268602.2015.1067132.

Kehkashan, T., Riaz, R. A., Al-Shamayleh, A. S., Akhunzada, A., Ali, N., Hamza, M., & Akbar, F. (2025). AI-generated text detection: A comprehensive review of methods, datasets, and applications. *Computer Science Review*, *58*, 100793. DOI: 10.1016/j.cosrev.2025.100793.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge University Press. DOI: 10.1017/CBO9780511620607.

Laufer, B., & Waldman, T. (2011). Verb–noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. DOI: 10.1111/j.1467-9922.2010.00621.x.

Liang, W., Yükselgonul, M., & Mao, Y. J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. DOI: 10.1016/j.patter.2023.100779.

Lillis, T. M., & Curry, M. J. (2010). *Academic writing in global context*. Routledge.

Liu, J. Q., & Hui, K. A. Y. (2024). The great detectives: Humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, *20*(1), 8. DOI: 10.1007/s40979-024-00155-6.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. DOI: 10.5054/tq.2011.240859.

Mavrou, I., & Chao, J. (2023). What does linguistic distance predict when it comes to L2 writing of adult immigrant learners of Spanish? *Written Communication*, *40*(3), 943–975. DOI: 10.1177/07410883231169511.

Miralles-González, P., Huertas-Tato, J., Martín, A., & Camacho, D. (2025). Not all tokens are created equal: Perplexity Attention Weighted Networks for AI generated text detection. *arXiv:2501.03940*. DOI: https://doi.org/10.48550/arXiv.2501.03940.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature* (Version 2). *arXiv:2301.11305*. DOI: 10.48550/ARXIV.2301.11305.

Mu, C., & Zhang, L. J. (2018). Understanding Chinese multilingual scholars' experiences of publishing research in English. *Journal of Scholarly Publishing*, *49*(4), 397–418. DOI: 10.3138/jsp.49.4.02.

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, *57*(10), 265. DOI: 10.1007/s10462-024-10903-2.

Murdoch, Y. D., Lim, H., & Cho, J. (2021). Writing center visitors: Influence of L1 writing skills on students' exophonic writings. *SAGE Open*, *11*(4). DOI: 10.1177/21582440211062234.

Odri, G. A., & Yoon, D. J. Y. (2023). Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics & Traumatology: Surgery & Research*, *109*(8), 103706. DOI: 10.1016/j.otsr.2023.103706.

Opara, C. (2025). Distinguishing AI-generated and human-written text through psycholinguistic analysis. In: *International Conference on Artificial Intelligence in Education* (pp. 212–219). Cham: Springer Nature Switzerland. DOI: 10.48550/arXiv.2505.01800.

Öztürk, Y., & Taşçı, S. (2023). A corpus-based analysis of lexical bundles in non-native post graduate academic writing and a potential L1 influence. *REFLections*, *30*(2), 488–505. DOI: 10.61508/refl.v30i2.267463.

Pan, F. (2018). A multidimensional analysis of L1–L2 differences across three advanced levels. *Southern African Linguistics and Applied Language Studies*, *36*(2), 117–131. DOI: 10.2989/16073614.2018.1476162.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. DOI: 10.1017/S0267190512000098.

Pedersen, A. M. (2011). Writing Across Languages, Disciplines, and Sources: Second Language Writers in Jordan. *Across the Disciplines*, *8*(1), 109–119. DOI: 10.37514/ATD-J.2011.8.1.06.

Picazo-Sanchez, P., & Ortiz-Martin, L. (2024). Analysing the impact of ChatGPT in research: P. Picazo-Sanchez and L. Ortiz-Martin. *Applied Intelligence*, *54*(5), 4172–4188. DOI: 10.1007/s10489-024-05298-0.

Popkov, A. A., & Barrett, T. S. (2025). AI vs academia: Experimental study on AI text detectors' accuracy in behavioral health academic writing. *Accountability in Research*, *32*(7), 1072–1088. DOI: 10.1080/08989621.2024.2331757.

Shin, Y. K. (2019). Do native writers always have a head start over nonnative writers? The use of lexical bundles in college students' essays. *Journal of English for Academic Purposes*, *40*, 1–14. DOI: 10.1016/j.jeap.2019.04.004.

