Journal of Humanities and Social Sciences Studies
ISSN: 2663-7197
DOI: 10.32996/jhsss
Journal Homepage: www.al-kindipublisher.com/index.php/jhsss

JHSSS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Comparing Linguistic Features between Human-written High-Scoring IELTS Essays and AI-Generated ones

Jinliang Wu

Undergraduate Student, School of Foreign Languages, Guizhou Medical University, Guiyang, China

Corresponding Author: Jinliang Wu**, E-mail:** 2093313576@qq.com

| ABSTRACT

Using two self-constructed corpora (25 essays each for Task 1 and Task 2), this study systematically compares the linguistic characteristics of high-scoring human IELTS essays (Simon's model essays, Band 7+) with Doubao-generated texts across four core dimensions: lexical resources, syntactic complexity, cohesion and coherence, and task response. Quantitative analyses employ lexical indices from McCarthy & Jarvis (2010), syntactic complexity metrics from Lu (2010), Halliday & Hasan's (1976) cohesion framework, and Stapleton & Wu's (2015) task response evaluation model. Results indicate that AI-generated essays slightly outperform human essays in lexical diversity and complexity (e.g., Shannon Entropy, MM Entropy) and maintain high stability in task response metrics (SSI, RQA, SRM). Syntactic complexity, however, exhibits task-specific patterns: in Task 1, AI demonstrates "breadth complexity," expanding surface structures with 1.7–2 times more basic units (e.g., words, verb phrases) than humans and relying heavily on coordination and complex noun phrases. In Task 2, humans achieve "deep logical complexity," constructing nested dependent structures (higher DC/C and DC/T) and producing denser verb phrases (VP/T), supporting more efficient and precise argumentation. In cohesion and coherence, AI exhibits notable weaknesses, including limited referential diversity, repetitive conjunctions, and insufficient implicit cohesion, whereas human texts display superior logical layering, natural collocation, and deep semantic association. These findings highlight the complementarity between AI and human writing, offering empirical support for leveraging AI in IELTS teaching and guiding future model optimization to balance structural precision with nuanced contextual contextual adaptability.

| KEYWORDS

IELTS writing; AI-generated texts; human high-scoring essays; lexical resources; syntactic complexity; cohesion and coherence; task response

| ARTICLE INFORMATION

ACCEPTED: 19 July 2025 **PUBLISHED:** 26 August 2025 **DOI:** 10.32996/jhsss.2025.7.8.8

1. Introduction

With the breakthrough progress of Large Language Models (LLMs) represented by Doubao in natural language processing, their application in education, especially in language learning and writing assistance, is rapidly expanding. These AI tools can quickly generate grammatically accurate and structurally complete texts, offering unprecedented convenience to language learners. However, this AI boom has also triggered academic reflection: Do the deep linguistic features of high-quality AI-generated texts share homogeneity with those of high-level human writers? Particularly in the context of IELTS, a rigorous international standardized test, exploring whether AI texts can truly replicate the linguistic essence of high-scoring human essays has become a topic of significant theoretical and practical significance.

Current research on AI writing mainly focuses on three areas: first, macro-level comparisons of features between AI and human texts, such as differences in lexical diversity and syntactic complexity (Zindela, 2024); second, exploring the impact of AI writing

tools on learners'writing output and teaching effectiveness (Wei, Wang, & Dong, 2023; Zare, Al-Issa, & Ranjbaran Madiseh, 2025); third, analyzing the generative logic, controllability, and potential risks of AI texts (Derner & Batistič, 2023; Weidinger et al., 2021). These studies provide valuable insights into AI writing but are generally unspecific, using general corpora or focusing on teaching applications, with few in-depth, multi-dimensional quantitative analyses in specific test contexts.

Additionally, few studies directly and systematically compare outputs from specific AI models (e.g., Doubao) with specific high-scoring human corpora (e.g., Simon's model essays). This study aims to fill this gap by empirically analyzing 50 matched corpora to systematically compare the linguistic features and advantages/disadvantages of high-scoring human IELTS essays and AI-generated essays across four core evaluation dimensions.

This study goes beyond macro comparisons by introducing a fine-grained quantitative indicator system: lexical resources are measured using indicators by McCarthy & Jarvis (2010); syntactic complexity adopts Lu's (2010) computational model; cohesion and coherence are analyzed based on Halliday & Hasan's (1976) classic theory; task response follows Stapleton & Wu's (2015) evaluation model. The innovation of this study lies in its focus on a highly structured writing scenario (IELTS) and its provision of a new empirical perspective on the complementarity and differences between AI and humans in high-scoring writing through precise corpus matching and multi-dimensional quantitative analysis. The results will provide solid theoretical support for the effective use of AI tools in IELTS teaching and the optimization of language models in academic writing.

