Journal of Pragmatics and Discourse Analysis
ISSN: 2755-3426
DOI: 10.32996/jpds
Journal Homepage: www.al-kindipublisher.com/index.php/jpds

JPDS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Cohesion and Coherence in AI-generated Narrative Texts: ChatGPT vs Grok

Youssef Hilmi[1] and Kenza Saadani Hassani[2]
[1]*PhD Student, English department, Sidi Mohamed Ben Abdellah University, Fes Morocco*
[2]*PhD, English department, Sidi Mohamed Ben Abdellah University, Fes Morocco*
Corresponding Author: Youssef HILMI, **E-mail**: Youssef.hilmi@usmba.ac.ma

| ABSTRACT
The rapid advancement of large language models has raised questions about their ability to produce cohesive and coherent texts. Our exploratory study examines and compares referential and temporal cohesion in 20 narrative texts, approximately 500 words each, generated by ChatGPT (OpenAI) and Grok (x.AI). Coh-Metrix 3.0 was used as an automatic tool for text analysis. 8 Coh-Metrix referential cohesion indices (Local and global noun, argument, stem and content word overlap) and two temporal cohesion indices (incidence of temporal connectives and semantic temporal overlap /tense aspect consistency) relevant for the study were selected. Given the small sample size, a series of Mann Whitney U tests along with Benjamini-Hochberg false discovery rate (FDR) correction were applied. The results revealed large similarity between local and global referential cohesion across models, with a small to moderate advantages for Grok in global noun and stem overlap (all FDR-adjusted p>.05). In contract, temporal cohesion showed a clear divergence between the two models: ChatGPT exhibited considerably higher use of explicit temporal connectives (CNCTemp, rank bi-serial r=0.73, large effect; FDR-adjusted p=.010), whereas Grok demonstrated stronger implicit temporal consistency (SMTEMP, r=.044, moderate-to-large effect). The latter difference, however, did not remain statistically significant after FDR correction (FDR-adjusted p=.245). Overall, these findings suggest that the two models espouse two distinct strategies to achieve text coherence in narrative texts with explicit cue reliance from ChatGPT versus deeper situational model coherence from Grok.

| KEYWORDS

Cohesion, coherence, ChatGPT, Grok, Coh-Metrix, AI-generated texts

| ARTICLE DOI:

ACCEPTED: 01 April 2026 **PUBLISHED:** 03 May 2026 **DOI:** 10.32996/jpda.2026.5.2.3

1. Introduction

Artificial intelligence, henceforth AI, may be considered as the most technological breakthrough since the invention of the internet. Its transformative impact spans almost every field of knowledge and professional practice. The public release of ChatGPT-3.5 by OpenAI in late 2022, with its unprecedented conversational capabilities, re-kindled public attention and debate about the potential of advanced generative AI systems.

A striking feature of generative AI tools such as ChatGPT and Grok that stands out is their highly sophisticated conversational and linguistic capabilities. In addition to their excellence in tasks like solving complex mathematical problems or performing well on professional exams, these systems can produce very well-crafted articles, essays and stories. This proficiency has ignited intense scholarly debate over the extent to which ChatGPT and other AI systems can genuinely generate human-like text. For some researchers, such output is only a sophisticated statistical mimicry of human language without any depth and genuine understanding (Chomsky et al, 2023). In contrast, others argue that large language models not only simulate but even rival or surpass human performance in certain linguistic tasks, which according to this view, would challenge innateness theories and position statistical models as better alternatives for explaining language competence (Piantadosi, 2024). Some critical views of AI systems have gone even further to express their deep concern that AI's mastery of language is a major threat to core human systems and that these systems may be in the position of "hacking the operating system of civilization" (Harari, 2023).

This ongoing debate provides the backdrop for the present study, which examines the extent to which contemporary generative AI systems produce cohesive and coherent text. The study addresses two main questions:

(1) How is referential cohesion distributed across ChatGPT and Grok's generated narrative texts? Are they similar or different?

(2) How is temporal cohesion distributed across ChatGPT's and Grok's generated narrative texts? Are they similar or different?

To answer these questions, these hypotheses are put forward:

RQ1 H0 hypothesis: Referential cohesion is equally or similarly distributed across narrative texts generated by ChatGPT and Grok.

RQ1 H1 referential cohesion is NOT equally or similarly distributed across narrative texts generated by ChatGPT and Grok.

RQ2 H0 temporal cohesion is equally and similarly distributed across narrative texts generated by ChatGPT and Grok.

RQ2 H1 temporal cohesion is NOT equally or similarly distributed across narrative texts generated by ChatGPT and Grok.

