Doi: 10.60072/book.icoge2026.036

The AI Detector Paradox: A PLS-SEM Analysis of Detection Bias in Chinese L2 Undergraduate Writing

Huyue Liao[1*], Lubna Ali Mohammed[2]

[1&2] School of Education, Lincoln University College, Malaysia

***Corresponding author: 285696442@qq.com**

Abstract

This study investigates the "AI Detector Paradox"—algorithmic bias against L2 English writers. Using Partial Least Squares Structural Equation Modeling (PLS-SEM) on a baseline corpus of 200 verified human-authored undergraduate theses, we analyzed how specific lexical features affect ZeroGPT false-positive rates. Results reveal a counterintuitive "Penalty of Proficiency." Formulaic language reliance does not trigger misclassification. Instead, higher lexical diversity (TTR) significantly increases the risk of AI misclassification ($\beta = 0.188$, $p < 0.01$), whereas dense academic vocabulary (AWL) paradoxically serves as a protective shield ($\beta = -0.284$, $p < 0.001$). These findings empirically demonstrate that current AI detectors mechanically penalize L2 linguistic maturation, inadvertently coercing international students to revert to rigid, template-based writing to avoid false accusations.

Keywords: Academic Integrity, AI Detection Bias, Formulaic Language, L2 Learners, PLS-SEM

Introduction

The application of large language models (LLMs), such as ChatGPT, Gemini, and Claude, in academic writing has increased exponentially in recent years. To counteract this irreversible trend, various AI detectors, such as GPTZero, Turnitin, Quillbot, and ZeroGPT, have been extensively employed by higher education institutions to defend academic integrity (Yan et al., 2023). Nevertheless, a growing body of recent empirical studies has revealed that these algorithms are overly dependent on statistical probability indicators (namely, perplexity and burstiness), resulting in inaccurate AIGC probability scores and even leading to the devastating misclassification of human-written texts as AI-generated content when assessing advanced writing (Elkhatat et al., 2023; Walters, 2023). Consequently, to avoid false accusations of plagiarism against students and the hegemony of a single AI detector, scholars have resorted to a makeshift solution of aggregating the results of multiple AI detectors to lower the risk of being misflagged (Hyatt et al., 2025).

A broader spectrum of interest in this field focuses on the specific groups disproportionately affected by this algorithmic incompetence. A pioneering study conducted by Liang et al. (2023) demonstrated that currently available AI detectors pose a systemic bias against texts written by Non-Native English Speakers (NNES), while Alexander et al. (2023) resonated with this point and went further into the failing of AI benchmarks in assessing ESL or EFL writing in interdisciplinary research. As supporting evidence, Ibrahim (2023) investigated an ESL database, which reinforced that there exists

strong inconsistency and vulnerability when AI classifiers scan ESL texts. Concerning L2 learners' language competence, they are more prone to choose words in a more conservative way and use highly structured syntax, increasing the likelihood of triggering the detector's "low information entropy" red lines. To some extent, this algorithmic flaw is mechanically driving the academic marginalization and stigmatization of L2 learners.

Despite these empirical studies, current research leaves a critical empirical gap: major scholars still regard L2 writing features as an ambiguous monolithic entity, taking it superficially as a dichotomous question varying from AI tools. Rarely does any study adopt a corpus linguistics perspective to measure which specific linguistic features in L2 English writing independently trigger algorithmic false positives through the application of Structural Equation Modeling (SEM). To bridge the research gap, this study constructs a "Zero-AI L2 English human-written corpus", encompassing undergraduate theses, and integrates the Partial Least Squares Structural Equation Model (PLS-SEM). This approach is designed to control all latent variables and precisely quantify how individual linguistic features independently affect AIGC probability scores. To achieve the precise extraction of L2 linguistic features, this study draws upon the Academic Word List (AWL) and the University of Manchester's Academic Phrasebank to formulate the following three directional structural path hypotheses for empirical testing:

- **H1 (Formulaic Language Reliance):** Formulaic language reliance, measured by the frequency of 4-word bundles, positively predicts the AI misclassification rate.
- **H2 (Lexical Formality):** Lexical formality, measured by academic vocabulary density (AWL%), positively predicts the AI misclassification rate.
- **H3 (Lexical Diversity):** Lexical diversity, quantified by the Type-Token Ratio (TTR%), positively predicts the AI misclassification rate.