2. Literature Review

In recent years, the emergence of large pre-trained language models (LLMs), notably the GPT series, has made the linguistic characteristics of AI-generated texts a key research focus in computational and applied linguistics. Macro-level studies using lexical diversity metrics (e.g., TTR, MTLD) indicate that AI demonstrates advantages in lexical breadth but often lacks natural collocation and contextual adaptability (McNamara, Crossley, & McCarthy, 2010). At the syntactic level, AI-generated texts tend to exhibit moderate complexity, favoring standardized sentence patterns while underutilizing the nuanced structures and rhetorical devices that humans employ for specific communicative purposes (Lu, 2011). While these studies establish foundational insights, they are often single-dimensional and fail to provide a comprehensive evaluation framework. IELTS writing, with its multi-dimensional scoring system, offers a suitable context for systematic analysis.

Prior research on high-scoring human IELTS essays highlights that superior cohesion and coherence depend not only on explicit conjunctions but also on reference, substitution, and implicit logical relationships (Halliday & Hasan, 1976). High scores in task response require clear argumentation, well-supported evidence, and rigorous logical reasoning rather than mere task completion (Stapleton & Wu, 2015). However, most existing studies lack direct comparisons with AI-generated texts, leaving AI performance in complex academic writing largely unexplored.

Current research limitations include: (1) non-specific corpora—few studies compare high-scoring human essays (e.g., Simon's model essays) with advanced AI models (Doubao), reducing the applicability of conclusions; (2) single-dimensional frameworks—most focus on either vocabulary or grammar, neglecting an integrated evaluation of lexical resources, syntactic complexity, cohesion/coherence, and task response; (3) theoretical emphasis—findings often remain abstract, offering limited guidance for AI-assisted IELTS instruction.

This study addresses these gaps by situating IELTS writing as a specific research context and constructing a multi-dimensional framework encompassing lexical resources, syntactic complexity, cohesion/coherence, and task response. It systematically compares high-scoring human essays with AI-generated texts using rigorous quantitative measures to derive targeted, practical insights. The theoretical foundation for the four-dimensional framework is as follows: Task response: Based on Stapleton & Wu (2015), assessing not only task completion but also reasoning quality (RQA) and structure-reasoning matching (SRM), distinguishing whether AI outputs are superficial or logically substantive. Cohesion and coherence: Drawing on Halliday & Hasan (1976), including reference, substitution, and ellipsis, allowing quantification of cohesion diversity and discourse fluency, and identifying reliance on repetitive or rigid patterns. Lexical resources: Using McCarthy & Jarvis (2010) metrics, particularly MTLD, to capture both vocabulary quantity and variability, revealing whether AI's lexical richness reflects mechanical combination or contextual adaptability. Syntactic complexity: Based on Lu (2010), including measures such as mean length of sentence (MLS), mean length of T-unit (MLT), and clauses per sentence (C/S), to assess whether AI demonstrates syntactic variation comparable to high-scoring human writers.

In sum, this study integrates these models into a systematic, multi-dimensional analytical framework, aiming to uncover the comparative strengths and weaknesses of humans and AI in high-scoring IELTS writing and providing an empirical basis for optimizing AI applications in language education.

3. Methodology

This study employs a multi-dimensional analysis approach centered on the quantitative comparison of linguistic features. The research focuses on four core dimensions, with each dimension's analytical framework constructed based on key literature. The Lexical Resources dimension integrates lexical diversity metrics such as MTLD and vocd-D (McCarthy & Jarvis, 2010), cognitive information theory (lexical complexity and density) from Zhao et al. (2023), and lexical density research by Liu, Z., & Dou, J. (2023) to evaluate diversity, accuracy, complexity, and density. The Syntactic Complexity dimension utilizes 14 syntactic complexity metrics (e.g., MLS, MLT, C/T) from Lu (2010) to quantify syntactic complexity and accuracy. The Cohesion and Coherence dimension is based on the framework by Halliday & Hasan (1976). The Task Response dimension is based on the framework by Stapleton & Wu (2015), which has been refined with quantitative rules for surface-level completeness in IELTS Writing Task 1 and the quality of reasoning and structural matching in Task 2. The relative strengths and weaknesses of the two groups will be determined through statistical tests. The study systematically examines these differences between high-scoring human-written IELTS essays and AI-generated essays using a matched corpus of 50 essays. This corpus is composed of four parallel sub-corpora (two for Task 1 and two for Task 2, each containing 25 essays), with strict control over the comparability of writing tasks and scoring criteria.