The purpose of this study is to examine how referential and temporal cohesion are distributed in narrative texts generated by two prominent generative AI models, ChatGPT and Grok and whether these two tools deploy similar or different cohesive mechanisms to achieve text coherence. Through the analysis of these features, which are crucial for discourse quality and reader comprehension (Halliday & Hasan, 1976; Graesser et al., 2004), the study aims at exploring the linguistic capabilities of two major AI tools as featured in their ability to produce coherent narratives.

The study is of interests to several audiences. Computer scientists who work in the development of AI tools can gain insight into how these models incorporate and use linguistic elements that enhance text quality. Educators using these tools in their classrooms will better understand which model suits their best interest. Additionally, researchers who are interested in exploring AI's linguistic capabilities will benefit from the empirical comparison of cohesion and coherence in AI's generated text.

2. Literature Review

Cohesion has been a central concept in linguistics and discourse analysis since the publication of Halliday and Hasan's (1976) seminal work Cohesion in English. In this book, they define cohesion as a semantic relation that binds a text together and hence turns a string of sentences into a unified and meaningful unit. They argue that cohesion is realized through lexico-grammatical resources. Lexically, it is realized through vocabulary via two mechanisms: reiteration which includes repetition, synonymy, near synonymy and superordination; and collocation which refers to the tendency of certain words to co-occur in a predictable manner. Cohesion is also realized grammatically through reference, substitution, ellipsis and conjunction. Although cohesion contributes significantly to the textuality of the text, it is still insufficient on its own to achieve coherence which involves other elements such as world knowledge (Halliday and Hasan, 1976, pp. 4, 7).

Both text cohesion and coherence are major factors that influence writing quality (Witte & Faigley, 1981, Zhang et al. 2024) and text comprehension (McNamara et. al. 2014, Zwaan, 1996, Zwaan 2025, Zwaan & Radvansky, 1998, Kamalski et al., 2008). While Halliday and Hasan (1976) have underscored the role of text-based cohesion in promoting text coherence, cognitive approaches attribute text coherence primarily to the reader's mental representation rather than the inherent textual properties (Karoly, 2017; Martin, 2001; MAO, 2021; Navratilova et al., 2017; Van Dijk & Kintsch, 1983; Widdowson, 2004; Zwaan & Radvansky, 1998; Zwaan, 1996). An important contribution in this regard is Van Dijk and Kintsch's (1983) strategy-oriented model of discourse processing, which claims that comprehension is contingent on the construction of a propositional textbase and a situation model in episodic memory:

The cognitive representation of the events, actions, persons, and in general the situation, a text is about…. A situation model may incorporate previous experiences, and hence also previous textbases, regarding the same or similar situations. At the same time, the model may incorporate instantiations from more general knowledge from semantic memory about such situations (Van Dijk & Kintsch, 1983, pp. 11-12).

According to this framework, both local and global coherence conditions must be satisfied during textbase construction. However, full or better comprehension requires consistent and constant updating of the situation model through interaction with the reader's goals and background knowledge.

The construction of this model relies on several interrelated strategies. Propositional strategies involve deriving propositions from clauses. Each clause expresses one proposition, which is also composed of other smaller ones called atomic propositions that correspond to word meanings. Operating primarily in short term memory, local coherence strategies establish links between propositions and the world facts these propositions express. Examples of these strategies include coreferential ties for shared entities (persons or objects), explicit connectives, clause ordering, and activations from long-term memory (Van Dijk & Kintsch, 1983).

Moreover, macrostrategies enable the derivation and inference of higher-level macropropositions from the textbase. These macropropositions are sequenced in their turn to form the macrostructure which is basically the overall gist or topic of the text. Finally, schematic strategies draw on conventional organizational patterns, superstructure, to structure the text globally.

Along similar lines, the event-indexing model (Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998) posits that situation-model construction involves monitoring and updating events along five dimensions: time, space, protagonist, causality

and intentionality. As readers process the text, they extract events and integrate them into the evolving situation model by assessing continuity on these dimensions. Coherence is facilitated when overlap exist between events, but when shifts occur, update of the model is required.

One important type of textual cohesion that supports this integration is referential cohesion. According to Halliday and Hasan (1976), a basic feature of reference is that "some elements in the text instead of being interpreted semantically in their own right, they make reference to something else for their interpretation" (p.31). In other words, the interpretation of an item, e.g., a pronoun or a noun phrase, depends on another element elsewhere in the text, which creates a referential cohesive tie. For instance, a noun, a pronoun or a noun phrase argument may refer to a prior constituent in the text. However, when content words lack connections to surrounding elements, referential cohesion gaps arise. These gaps can disrupt comprehension by pushing readers to bridge discontinuities (McNamara et al., 2014, p. 50)

Another prominent type of cohesion in narrative texts is temporal cohesion. As McNamara et al. (2014) note, "Temporality in text is important because of its ubiquitous presence in organizing language and discourse. Time is represented through inflected tense morphemes (e.g., "-ed," "is," "has") in every sentence of the English language" (p. 56). Temporal cohesion is featured through tense and aspect. Verb tense is rapidly accessed during processing, which influence sentence and discourse comprehension (Trueswell & Tanenhaus, 1991).