Showemimo, P. (2025). The AI Paradox in Academia: When Writing Too Well Raises Red Flags. *SSRN*. DOI: 10.2139/ssrn.5255976.

Shchemeleva, I. (2021). There's no discrimination: The association between journals, title rephrasing of the given reviewer suggestions, prevalence of text-recycling, and publishers' policies in English. *Publications*, *9*(1), 8. DOI: 10.3390/publications9010008.

Shehata, A. M. K., & Eldakar, M. A. M. (2018). Publishing research in the international context: An analysis of Egyptian social sciences scholars' academic writing behaviour. *The Electronic Library*, *36*(5), 910–924. DOI: 10.1108/EL-01-2017-0005.

Subandowo, D., Sárdi, C., & Thresia, F. (2025). An investigation of English academic writing strategies employed by Indonesian graduate students in an English medium instruction (EMI) context. *Asian-Pacific Journal of Second and Foreign Language Education*, *10*(1), 38. DOI: 10.1186/s40862-025-00345-w.

Taşçı, S., & Öztürk, Y. (2021). Post-predicate that-clauses controlled by verbs in native and non-native academic writing: A corpus-based study. *Australian Journal of Applied Linguistics*, *4*(1), 18–33. DOI: 10.29140/ajal.v4n1.486.

Tufts, B., Zhao, X., & Li, L. (2025). A practical examination of AI-generated text detectors for large language models. In: *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 4824–4841). DOI: 10.48550/arXiv.2412.05139.

Uchendu, A., Le, T., Shu, K., & Lee, D. (2020). Authorship attribution for neural text generation. In: *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP) (pp. 8384–8395). Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.673.

Valipour, V., Assadi, N., & Asl, H. D. (2017). The generic structures and lexico-grammaticality in English academic research papers. *Southern African Linguistics and Applied Language Studies*, *35*(2), 169–182. DOI: 10.2989/16073614.2017.1373365.

Wang, X., Zhang, W., & Rajtmajer, S. (2024). Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. *arXiv:2410.18390*. DOI: 10.48550/arXiv.2410.18390.

Weber-Wulff, D., Anohina-Naumeca, A., & Bjelobaba, S. L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1–39. DOI: 10.1007/s40979-023-00146-z.

Won, D., Shin, Y. K., Kim, H. & Yoo, I. W. (2025). Advancing Language Assessment with GPT: Is It Nonnative-Language Friendly? *Language Assessment Quarterly*, *1*, 1–18. DOI: 10.1080/15434303.2024.2444349.

Zaini, A., & Ollerhead, S. (2019). Reverse contrastive rhetoric in expository writing: Transfer and power relations at work. *Southern African Linguistics and Applied Language Studies*, *37*(1), 41–61. DOI: 10.2989/16073614.2019.1609364.

Appendix A

This appendix provides a summary of the five proxies coded in this study, along with their computational operationalization and flagging criteria. All coding, feature extraction, and analysis were conducted programmatically in Python. Complete technical specifications, including software libraries, mathematical formulas, and reproducibility documentation, are available in the full Technical Codebook upon request.

Coding scheme

Feature	Definition	Computational Measure	Flagging Threshold
F1: Low Perplexity	Text predictability to language model	GPT-2 perplexity score	≤ 25th percentile
F2: Low Burstiness	Uniformity of word probability	Variance of GPT--2 log probabilities	≤ 25th percentile
F3: High Formulaic Language	Density of recurrent multi--word sequences	Bundle coverage (3-5 grams, 3+ occurrences)	≥ 75th percentile
F4: Consistent Formality	Uniformity of register across sentences	Variance of Heylighen--Dewaele formality scores	≤ 25th percentile
F5: Low Lexical Diversity	Repetitiveness of vocabulary	HD-D lexical diversity score	≤ 25th percentile

Appendix B

We present four examples representing different flag patterns across genres and sections. Example B is extracted from a technical paper. Example B is an excerpt from a technical paper coded as "Technical 1". The text is extracted from the results section. This extract represents Window 46, which received a Suspicious Score of 5/5, Flagging features F1, F2, F3, F4, F5 (consult coding scheme). Here is the excerpt:

> "enables a deeper understanding of the model's predictive and explanatory power, which will be discussed in the following section. To assess the significance of the path coefficients, bootstrapping with 5,000 sub--samples was conducted in SMART PLS 4.0. The path analysis revealed complex relationships..."