3.1 Corpora

This study constructed an AI corpus using Doubao-generated texts for 50 identical IELTS topics—25 Task 1 (chart analysis) and 25 Task 2 (argumentative)—matched with a human baseline of high-scoring model essays (Band 7+) from former IELTS examiner Simon. Prompts for AI generation explicitly incorporated IELTS scoring criteria—lexical resource, grammatical range and accuracy, coherence and cohesion, and task achievement—to ensure targeted, standardized outputs. Strict control of variables ensured both groups were aligned in topic, text type, and length, enabling valid inter-group (AI vs. human) and intra-group (Task 1 vs. Task 2) comparisons.

Task selection leveraged the distinct cognitive and linguistic demands of each writing type. Task 1 requires objective reporting of visual data, prioritizing linguistic precision, logical structure, and organizational clarity—ideal for testing AI's capacity to process and convey factual information. Task 2 demands persuasive argumentation on abstract or social issues, integrating complex reasoning, deeper argument development, and personal style—allowing assessment of AI's ability to emulate higher-order thinking and rhetorical strategy. This dual-task design enables a comprehensive mapping of comparative strengths and weaknesses across writing contexts.

Prompting strategies were task-specific. For Task 1, instructions required ≥150 words, accurate trend summaries, key feature highlighting, precise data references, and the avoidance of subjective opinions, alongside varied cohesive devices, precise and varied vocabulary, and error-free grammar. For Task 2, instructions required ≥250 words, a clear stance, well-developed supporting arguments, cohesive paragraphing, accurate academic vocabulary, and flexible complex sentence structures with full grammatical and punctuation accuracy.

Table 1. Corpora Size of four Groups (words)

	Simon-written compositions	Doubao-generated compositions	Simon-written compositions	Doubao-generated compositions
Category	Task1	Task1	Task2	Task2
1	159	305	288	290
2	170	333	271	333
3	170	347	273	313
4	159	295	272	324
5	184	319	268	347
6	178	310	257	339
7	157	303	241	306
8	176	336	253	338
9	150	248	256	333
10	170	272	264	327
11	156	337	274	302
12	168	309	252	296
13	189	322	259	284
14	162	358	257	297
15	162	322	263	355

16	162	316	250	315
17	163	268	263	262
18	179	337	258	291
19	167	298	248	278
20	155	428	289	282
21	188	336	267	292
22	175	284	267	342
23	181	336	279	429
24	181	342	272	550
25	157	345	303	519
Average	169.12	320.24	284.75	466.44
Sum	4218	8006	6644	8344

3.3 Data processing

This study employs a multi-dimensional framework to compare human-written and AI-generated IELTS essays across four linguistic dimensions: lexical resources, syntactic complexity, cohesion and coherence, and task response.

Quantitative analysis reveals that AI-generated texts maintain consistent advantages in lexical diversity, complexity, and task response scores, while human texts demonstrate superior adaptability in lexical density and contextual relevance, often using personalized expression to enhance logic and rhetorical impact.

In the lexical resources dimension, diversity is assessed through TTR, MTLD, HD-D, and Maas, following McCarthy and Jarvis (2010), while complexity is measured via Shannon and MM Entropy. Lexical accuracy is evaluated through BFF and COAF indicators, capturing the proportion of samples meeting set criteria. Lexical density is calculated from POS-tagged corpora, with both overall and standardized values compared for cross-task consistency.
Syntactic complexity, based on Lu (2010), shows that AI texts in Task 1 exhibit greater "breadth complexity," expanding surface structures and generating 1.7–2× more verb phrases than human texts. Conversely, high-scoring human Task 2 essays display "deep logical complexity" through more subordinate structures, balancing concision with argumentative depth.

Cohesion and coherence, analyzed via Halliday and Hasan (1976) and Coh-Metrix, indicate that human writers retain an irreplaceable advantage in contextual adaptation—employing varied cohesive devices, layered logic, and natural collocations—especially in argumentative contexts where flexible cohesion boosts persuasiveness.

Task response evaluation, drawing on Stapleton and Wu (2015), shows that both groups achieve 100% Surface Structural Integrity (SSI), while AI consistently attains full scores in Reasoning Quality (RQA) and Structure–Reasoning Match (SRM). Human texts score slightly lower due to occasional ambiguities, yet surpass AI in stylistic individuality and emotional resonance through precise deep-contextual adaptation.

4.Results and discussion

4.1 Comparison of Lexical resources between Human-written and Doubao-generated Compositions

Table 2 presents lexical diversity analysis for 25 IELTS Writing Task 1 and Task 2 essays each from Simon (human) and Doubao (AI), measured by four metrics: TTR (Type-Token Ratio), MTLD (Moving-Average Type-Token Ratio), HD-D (Herdan's D), and Maas (corrected lexical diversity).