Temporal cohesion is especially crucial in narratives because it enables readers to sequence events chronologically. Readers typically assume iconicity. This means they expect the described order to match the actual sequence of events. However, when this expectation is violated, textual cues are required to signal shifts and locate events on the timeline. (Zwaan, Madden, & Stanfield, 2001). Such cues serve multiple tasks such as reducing ambiguity, activation of relevant concepts, facilitating the integration of incoming information, and serving as processing signals. (Zwaan & Rapp, 2006; Zwaan, 1996). The cues also establish cohesive links between sentences and propositions and contribute to the organization of the text (McNamara et al., 2014, Van Dijk & Kintsch, 1983). However, they impose additional cognitive load and can negatively affect memory and text recall, particularly temporal connectives (Millis, Graesser, & Haberlandt, 1993; McNamara et al., 2014; Zwaan, 1996). These connectives are particularly abundant in narrative texts (McNamara et al., 2014).

The overarching theoretical question of whether conversational AI Chatbots can generate human-like texts has prompted researchers to examine the linguistic properties of these outputs and compare them to human-written ones. ChatGPT is still the most extensively studied generative AI model to date. For instance, Herbold et al., (2023) demonstrated in a large-scale comparison that ChatGPT-generated argumentative essays were rated higher in overall quality, logical structure and coherence than those written by high school students.

Distinctive stylistic features further differentiate AI from human writing. ChatGPT tends to use more nominalizations and lexical diversity while using fewer discourse markers (connectives); it also relies on paragraphing and syntactic complexity to maintain logical flow (Herbold et al., 2023; Georgiou, 2025).

ChatGPT also demonstrates strong narrative capabilities. It outperformed Chinese intermediate English learners in narrativity, word correctness and referential cohesion. However, it lagged behind in syntactic simplicity and deep cohesion (Zhou et al., 2023).

Furthermore, ChatGPT exhibits human-like patterns in several aspects of language use such as discourse level inferences, semantic priming which means updating word meanings based on recent encounters, and syntactic structure reuse (Cai et al., 2023). Nevertheless, it falls short in two key experiments: unlike humans, who tend to use shorter words in more predictable contexts, ChatGPT does not follow this principle of least effort. Additionally, it fails to resolve syntactic ambiguities using contextual information (Cai et al., 2023).

Emara (2025) conducted a mixed method stylometric study that compared nonnative ESL students' adaptations of classic Arabic short stories with ChatGPT-4 adaptations of the same texts. The findings revealed significant differences: ChatGPT-generated versions exhibited greater fidelity to the original theme, descriptive, unique and bias-free language. In contrast, student adaptations often deviated from the source texts and featured simpler, more repetitive structures, longer sentences with excessive coordinators, basic vocabulary, frequent intensifiers, and L1-induced features. Remarkably enough, students used more cohesive markers than ChatGPT.

In a complementary study, Revell et al., (2024) compared first-year undergraduate analyses of Old English poetry with those generated by ChatGPT-4o. The results indicated that AI-produced commentary essays were rated as effective and of comparable quality to student work, without any statistically significant difference in scores, with formal structure and tone. However, AI texts were more descriptive than analytical and lacked genuine critique of poetic style and effect compared to human productions.

Akinwande et al., (2024) conducted a large-scale comparative analysis of 500k human-written and AI-generated essays. Their findings revealed that human-authored essays have a higher average of word count and greater vocabulary diversity. On the other hand, AI-generated essays featured longer average word length. Similarly, Guo et al., (2023) examined responses to 40,000 questions across multiple domains including open-domain knowledge, finance, medicine, law, and psychology. They found that human experts delved into hidden or deeper meanings and employed more diverse vocabulary. In contrast, ChatGPT relied more on literal interpretations of questions and used more conjunctions to convey logical flow and structural clarity (Guo et al., 2023;

Liao et al., 2023). These patterns suggest that while AI systems excel at surface-level fluency and explicit connections, they still struggle with interpretive depth.

Early work on language models, such as that by Gehrmann et al., (2019) observed that humans varied expressions by using synonyms or pronouns for coreference. Language models, however, often repeated entity names verbatim instead of pronominalization.

In spite of ChatGPT's remarkable performance in many linguistic tasks, it remains limited at the pragmatic level. For instance, Qui et al., (2023) demonstrated that ChatGPT struggled to differentiate literal meaning from pragmatic inference. This reflects a deficiency in the model's ability to switch between semantic and pragmatic interpretations. Similarly, ChatGPT faces challenges with figurative language. Sahariet al., (2025) found that while ChatGPT-3.5 could handle simpler English-Arabic translation tasks adequately, it grappled with conveying the nuances of metaphors, idioms and other figurative elements.