The specifications of the flagged window are reported here for further illustration. This window achieves a perfect 5/5 score through extreme z-scores

across all features. The text is highly predictable (z = -1.88 perplexity), extremely smooth (z = -2.97 burstiness), formulaic ("path coefficients", "bootstrapping", "sub-samples"), maintains consistent formal register (z = -0.84), and uses limited vocabulary (z = -2.38 lexical diversity). This pattern is genre-appropriate for technical results reporting with statistical notation and methodological terminology.

FEATURE PROFILE:

F1 (Low Perplexity):	FLAGGED (z = -1.88, PPL = 364.2)
F2 (Low Burstiness):	FLAGGED (z = -2.97, Burst = 1.64)
F3 (High Formulaic):	FLAGGED (z = 2.06, Coverage = 0.176)
F4 (Formality Variance):	FLAGGED (z = -0.84, Var = 0.175)
F5 (Low Lexical Diversity):	FLAGGED (z = -2.38, HD-D = 0.783)

Example C is an excerpt from a technical paper coded as "Technical 1". The text is extracted from the methods section. This extract represents Window 16, which received a Suspicious Score of 2/5, Flagging features F2, F4 (consult coding scheme). Here is the excerpt:

> "understand how students engage with educational technologies (Granić & Maranguni, 2019). In the realm of language teaching and learning, TAM has been employed to examine the importance of technology acceptance among both educators and students. For educators, accepting technology is essential for..."

The specifications of the flagged window are reported here for further illustration. This window achieves a 2/5 score. This methodological description shows moderate flagging (2/5). The text maintains a consistent formal register (F4 flagged, z = -2.25) and smooth flow (F2 flagged), but perplexity is actually high (z = 1.26), indicating varied sentence structures and conceptual explanations. This is consistent with the possibility that not all technical writing triggers all

flags; the literature review and methods sections show more linguistic variation than pure results reporting.

FEATURE PROFILE:

F1 (Low Perplexity):	NOT FLAGGED (z = 1.26, PPL = 568.8)
F2 (Low Burstiness):	FLAGGED (z = -0.83, Burst = 2.30)
F3 (High Formulaic):	NOT FLAGGED (z = -1.30, Coverage = 0.081)
F4 (Formality Variance):	FLAGGED (z = -2.25, Var = 0.154)
F5 (Low Lexical Diversity):	NOT FLAGGED (z = 0.16, HD-D = 0.853)

Example D is an excerpt from a technical paper coded as "Linguistics 2". The text is extracted from the literature review section. This extract represents Window 24, which received a Suspicious Score of 3/5, Flagging features F3, F4, F5 (consult coding scheme). Here is the excerpt:

> "reflects the overall acceptability of the pragmatic functions linked to the religious marker … in everyday spoken Jordanian Arabic (SJA). This benchmark precedes the analysis of how face-related factors influence acceptability rates. To capture these preferences, a 5-point L…"

The specifications of the flagged window are reported here for further illustration. This window achieves a 3/5 score. This theoretical framework section shows a different flag pattern (F3, F4, F5 without F1/F2). The text is formulaic (repeated conceptual terminology: "acceptability", "pragmatic functions", "face-related factors") and maintains consistent formality, but perplexity and burstiness are NOT flagged because the author introduces new concepts and examples (phonetic transcription, language-specific terms). This illustrates how theoretical exposition differs from statistical reporting.

FEATURE PROFILE:

F1 (Low Perplexity):	NOT FLAGGED (z = -0.61, PPL = 437.8)
F2 (Low Burstiness):	NOT FLAGGED (z = 1.04, Burst = 2.77)
F3 (High Formulaic):	FLAGGED (z = 1.47, Coverage = 0.179)
F4 (Formality Variance):	FLAGGED (z = -1.01, Var = 0.003)
F5 (Low Lexical Diversity):	FLAGGED (z = -0.83, HD-D = 0.764)

Our conclusions are constrained by the small, personal nature of the corpus, the focus on a single L1 background and disciplinary area, and the use of a simulator rather than direct access to commercial detectors. All analyses are cross-sectional and observational and thus cannot establish causal relationships between specific textual properties and detector behaviour. Future research with larger, more diverse samples and direct access to detection outputs is needed to test and refine the hypotheses advanced here.