Table 2. Lexical diversity of task1 and task 2 between Simon-written and Doubao-generated Compositions

Group	Simon-written compositions				Doubao-generated compositions			
Metrics	TTR	MTLD	HD-D	Maas	TTR	MTLD	HD-D	Maas
Task1	0.5876	46.1516	35.472	0.05348	0.59148	46.3736	35.7012	0.05424
Task2	0.58808	46.0564	35.366	0.05344	0.5876	46.1516	35.472	0.05348

Table 2 compares lexical diversity between human-written (Simon) and AI-generated (Doubao) essays in IELTS Writing Tasks 1 and 2 using four metrics: TTR, MTLD, HD-D, and Maas.

For Task 1, AI slightly outperforms humans across all indicators (e.g., TTR: 0.59148 vs. 0.5876; MTLD: 46.3736 vs. 46.1516), reflecting AI's stable lexical richness and diversity, which suits descriptive tasks like chart interpretation. Humans, however, possess superior lexical adaptation and deeper understanding of social implications behind data.

In Task 2, differences diminish. Humans lead marginally in TTR (0.58808 vs. 0.5876), indicating greater basic vocabulary flexibility in argumentative writing, while AI maintains minor advantages in MTLD, HD-D, and Maas. Humans demonstrate better integration of cultural and contextual nuances, whereas AI upholds stable diversity but lacks emotional and creative adaptation for deeper engagement.

Table 3 presents lexical complexity measured by Shannon Entropy and MM Entropy for 25 essays per task. Doubao's texts consistently score higher, confirming its advantage in vocabulary complexity and balanced lexical distribution across both tasks.

Table 3. Lexical complexity of task1 and task 2 between Simon-written and Doubao-generated Compositions

Group	Simon-written compositions		Doubao-generated compositions	
Metrics	Shannon Entropy	MM Entropy	Shannon Entropy	MM Entropy
Task1	4.5596	4.8268	4.6328	4.8828
Task2	5.212	5.4624	5.6624	5.8624

Based on 25 Task 1 essays, lexical complexity was compared between Simon's human-written and Doubao's AI-generated texts using Shannon Entropy and MM Entropy.

Doubao scored higher on both metrics (4.6328 and 4.8828) than Simon (4.5596 and 4.8268), indicating greater lexical diversity and balance. In Task 2, entropy values for both writers increased, reflecting higher lexical demands of argumentative prompts. Doubao maintained its advantage, suggesting effective adaptation to complexity. Lexical density analysis (Table 4) showed Doubao's overall density and standardized scores slightly exceeded Simon's in Task 1, with similar variability (SD). In Task 2, Simon achieved marginally higher density scores, reflecting human strengths in dense vocabulary use for complex arguments. Stable SDs across tasks indicate comparable lexical fluctuation patterns.

Table 4 presents the lexical density, including Overall Density (OD), Standard Deviation (SD), and Standardization Value (SP), of Task 1 and Task 2 between Simon - written compositions and Doubao - generated compositions.

Table 4. Lexical density of task1 and task 2 between Simon-written and Doubao-generated Compositions

Group	Simon-written compositions			Doubao-generated compositions		
Metrics	OD	SD	SP	OD	SD	SP
Task1	0.4386	0.07316	438.6	0.44052	0.0736	440.52
Task2	0.42992	0.07166	429.92	0.42684	0.070696	426.84

Note: OD: Overall Density. SD: Standard Deviation. SP: Standardization Value

In Task 1 lexical density, human essays had a mean OD of 0.4386 (SD 0.07316; SP 438.6), while Doubao scored slightly higher at 0.44052 (SD 0.0736; SP 440.52), indicating denser and more stable vocabulary use by AI. Humans benefit from flexible lexical adjustment enhancing logic and appeal, but with marginally lower information density.

In Task 2, human essays outperformed Doubao with an OD of 0.42992 (SD 0.07166; SP 429.92) versus 0.42684 (SD 0.0707; SP 426.84), reflecting concentrated, experience-based vocabulary suited to complex arguments. Doubao maintained stable but more formulaic lexical organization. Lexical accuracy, measured by BFF, COAF, FMAF, CAF, and DDF metrics, showed Doubao's slight advantage in precision and stability (e.g., BFF 99.548 vs. 99.44; CAF 86.32 vs. 85.86; DDF 0.208 vs. 0.26). Simon's scores exhibited greater variability, reflecting human responsiveness to task complexity and creativity.
Table 5 presents lexical accuracy of Task 1 and Task 2 between Simon - written and Doubao - generated compositions, involving metrics (BFF, COAF, FMAF, CAF, DDF).