Large language model's ability to generate cohesive and coherent texts remains central to their overall success as human-like language producers. Ismail (2023) examined cohesion and coherence in ChatGPT-generated essays versus those written by Egyptian university students. The results showed that ChatGPT outperformed students in global coherence, situation model and connectives and lexical choices.

These positive findings for ChatGPT contrast with Ripoll Y Schmitz and Sonnleitner (2025), who found that human-written reading comprehension passages were rated significantly more coherent than those generated by AI ($p=.023$), particularly in informative genres. The contradictory results may stem from methodological differences such as Ismail's use of non-native English student texts versus the native-level in Ripoll Y Schmitz and Sonnleitner.

In conclusion, the reviewed literature reveals two primary trends. First, research has predominantly compared human-written texts with those generated by AI. This is comprehensible as researchers are interested in examining whether AI can produce authentically human-like texts. Second, ChatGPT has been the central focus of these investigations. This is probably due to the fact that ChatGPT was the first-large-scale generative model released to the public and it instantly attracted a widespread academic scrutiny. This focus has largely overlooked direct comparisons between different AI systems which could offer insight into similarities or difference within AI systems. Therefore, the present study addresses this gap by comparing the distribution of cohesion mechanisms in narrative texts generated by ChatGPT and Grok. This exploratory study offers insights into model-specific differences that previous human-AI comparisons have not explored.

3. Methodology

The corpus

The data for this study consists of a corpus of 20 narrative texts, 10 generated by ChatGPT from OpenAI and 10 by Grok from x.AI. These two specific AI tools were selected for the following reasons. First, ChatGPT was the first generative AI tool to achieve widespread public adoption following its launch in November 2022, and it remains the most extensively used model, with an estimated 700-900 million weekly active users as of early 2026 (OpenAI, 2025; Altman,2025).

ChatGPT is a large language model developed by OpenAI that employs transformer-based neural network architectures to understand and generate human-like text. It functions as a conversational agent capable of answering questions, producing coherent narratives and supporting language processing tasks across diverse domains.

On the other hand, Grok represents rising contender in the AI landscape. It was released in November 2023, one year after ChatGPT. While Grok's user base remains significantly smaller than ChatGPT's, with 30-35 million monthly active users as of early 2026 (Exploding Topics, 2025), its popularity is growing rapidly, particularly among X users due to its integration with the platform. Grok is also a conversational large language model developed by xAI and built on advanced transformer architecture (xAI, n.d.). This comparison between ChatGPT and Grok provides a meaningful insight into how different models handle two types of cohesion in narrative texts.

To ensure comparability, the same 10 prompts were used for both models. Each prompt required 500-word narrative story on a specified topic and was delivered in zero-shot mode, which means that no examples or additional guidance were provided. Texts were generated independently in separate sessions to avoid cross-context influence. All texts were generated in October 2025 using free versions available at that time ChatGPT 5 and Grok 4.

The prompts were as follow:

S1: "Write a story of 500 words about a Moroccan young man named Youssef"
S2:"Write a story of 500 words about a Moroccan old man"
S3: "Write a story of 500 words about a young explorer discovering a hidden ancient city in the jungle."
S4: "Write a story of 500 words about a chef who invents a magical recipe that changes people's emotions."
S5:"Write a story of 500 words about a musician reuniting with a long-lost friend during a storm."
S6:"Write a story of 500 words about a child who befriends a stray robot in a futuristic city."
S7:"Write a story of 500 words about a librarian uncovering a cursed book that alters reality."
S8:"Write a story of 500 words about a sailor surviving a shipwreck on a mysterious island."
S9:"Write a story of 500 words about an artist whose paintings come to life at midnight."
S10:"Write a story of 500 words about a detective solving a puzzle involving forgotten family secrets."

After generation, the 20 texts were pasted into Microsoft Word document for initial data cleaning. Non-linguistic marks such as formatting artifacts, special characters and symbols were removed to ensure compatibility with Coh-Metrix 3.0 processing. Text cleaning is essential for accurate index computation and processing (McNamara et al., 2014). Punctuation marks were kept intact. The cleaned texts were then converted to plain text (.txt) format as a final step before processing.

Coh-Metrix 3.0 indices measure multiple discourse constructs. Given the study's focus on referential and temporal cohesion, only relevant indices were selected. The chosen indices were then exported and statistically analyzed using IBM SPSS statistics 26. Descriptive statistics, such as median and interquartile ranges, and inferential tests namely Mann-Whitney U test were used to compare performance between ChatGPT and Grok. Since a series of Mann Whitney U tests were conducted, a Benjamini-Hochberg correction was applied to control false positives (FDR).

The tool

Coh-Metrix is a computational tool which was developed by researchers at the Institute for Intelligent Systems at The University of Memphis. Drawing on advances in computational linguistics and discourse processing, the developers created a suite of sophisticated indices and compiled them into a single automated system called Coh-Metrix (Graesser et al., 2004).