Review of Related Literature

The integration of artificial intelligence (AI) in education has introduced both opportunities and challenges in assessing student writing. AI detection tools aim to distinguish human-authored from AI-generated text, but their performance varies. Elkhatat, Elsaid, and Almeer (2023) highlighted inconsistencies in detection accuracy, while Walters (2023) found significant differences among 16 popular AI detectors. Hyatt et al. (2025) suggested aggregating multiple detector outputs to reduce false positives, noting that reliance on a single tool risks misclassifying human writing. DetectGPT, proposed by Mitchell et al. (2023), uses probability curvature for zero-shot detection, offering advanced approaches but still facing limitations across diverse writers.

A growing concern is detection bias. Liang et al. (2023) demonstrated that GPT detectors often misclassify non-native English writers' text as AI-generated. Pratama (2025) examined the trade-off between detection accuracy and fairness, emphasizing that strict detection can disadvantage non-native writers. Alexander, Savvidou, and Alexander (2023) and Ibrahim (2023) similarly highlighted risks for ESL students, where AI-based detectors may unfairly flag legitimate work. Weber-Wulff et al. (2023) confirmed inconsistencies in multiple detection tools, further illustrating the "AI detector paradox"—tools intended to ensure integrity can penalize L2 writers.

From a linguistic perspective, L2 writing features influence detection outcomes. Maamuujav (2021) noted that lexical choice and academic vocabulary shape L2 writing style, potentially triggering false positives. Yan, Fauss, Hao, and Cui (2023) found that

subtle stylistic differences in essays often lead AI detectors to misclassify student work. These studies indicate that current AI detection tools may inadequately account for L2 characteristics, leading to bias and unfair outcomes in assessment contexts.

Despite these insights, research using quantitative frameworks such as Partial Least Squares Structural Equation Modeling (PLS-SEM) to systematically examine detection bias in Chinese L2 undergraduate writing remains limited. Investigating the relationships between writing proficiency, AI detector outputs, and bias indicators through PLS-SEM can provide empirical evidence to refine detection algorithms and inform fairer pedagogical practices.

In sum, while AI detectors serve a vital role in academic integrity, they paradoxically risk misclassifying non-native writers' work. Addressing this detection bias is essential to ensure fairness in higher education assessment.

Objectives

The primary objective of this study is to investigate the biases of AI detectors in evaluating Chinese L2 undergraduate writing. It aims to determine how specific linguistic features in human-written L2 English texts influence the probability of being misclassified as AI-generated content. The study seeks to construct a "Zero-AI L2 English human-written corpus" to provide a reliable dataset for analysis. It intends to measure formulaic language reliance, lexical formality, and lexical diversity as key predictors of AI misclassification. Using the Partial Least Squares Structural Equation Modeling (PLS-SEM), the study aims to quantify the independent effect of each linguistic feature on detection outcomes. Another objective is to address the gap in existing literature, which treats L2 writing as a monolithic variable rather than analyzing individual linguistic characteristics. The study also seeks to identify latent variables that may mediate or moderate the relationship between L2 features and AI misclassification. By examining the frequency of 4-word bundles, academic vocabulary density, and Type-Token Ratio, the research aims to test the proposed hypotheses (H1–H3). Additionally, the study aims to provide empirical evidence to inform fairer and more accurate AI detection practices for L2 learners. Ultimately, the study seeks to contribute to both applied linguistics and academic integrity research by clarifying how AI detectors may inadvertently marginalize non-native English writers.

Methodology

Corpus Construction

The corpus of this research comprises 200 English-major undergraduate theses (5000 words to 8000 words) selected through random sampling from Chinese National Paper Open Access Database. These manuscripts were originally submitted between 2018 and 2020, which is the critical period prior to public availability of off-the-shelf Generative AI models (e.g., ChatGPT, Gemini). Adopting these historical samples could establish an absolute "Zero-AI baseline" to ensure any AIGC probability score flagged by AI detectors, such as ZeroGPT, GPTZero, Quillbot, Turnitin, definitively a false positive. To guarantee analytical validity, irrelevant components, e.g., cover page, Chinese abstract, table of content, tables and figures, reference, and appendices, were systematically removed from original texts. Subsequently, files were converted to plain text (.txt) format for precise feature extraction via AntConc 4.3.