Table 5. Lexical accuracy of task1 and task 2 between Simon-written and Doubao-generated Compositions

Group	Simon-written compositions					Doubao-generated compositions				
Metrics	BFF	COAF	FMAF	CAF	DDF	BFF	COAF	FMAF	CAF	DDF
Task1	99.44	99.2	99.448	85.86	0.26	99.548	99.364	99.632	86.32	0.208
Task2	99.192	98.872	99.244	79.6	0.336	99.472	99.336	99.532	86.324	0.224

Note: BFF: Basic Frequency Features. COAF: Co-occurrence-Association Feature. FMAF: Form-Meaning Association Feature. CAF: Context Adaptation Feature. DDF: Deviation Degree Feature

In Task 1 lexical analysis, Simon's essays achieved high scores in BFF (99.44), COAF (99.2), and FMAF (99.448), indicating precise core vocabulary use. CAF (85.86) reflected adaptability to complex contexts, while low DDF (0.26) confirmed strong lexical norm adherence. Doubao slightly outperformed Simon in BFF (99.548), FMAF (99.632), and CAF (86.32), with an even lower DDF (0.208), demonstrating algorithmic precision and control.

In Task 2, Simon's scores showed greater variability with declines in BFF (99.192), COAF (98.872), FMAF (99.244), and CAF (79.6), and a higher DDF (0.336), reflecting increased errors in complex contexts. Doubao maintained stable, superior performance across accuracy metrics, with CAF at 86.324 and DDF at 0.224.

Overall, Doubao exhibits consistent advantages in lexical precision and error control, while maintaining comparable collocation and context adaptation. Simon's more variable performance under complex tasks suggests potential complementarity: AI ensuring accuracy, humans contributing creativity and nuanced language.

4.2 Comparison of Syntactic complexity between Simon-written compositions and Doubao-generated compositions Compositions

Figure 1 conducts correspondence analysis on syntactic complexity between human - written and Doubao - generated compositions. In the two - dimensional space structured by Dimension 1 (with higher variance interpretation) and Dimension 2, the closer the points (representing composition types or syntactic features) are to one another, the stronger the association between their corresponding categories; conversely, the farther they are, the weaker the assocation.

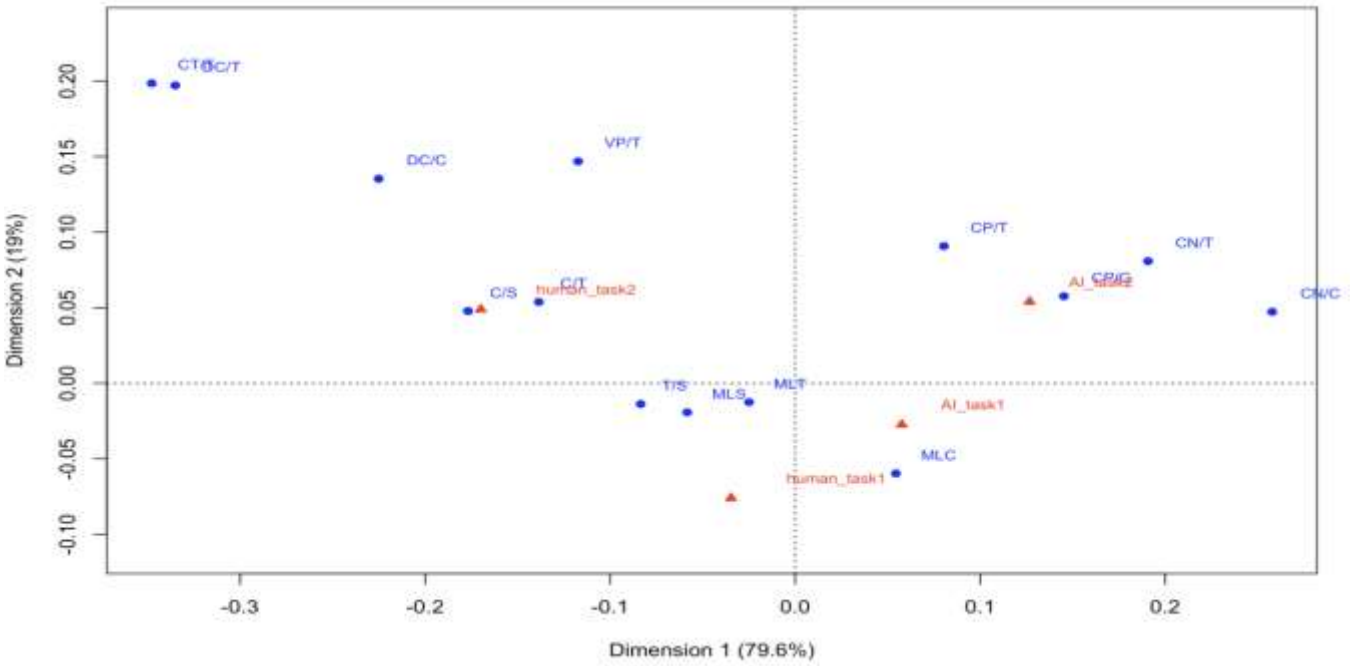

Figure 1. Correspondence Analysis of Syntactic complexity between Human-written Compositions and Doubao-generated Compositions