Coh-Metrix analyzes texts for a wide array of linguistic and discourse features. Its primary goal is the measurement of text cohesion. However, it also includes indices that assess readability, syntactic complexity, lexical diversity, and aspects of world knowledge. The tool employs established computational methods including syntactic parsers (Charniak, 2000), and Latent Semantic Analysis (LSA; Landauer et al., 2007) to compute these measures. In addition, Coh-Metrix provides basic textual descriptions including the average of word length and sentence length and word count.

Coh-Metrix has proven to be a powerful and reliable tool for automated textual analysis. Numerous studies have validated its efficacy in measuring intended constructs, including cohesion and Latent Semantic Analysis (LSA) indices (McNamara et al., 2010), lexical diversity indices (McCarthy & Jarvis, 2010), and the L2 lexical growth indices (Crossley, Salsbury, & McNamara, 2009). Coh-Metrix has also been successfully applied to detect authorship and stylistic differences (McCarthy, Lewis et al., 2006), measure structural cohesion (McCarthy et al., 2007), assess temporal cohesion across genres (Duran et al., 2006) distinguish high from low cohesion texts (McNamara, Ozuru, Graesser, & Louwerse, 2006), and classify textual genre (McCarthy et al., 2006).

Coh-Metrix 3.0 provides approximately 100 indices measuring various constructs related to language, discourse, cohesion and readability. Given the study's focus on referential and temporal cohesion in AI-generated narrative texts, only relevant indices from Coh-Metrix were selected and extracted for analysis. These indices are as follows:

Referential cohesion:

CRFNO1: Adjacent sentence noun overlap

CRFAO1: Adjacent sentence argument overlap

CRFSO1: Adjacent sentence stem overlap

CRFNOa: Global (all-sentence) noun overlap

CRFAOa: Global argument overlap

CRFSOa: Global stem overlap

CRFCWO1: Adjacent sentence content word overlap

CRFCWOa: Global content word overlap

Temporal cohesion:

CNCTemp: Incidence of temporal connectives (e.g., before, after, then)

SMTEMP: Semantic temporal overlap (indicating deeper temporal coherence across the text)

Referential cohesion is assessed through eight indices in Coh-Metrix 3.0 which measures four types of coreference at two levels: local (adjacent sentences) and global (across all sentences in the text). These four types are noun overlap, argument overlap, stem overlap and content overlap. The first type measures the proportion of sentences sharing identical nouns with no morphological variations (e.g., "table"/ "table"). Argument overlap captures overlap between head nouns with no morphological restrictions, (e.g., "table"/ "tables") and includes pronouns (e.g., "they"/ "they"). It is worth to note that Coh-Metrix does not resolve the anaphoric relationship between head nouns and their referents (McNamara et al, 2014). Stem overlap is more flexible. It detects shared lemmas (core morphological units) between a noun in one sentence and any content word (noun, adjective, verb or adverb) in another sentence regardless of morphological differences. Last but not least, content word overlap measures shared content words across sentences.

Coh-Metrix also measures temporal cohesion through two primary indices that capture both explicit and implicit temporal relations in text. The first index tracks tense and aspect repetition across the text. Higher numbers of tense shifts result in lower repetition scores, which indicates less consistent temporal continuity (McNamara et al., 2014). The second index provides an incidence score of occurrence per 1,000 words for temporal connectives which serve as explicit cues to signal time relationships. These measures allow Coh-Metrix to quantify both surface-level temporal markers and deeper semantic temporal coherence.

4. Results and Discussion

The primary objective of the analysis was to determine whether the distribution of referential and temporal cohesion differed between the narrative texts generated by ChatGPT and those generated by Grok. Given the small sample size (n=10 per model), a non-parametric approach was adopted. A series of Mann-Whitney U test were used to compare the two groups, as it is

appropriate for ordinal or non-normally distributed data and does not assume equal variance. A Benjamini-Hochberg correction (FDR) was applied to control false positives.

Referential cohesion

Descriptive statistics, medians and interquartile ranges, were first calculated for referential cohesion at the local (adjacent sentences) and global (all sentences) levels separately. Table 1 presents the results for local referential cohesion indices.