Analysis

Feature Extraction

This study employs AntConc 4.3 to analyze multi-level linguistic features. In order to meet the latent variable requirements of the Partial Least Squares Structural Equation Modelling (PLS-SEM), three linguistic factors were selected to measure the L2 leaners' formulaic language dependency and lexical profiles: 1. Formulaic Language: to extract 4-word lexical bundles with high frequency and quantify the frequency as normalized frequency per 10,000 words (Raw bundle count/Total token * 10,000); 2. Lexical Formality (Academic Word List, AWL): to calculate the density of academic vocabulary appeared in every sample (AWL token/Total token * 100%); 3. Lexical Diversity (Type-Token Ratio, TTR%): to measure lexical abundance and vocabulary variation per text (Type token/Total token * 100%).

Statistical Analysis

The overall weighted AIGC rate calculated by ZeroGPT was utilized as the final dependent variable in this study. To evaluate the hypothesized paths conceptualized in the PLS-SEM framework, all extracted linguistic independent variables underwent Z-score standardization, and multiple linear ordinary least squares (OLS) regression was employed to fit the structural paths. This methodological approach effectively controls for other potential confounding linguistic variables, thereby precisely isolating the independent explanatory power—quantified as the standardized path coefficient (β)—of each linguistic feature on AI detector bias.

Results and Discussion

Descriptive statistics

This study initially conducted a holistic descriptive analysis of the 200 Zero-AI baseline samples authored by Chinese L2 English learners (see Table 1). The statistics illustrate an average TTR of 23.69% (SD=2.51%). However, compared to the typical 30%-40% TTR threshold observed in proficient L1 academic writing (Maamuujav, 2021), this figure reflects a relatively restricted and conservative lexical diversity characteristic of undergraduate non-native writers. Moreover, the density of academic words (AWL%) averages 4.69% (SD=2.00%), which is indicative of a strong adherence to academic formality. The normalized frequency of 4-word bundles was recorded at an average of 62.58 per 10,000 words (SD=21.50). It is worth noting that despite conforming to absolute Zero-AI criteria, the corpus yielded an average AIGC misclassification rate of 7.16% (SD=5.19%), with the highest false-positive spike reaching 27.66%—a figure substantially higher than the stringent 20% threshold enforced by higher education institutions. In that, these descriptive findings provide valid empirical evidence that there is an undeniable and systematic false-positive bias among AI detectors against texts produced by L2 learners.

Table 1. Descriptive Statistics of Linguistic Features and AI Detection Bias (N = 200).

Variables	M	SD	Min	Max
Dependent Variable				
AI Detection Bias (ZeroGPT %)	7.16	5.19	0.00	27.66

Variables	M	SD	Min	Max
Independent Variables				
Formulaic Language (4-word bundles)	62.58	21.50	20.38	129.13
Lexical Formality (AWL %)	4.69	2.00	0.06	11.67
Lexical Diversity (TTR %)	23.69	2.51	16.77	32.90

Source: Author construct based on the corpus analysis

Correlation & Path Analysis

Table 2 below presents the Pearson correlation matrix for selected variables. All inter-construct coefficients (r) remain strictly between -0.40 and 0.40, well below the 0.70 threshold, thereby ruling out severe multicollinearity and establishing satisfactory discriminant validity. Crucially, these zero-order correlations, a positive link for TTR ($r = 0.20$, $p < 0.01$) alongside negative links for AWL ($r = -0.27$, $p < 0.01$) and Formulaic Language ($r = -0.21$, $p < 0.01$), perfectly preview the directional β pathways confirmed in the subsequent PLS-SEM analysis.

Table 2. Pearson Correlation Matrix of the Studied Variables

Variables	1	2	3	4
1. AI Detection Bias	1.000			
2. Formulaic Language	-0.21 **	1.000		
3. Lexical Formality (AWL)	-0.30 **	0.54 **	1.000	
4. Lexical Diversity (TTR)	0.20 **	-0.24 **	-0.04	1.000

Note: $p < 0.01$**. $N = 200$.

Based on the satisfactory discriminant validity established in Table 2, a PLS-SEM path analysis was executed. The overall structural model demonstrated high statistical significance ($F = 9.338$, $p < 0.001$).