In Writing Task 1, Doubao-generated texts exhibit a "scale-first" syntactic pattern, with significantly higher surface-level metrics such as Mean Length of Sentence (MLS), Mean Length of T-unit (MLT), Coordinating Phrases per Clause (CPC), and Complex Noun Phrase ratio (CN/C) compared to human texts. In contrast, human texts outperform Doubao in deeper syntactic metrics

like Clause Density per T-unit (CT/T) and Dependency Complexity per Clause (DCC). This reflects Doubao's reliance on surface expansion versus humans' deeper structural optimization.

In Task 2, these differences are amplified due to greater semantic and logical demands. Human texts demonstrate enhanced deep-structure configurations with hierarchical clause nesting, clustering on the "deep complexity" side, while Doubao maintains a scale-dominant but more redundant syntactic style, showing greater divergence under increased task constraints.

Overall, human syntactic complexity better meets IELTS criteria by balancing sentence variety, precision, and logic with low redundancy and context-adaptive subordination. Doubao produces longer, more phrase-rich sentences but depends heavily on coordination and displays high redundancy, impacting coherence and clarity.
Table 6 illustrates the distribution of various textual cohesion devices (including Reference, Substitutions and Ellipsis, Conjunction, Repetition, Synonymy, Hyponymy, Meronymy, and Collocation) between Simon - written compositions and Doubao - generated compositions across Task 1 and Task 2.

Table 6. Textual Cohesion between Simon-written and Doubao-generated Compositions

Cohesion Devices	Simon-written compositions		Doubao-generated compositions	
	Task 1	Task 2	Task 1	Task 2
Reference	745	720	682	642
Substitutions and Ellipsis	0	8	0	2
Conjunction	362	415	318	338
Repetition	432	390	495	482
Synonymy	195	225	156	168
Hyponymy	128	140	102	108
Meronymy	39	45	31	32
Collocation	365	380	330	325

Based on a comparative analysis of human-written and AI-generated texts for IELTS Writing Task 1 and Task 2, a clear disparity in cohesive strategies and their effectiveness emerges. In descriptive writing (Task 1), human texts demonstrate superior coherence by employing a wider range of precise cohesive devices, including more frequent and accurate referencing (e.g., "the figure"), a greater variety of connectors to establish logical layering (e.g., "by contrast," "finally"), and more contextually appropriate collocations.

AI-generated texts, conversely, rely heavily on repetition of core terms, which, while reinforcing the topic, leads to redundancy and a lack of expressive fluidity. Similarly, in argumentative writing (Task 2), human texts achieve a more rigorous logical flow and persuasive power through sophisticated cohesive techniques such as precise referencing to maintain a clear argument, diverse connectors to introduce counter-arguments and advance propositions (e.g., "however," "furthermore"), and a richer use of synonymy and hyponymy. AI texts, again, primarily resort to high-frequency repetition of keywords, resulting in a flattened logical structure and a monotonous style.

The findings reveal a consistent pattern: human writers utilize a combination of explicit and implicit, context-adapted cohesive devices to build a tightly-knit, dynamic discourse, whereas AI-generated outputs, constrained by a rule-based methodology, exhibit a pronounced over-reliance on repetition, rigid connectors, and an overall deficiency in lexical variety and semantic depth, which ultimately compromises the fluency and rhetorical effectiveness of the text.

4.3 Comparison of Coherence between Simon-written compositions and Doubao-generated compositions Compositions

In IELTS Writing Task 1 (descriptive essays), human-written texts exhibited a clear advantage in coherence, achieved through a richer and more diverse use of cohesive devices. At the referential level, humans employed 745 instances of referencing, exceeding AI's 682 (+63), using precise terms such as "the figure" and "this process" to consistently anchor descriptions to chart data or procedural steps. In connector usage, humans used 362 instances versus AI's 318 (+44), employing devices such as "by contrast" for comparisons and "finally" for chronological sequencing, thereby structuring information hierarchically. Synonym/paraphrase use (195 vs. 156, +39) avoided redundancy through lexical variation (e.g., "rise," "increase," "jump"), while

collocations relevant to descriptive contexts (365 vs. 330, +35) such as "consumer spending" and "water cycle" reinforced semantic networks.