Local referential cohesion was assessed using four Coh-Metrix indices: Noun overlap (CRFNO1), argument overlap (CRFAO1), stem overlap (CRFSO1) and content word overlap (CRFCWO1). These indices are calculated at the level of adjacent sentences. As shown in table 1, the results revealed minimal difference between ChatGPT and Grok across these measures. For noun overlap, Grok exhibited slightly higher scores (Md=0.065, IQR=0.09) than ChatGPT (Md=0.048, IQR=0.08). Similarly, Grok and ChatGPT showed similar tendency regarding argument overlap where Grok (Md=0.250, IQR=0.16) scores slightly higher than ChatGPT (Md=0.229, IQR=0.25). Stem overlap also favors Grok (Md=0.094, IQR=0.12) over ChatGPT (Md=0.060, IQR= 0.07). However, content word overlap was nearly identical between the two models with Grok scoring (Md=0.033, IQR=0.02) and ChatGPT (Md=0395, IQR=.04). The Mann-Whitney U test showed no statistically significant difference after Benjamini-Hochberg FDR correction (all FDR-adjusted p>.05). Rank biserial effect sizes were small (r=0.00-0.10). Overall, these findings suggest that ChatGPT and Grok's narrative texts are highly comparable in terms of local referential cohesion.

Table 1

Local Referential Cohesion in AI-Generated Narratives: ChatGPT vs. Grok (n = 10 per model)

Metric	Description	ChatGPT Md (IQR)	Grok Md (IQR)	U	p	FDR-adjusted p	r
CRFNO1	Noun overlap, adjacent sentences	.048 (.08)	.065 (.09)	45.000	.704	1.000	.08
CRFAO1	Argument overlap, adjacent sentences	.2290(.25)	.2500 (.16)	44.000	.650	1.000	.10
CRFSO1	Stem overlap, adjacent sentences	.0605 (.07)	.0940 (.12)	44.500	.677	1.000	.09
CRFCWO1	Content word overlap, adjacent sentences	.0395 (.04)	.0335 (.02)	50.000	1.000	1.000	0.00

Note. Md=median; IQR=interquartile range. Comparisons were performed using the Mann-Whitney U test. FDR-adjusted p-values account for multiple testing. r=rank-biserial correlation effect size.

Global referential cohesion was also examined using four Coh-Metrix indices that assess overlap across all sentences in the text: noun overlap (CRFNOa), argument overlap (CRFAOa), stem overlap (CRFSOa) and content word overlap (CRFCWOa). As shown in table 2, the results indicated largely similar levels of global referential cohesion between ChatGPT and Grok, with some minor differences.

Grok produced higher scores than ChatGPT on noun overlap (Md= 0.143, IQR=0.07 vs. Md= 0.111, IQR=0.05, r=0.41), which reflects a moderate effect size in favor of Grok. Argument overlap was nearly identical between the models (Grok: Md=0.300, IQR=0.13; ChatGPT: Md=0.307, IQR =0.12) with a negligible effect size (r=0.05). Stem overlap also favored Grok (Md=0.173, IQR=0.08) over ChatGPT (Md=0.146, IQR=0.07) with a small-to-moderate effect size (r=0.30). In contrast, content word overlap showed almost no difference (Grok: Md=0.044, IQR=0.01; ChatGPT: Md=0.044, IQR=0.01; r=0.04) with insignificant effect size difference. Mann Whitney U test revealed no statistically significance between the two models after the Benjamini-Hochberg FDR correction.

Globally, these findings suggest that global referential cohesion is also comparable to a large extent across the two AI models, though Grok exhibited noticeable advantages in noun overlap and stem overlap, which indicates a slightly stronger global referential ties in its narrative outputs.

Table 2
Global Referential Cohesion in AI-Generated Narratives: ChatGPT vs. Grok (n = 10 per model)

Metric	Description	ChatGPT Md (IQR)	Grok Md (IQR)	U	p	FDR-adjusted p	r
CRFNOa	Noun overlap, all sentences in text	0.112 (.05)	0.143 (.07)	26.00	.070	.233	0.41
CRFAOa	Argument overlap, all sentences in text	0.3070 (.12)	0.301 (.13)	47.00	.821	1.000	0.05
CRFSOa	Stem overlap, all sentences in text	0.146 (.07)	0.173 (.08)	32.00	.173	.433	0.30
CRFCWOa	Content-word overlap, all sentences in text	0.045(.01)	0.044 (.01)	47.50	.850	1.000	0.04

Note. Md=median; IQR=interquartile range. Comparisons were performed using the Mann-Whitney U test. FDR-adjusted p-values account for multiple testing. r=rank-biserial correlation effect size.

Temporal cohesion

Two Coh-Metrix indices were used to measure temporal cohesion: incidence of temporal connectives per 1,000 words (CNCTemp) and semantic temporal overlap/consistency of tense and aspect across the text (SMTEMP). As shown in Table 3, ChatGPT produced a substantially higher frequency of temporal connectives (Md=25.09, IQR= 9.08) than Grok (Md=18.15, IQR=10.42), with a large effect size (r=0.73). In contrast, Grok achieved higher scores on semantic temporal overlap/tense-aspect (Md=0.928, IQR=0.04) than ChatGPT (Md=0.890, IQR=0.09), with a moderate-to-large effect size (r=0.44). Mann Whitney U test showed that the difference in CNCTemp remained statistically significant after Benjamini-Hochberg FDR correction (FDR-adjusted p=.010) while the difference in SMTEMP did not (FDR-adjusted p=.245). These findings suggest distinct strategies for achieving temporal cohesion in narrative texts. ChatGPT depends more heavily on overt connectives for explicit temporal signaling, whereas Grok achieves temporal unity through greater tense/aspect consistency.