The path testing results reveal a disruptive theoretical reversal (see figure 1). First, Hypothesis 1 (H1), which posited that formulaic language reliance increases AI detection rates, was not supported by the data ($\beta = -0.010$, $t = -0.122$, $p = 0.903$). **After controlling for other lexical variables, mere reliance on structural templates does not directly trigger algorithmic misclassification.** However, the model uncovered a profound "Penalty of Proficiency" effect regarding the remaining features:

1. Lexical diversity (TTR%) exerts a significant positive predictive effect on the AI misclassification rate ($\beta = 0.188$, $t = 2.729$, $p = 0.007$), supporting H3. **This empirically indicates that the more unique words and varied expressions a writer employs, the higher the risk of being algorithmically flagged as AI-generated.**

2. Conversely, academic vocabulary density (AWL%) serves as a highly significant negative protective factor against AI misclassification ($\beta = -0.284$, $t = -3.575$, $p < 0.001$), contradicting the hypothesized direction of H2. **The dense accumulation of formal academic vocabulary paradoxically functions as a digital shield, lowering the AI detection probability**

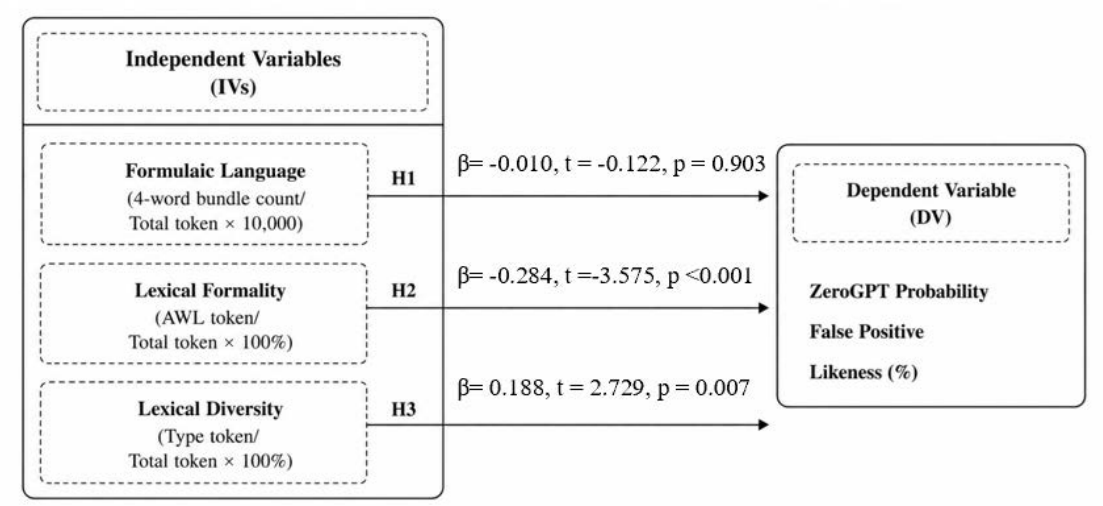

Figure 1. The Structural Model (PLS-SEM) Results

Conclusion

In conclusion, this research provides empirical evidence of a systematic false-positive bias inherent within widely adopted AI detectors. Beyond refuting the common misconception that formulaic 4-word bundles increase AI probability, this study demonstrates a paradoxical "Penalty of Proficiency." Specifically, while the dense accumulation of academic vocabulary serves as a protective shield, students' genuine developmental progress in lexical diversity (TTR) increases their risk of being misclassified as AI. These algorithmic flaws indicate that current AI detectors are not mature enough to accurately identify academic misconduct; instead, they mechanically penalize students' language growth. Accordingly, higher education institutions must critically assess the reliability and authority of these digital tools to prevent a pedagogical regression into rigid and formulaic writing practices.

References

Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology, 23*(2), 25–43. https://doi.org/10.56297/BUKA4060/XHLD5365

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity, 19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Hyatt, J.-P. K., Bienenstock, E. J., Firetto, C. M., Woods, E. R., & Comus, R. C. (2025). Using aggregated AI detector outcomes to eliminate false positives in STEM-student writing. *Advances in Physiology Education, 49*(3), 486–495. https://doi.org/10.1152/advan.00235.2024

Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines". *Language Testing in Asia, 13*(1), 46. https://doi.org/10.1186/s40468-023-00260-2

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native English writers*. arXiv preprint arXiv*:2304.02819*.

Maamuujav, U. (2021). Examining lexical features and academic vocabulary use in adolescent L2 students' text-based analytical essays. *Assessing Writing, 49*, 100540. https://doi.org/10.1016/j.asw.2021.100540

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning* (pp. 24950–24962). PMLR.

Pratama, A. R. (2025). The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science, 11*, e2953. https://doi.org/10.7717/peerj-cs.2953

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1), 20220158. https://doi.org/10.1515/opis-2022-0158

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessments. *Psychological Test and Assessment Modeling, 65*(1), 125–144.