AI-generated texts outperformed humans only in repetition (495 vs. 432, +63), frequently reiterating core terms like "percentage" and "stage." While this reinforced topical focus, excessive repetition reduced expressive variety and introduced redundancy. Across other cohesive dimensions, AI lagged: referencing lacked precision, connectors were limited in range, and synonym/paraphrase usage was insufficient, resulting in rigidity and reduced integration of complex information.

In IELTS Writing Task 2 (argumentative essays), human-written texts again demonstrated superior logical coherence. Referential usage was 720 vs. AI's 642 (+78), enabling precise tracking of argumentative elements (e.g., "they" clearly referring to supporters of a policy). Connector usage (415 vs. 338, +77) allowed humans to flexibly layer logic—using "however" for counterarguments, "furthermore" to extend reasoning, and "for example" for evidence—creating multi-tiered arguments. Synonym/paraphrase use (225 vs. 168, +57) enriched discourse with lexical shifts such as "argue," "claim," and "maintain." Clarification of hyponymy/hypernymy (140 vs. 108, +32), such as refining "transport" to "cars" and "trains," enhanced conceptual precision and argumentative depth.

AI argumentative texts relied heavily on repetition (482 vs. 390, +92) of core terms like "policy" and "problem." Connectors were formulaic and limited, constraining logical versatility. Semantic refinement and conceptual layering were underdeveloped, producing single-layer arguments that lacked depth and reader engagement.

Overall, human writers displayed superior coherence-building in both descriptive and argumentative tasks through context-sensitive explicit cohesion (precise referencing, varied connectors) and implicit cohesion (semantic layering, natural collocations, lexical variety). AI outputs, while structurally stable, over-relied on repetition, employed connectors rigidly, and failed to develop semantic depth or multi-layered logic. These limitations resulted in noticeable deficits in fluency, logical rigor, and richness of content, underscoring the need for continued optimization of AI's capacity to simulate human-like coherence.

4.4 Comparison of Task response between Simon-written compositions and Doubao-generated compositions Compositions

Table 7 compares human-written and AI-generated IELTS Writing Task 1 (descriptive) and Task 2 (argumentative) essays on three core dimensions. The dataset includes 50 texts (25 human, 25 AI per task).

Table 7. Task response between Simon-written and Doubao-generated Compositions

	Simon-written compositions			Doubao-generated compositions		
Category	Task1			Task1		
Dimensions	SSI	ROA	SPM	SSI	ROA	SPM
Average	100%	98%	96%	100%	100%	100%
File Number （25）	Simon-written compositions			Doubao-generated compositions		
Category	Task 2			Task 2		
Dimensions	SSI	ROA	SPM	SSI	ROA	SPM
Average	100%	97%	96%	100%	100%	100%

Note：SSI（Surface Structure Integrity）、RQA（Reasoning Quality Assessment）、SRM（Structure-Reasoning Matching）

This quantitative comparison examines the performance of human-written and AI-generated IELTS essays across three core task response dimensions—Surface Structure Integrity (SSI), Reasoning Quality Assessment (RQA), and Structure–Reasoning Matching (SRM)—in both Task 1 (descriptive reports of charts) and Task 2 (argumentative essays), using a corpus of 25 texts per group for each task type.

The evaluation, based on Stapleton and Wu's (2015) argument quality framework and official IELTS scoring criteria, shows that in the SSI dimension, both human and AI texts achieved 100% in both tasks, indicating complete coverage of all required components. For Task 1, this included identifying the chart type, presenting key data, and summarizing overall trends; for Task 2, it encompassed stating a claim, providing supporting arguments, addressing counter-claims, and offering refutations.

In the RQA dimension, AI-generated texts attained perfect scores (100%) across both tasks, reflecting consistently relevant, accurate, and non-redundant reasoning. Human-written texts scored marginally lower—98% for Task 1 and 97% for Task 2—due

to occasional vague phrasing or underdeveloped arguments. Similarly, in the SRM dimension, AI texts again scored 100% for both tasks, evidencing precise logical correspondence between structural elements—such as the alignment of data and trends in Task 1 or the consistency between counter-claims and refutations in Task 2—and strong overall coherence. Human-written texts scored 96% in both tasks, with the slight gap arising from minor logical inconsistencies or refutations that only partially addressed the counter-claim in a few cases.