Table 3
Temporal Cohesion in AI-Generated Narratives: ChatGPT vs. Grok (n = 10 per model)

Metric	Description	ChatGPT Md (IQR)	Grok Md (IQR)	U	p	FDR-adjusted p	r
CNCTemp	Temporal connectives incidence (per 1,000 words)	25.10(9.08)	18.15(10.42)	7.00	.001	.010	0.73
SMTEMP	Temporal cohesion, tense and aspect repetition, mean	0.890(.09)	0.928 (.04)	24.00	.049	.245	0.44

Note. Md=median; IQR=interquartile range. Comparisons were performed using the Mann-Whitney U test. FDR-adjusted p-values account for multiple testing. r=rank-biserial correlation effect size.

5. Conclusion

In this study we examined two types of cohesion in narrative texts generated by Grok and ChatGPT: referential cohesion and temporal cohesion. We used a corpus of 20 narrative texts. Each model generated 10 texts based on the same prompts. The texts were analyzed using Coh-Metrix 3.0. The statistical analysis was carried out using IBM SPSS 26. The findings revealed a nuanced pattern. While referential cohesion was comparable in this sample across models and showed no significant difference, temporal cohesion, particularly CNCtemp index, showed clear divergence between the two models.

Although Grok's scores are relatively higher than ChatGPT, referential cohesion at both levels (local and global) did not differ significantly between Grok-generated texts and those of ChatGPT. This similarity suggests that both models have achieved a similar level of performance in maintaining entity continuity and anaphoric links in these texts. This implies that readers may experience comparable ease with tracking characters, objects or ideas regardless of which model generated the text.

In contrast, temporal cohesion emerged as a strong differentiating factor in this sample with each models adopting a different strategy. At the surface level, ChatGPT employed significantly more explicit temporal connectives which resulted in highly more

transparent chronological signaling through markers such as "then" "after" "next". This finding aligns with previous results from (Guo et al., 2023; Liao et al., 2023) who found that ChatGPT used more conjunctions to maintain logical flow. This strategy of overt use of discourse temporal markers may enhance readers comprehension especially unskilled readers or low-knowledge readers by helping them navigate the text through these cohesive cues and making time flow more predictably especially in narrative texts where sequential progression is crucial (Kamalski et al., 2008; McNamara et al., 2014).

At the deeper level, however, Grok, showed numerically higher semantic temporal overlap and tense-aspect consistency, (SMTEMP: Md = 0.928 vs. 0.890; Mann–Whitney U = 24.00, p = .049, r =0.44). However, this difference did not remain statistically significant after Benjamini-Hochberg FDR correction (FDR-adjusted p=.245). Nonetheless, the moderate-to large effect size points to a potential advantage for Grok in achieving temporal coherence through greater tense and aspect repetition with less disruptions such as tense shifts or flashback. In other words, Grok appears to achieve chronological coherence more implicitly by relying on tense and aspect uniformity instead of using explicit temporal connectives. This consistency in tense and aspect with fewer time shifts has an effect on narrative text comprehension as suggested by the event-indexing model (Zwaan, 2024). According to Zwaan (2024) construction of the situation model involves five dimensions including time, location, entity, motivation, and physical causation. The overlap on these five dimensions between events makes the processing of the text easier for readers. Following this line of reasoning, Grok texts are easier to process as there is more temporal overlap achieved through tense and aspect consistency along with to a less processing load imposed by explicit markers.

Overall, these findings suggest that Grok's generated texts exhibit stronger referential cohesion, though non-significant, and a higher semantic temporal cohesion than those generated by ChatGPT. These results align with previous research findings suggesting that more recent and advanced AI models achieve better coherence through deeper referential cohesion and semantic structures rather than explicit discourse markers (Herbold et al., 2023; Zhou et al., 2023). Herbold et al. (2023) reported that ChatGPT-4 generated essays employed less connectives while receiving higher coherence ratings. A pattern that converges with Grok's reduced usage of temporal connectives aligned with enhanced deeper cohesion. Similarly, Zhou et al. (2023) found that although ChatGPT outperformed human writers in referential cohesion, it was limited in deep cohesion. These results suggest that Grok may reflect a subsequent development in AI text generation characterized by more reliance on deep cohesion mechanisms that approximate features of high-quality human discourse (Liao et al., 2023; Revell et al., 2024; Ripoll Y Schmitz & Sonnleitner, 2025). This means that as AI technology advance, new models would be able to generate more coherent texts and therefore attain a human-like output.