5. Conclusion

This study conducted a multi-dimensional analysis of the linguistic features of high-scoring human-written and AI-generated IELTS essays, revealing distinct performance differences. AI-generated texts demonstrated notable advantages in surface-level features—lexical resources, syntactic complexity, and task response—producing vocabulary-rich, grammatically accurate, and task-aligned writing. These strengths underscore the potential value of AI tools in supporting grammar and vocabulary acquisition.

However, clear shortcomings emerged in deeper linguistic features and pragmatic adaptability. AI essays tended to display high repetition, rigid transitions, and weak implicit cohesion, struggling to employ varied cohesive devices to establish rigorous logical hierarchies and deep semantic links. This reflects AI's incomplete mastery of contextual adaptation, rhetorical strategy, and an individualized authorial voice. In contrast, high-scoring human essays excelled in creative and critical thinking, emotional resonance, and natural fluency—qualities that remain beyond AI's reach. At the syntactic level, AI's strength lies in accuracy and structural stability, whereas human writers demonstrate greater flexibility, adjusting structures dynamically to suit context and communicative intent, resulting in more natural and expressive prose. This contrast underscores the distinction between AI's "technical correctness" and human "contextual adaptability."

Overall, the findings highlight a complementary relationship rather than a replacement dynamic between AI and human writing. AI can function as an efficient language generator, offering standardized models to strengthen foundational skills, but its discourse-level limitations necessitate its role as an "intelligent assistant" rather than a substitute. Learners may use AI outputs as templates for refining syntax and vocabulary while focusing on cultivating higher-level writing abilities. Pedagogically, teachers should serve as mediators, leveraging AI-generated texts alongside high-scoring human essays for comparative analysis. Tasks designed to prompt refutation, supplementation, or reconstruction of AI arguments, or to produce texts with strong personal voice, can both address AI's weaknesses and enhance teaching effectiveness.

Despite its contributions, this study has limitations. The corpus, drawn solely from a single Simon's dataset, restricts generalizability; future research should expand the dataset and include second-language learner texts. The analysis focused on static textual features, excluding prompt effects and dynamic writing processes; subsequent studies could integrate eye-tracking, think-aloud protocols, or interviews to examine cognitive processes in human and AI writing. Finally, the evaluation relied on linguistic feature analysis aligned with IELTS criteria but lacked third-party examiner scoring; future work should incorporate examiner assessment and explore fine-tuning language models to better align with the developmental trajectory of human writing, thereby yielding more robust and pedagogically relevant conclusions.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Derner, E., & Batistič, K. (2023). Beyond the Safeguards: Exploring the Security Risks of ChatGPT. arXiv. https://doi.org/10.48550/arXiv.2305.08005
[2] Liu, Z., & Dou, J. (2023). Lexical Density, Lexical Diversity, and Lexical Sophistication in Simultaneously Interpreted Texts: a Cognitive Perspective. *Frontiers in psychology, 14*, 1276705. https://doi.org/10.3389/fpsyg.2023.1276705
[3] Lu, X. (2010). Automatic Analysis of Syntactic Complexity in Second Language Writing. In*ternational journal of corpus linguistics, 15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu
[4] 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. https://doi.org/10.5054/tq.2011.240859
[5] McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381−392. https://doi.org/10.3758/BRM.42.2.381
[6] McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication, 27*(1), 57−86. https://doi.org/10.1177/0741088309356674
[7] Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

[8] Stapleton, P., & Wu, Y. A. (2015). Assessing the Quality of Arguments in Students' Persuasive Writing: A Case Study Analyzing the Relationship between Surface Structure and Substance. *Journal of English for Academic Purposes, 17*, 12–23. https://doi.org/10.1016/j.jeap.2014.10.002

[9] Stapleton, P., & Wu, W. (2015). Assessing Task Response in IELTS Writing: A New Model for Evaluation. *Applied Linguistics, 36*(5), 589 - 610.

[10] Wei, P., Wang, X., & Dong, H. (2023). The impact of Automated Writing Evaluation on Second Language Writing Skills of Chinese EFL Learners: A Randomized Controlled Trial. *Frontiers in Psychology, 14*, 1249991. https://doi.org/10.3389/fpsyg.2023.124999

[11] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and Social Risks of Harm From Language Models. arXiv. Retrieved from https://doi.org/10.48550/arXiv.2112.04359

[12] Zare, J., Al-Issa, A., & Madiseh, F. R. (2025). Interacting with ChatGPT in Essay Writing: A Study of L2 Learners' Task Motivation. *ReCALL, 37*(3), 385–402. https://doi.org/10.1017/S0958344025000035

[13] Zindela, N. (2024). Comparing measures of syntactic and lexical complexity in artificial intelligence and L2 human-generated argumentative essays. *International Journal of Education and Development using Information an Communication Technology, 19*(3), 50–68