This has implications for educators who use AI generated texts in their classrooms to teach narrative texts. ChatGPT texts which contain more temporal connectives and hence more cohesive cues would be a very good choice for unskilled or low-knowledge readers as it will save them the extra effort of trying to put events together on a timeline (Kamalski et al., 2008; McNamara et al., 2014). On the other hand, Grok's texts would be a better choice for skilled or high-knowledge readers for whom, as studies have suggested, cohesive cues may form an addition burden and who would be better off making inferences themselves (Kamalski et al., 2008; McNamara et al., 2014).

Numerous limitations of our study should be acknowledged. First, the size of the corpus sample was relatively small (n=10 texts per model). Although sufficient for an exploratory study, this limited statistical power and might have contributed to the non-significance observed in the referential cohesion indices. Future studies with larger corpora would provide greater power to detect subtle model differences. Second, all texts were generated using zero-shot prompts strategy with a text length constraint of 500 words. different prompting techniques (e.g., few-shot examples, persona adoption or explicit stylistic instructions) may yield different patterns of cohesion. Future research may explore different prompt engineering strategies such as persona adoption or explicit stylistic constraints. Third, texts were generated through standard web interface rather than the API. Consequently, parameters such as temperature, top-p and presence/frequency penalties were fixed at default values. Systematic manipulation of these parameters of these parameters via the API could produce different cohesion outcome.

Funding: This research received no external funding

Conflicts of Interest: The authors hereby declare that there is no competing interest to declare.

ORCID iD: 0009-0004-9839-7575

References

[1] Akinwande, M. J., Adeliyi, O., & Yussuph, T. T. (2024). Decoding AI and human authorship: Nuances revealed through NLP and statistical analysis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4895334

[2] Cai, T., Wang, X., Ma, T., Chen, X., & Zhou, D. (2024). Large language models as tool makers. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=qV83K9d5WB

[3] Charniak, E. (2000). A maximum-entropy-inspired parser. In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (pp. 132–139). Association for Computational Linguistics.

[4] Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). The false promise of ChatGPT. The New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

[5] Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. Language Learning, 59(2), 307–334. https://doi.org/10.1111/j.1467-9922.2009.00508.x
[6] Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2006). Using temporal indices to predict temporal segments in narrative and expository texts. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 1261–1266). Cognitive Science Society.
[7] Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. Behavior Research Methods, 39(2), 212–223. https://doi.org/10.3758/BF03193150
[8] Emara, I. F. (2025). A linguistic comparison between ChatGPT-generated and nonnative student-generated short story adaptations: A stylometric approach. Smart Learning Environments, 12(1), Article 36. https://doi.org/10.1186/s40561-025-00388-z
[9] Exploding Topics. (2025, September 30). Number of Grok users (Grok statistics 2025). https://explodingtopics.com/blog/grok-users
[10] Faigley, L., & Witte, S. P. (1981). Coherence, cohesion, and writing quality. College Composition and Communication, 32(2), 189–204. https://doi.org/10.2307/356693
[11] 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. https://doi.org/10.18653/v1/P19-3019
[12] Georgiou, G. P. (2025). Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool. Information, 16(11), Article 979. https://doi.org/10.3390/info16110979
[13] Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193–202. https://doi.org/10.3758/BF03195564
[14] 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 preprint arXiv:2301.07597.
[15] Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. Longman.
[16] Harari, Y. N. (2023, April 28). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. The Economist. https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation
[17] Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific Reports, 13(1), Article 18617. https://doi.org/10.1038/s41598-023-45644-9
[18] Ismail, H. Y. S. (2023). Cohesion and coherence in essays generated by ChatGPT: A comparative analysis to university students' writing. CDELT Occasional Papers in the Development of English Education, 83, 143–165. https://opde.journals.ekb.eg/article_325331.html
[19] Kamalski, J., Lentz, L., Sanders, T., & Zwaan, R. A. (2008). The forewarning effect of coherence markers in persuasive discourse: Evidence from persuasion and processing. Discourse Processes, 45(6), 545–579. https://doi.org/10.1080/01638530802069983
[20] Károly, K. (2017). Aspects of cohesion and coherence in translation: The case of Hungarian-English news translation. John Benjamins Publishing Company. https://doi.org/10.1075/btl.134
[21] Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). Handbook of latent semantic analysis. Lawrence Erlbaum Associates.
[22] Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Li, Q., Liu, T., & Li, X. (2023). Differentiating ChatGPT-generated and human-written medical texts: Quantitative study. JMIR Medical Education, 9, Article e48904. https://doi.org/10.2196/48904
[23] Mao, F. (2021). A re-analysis of cohesion and coherence. International Journal of Social Sciences and Management Review, 4(4), 37–45. https://doi.org/10.37602/IJSSMR.2021.4404
[24] Martin, J. R. (2001). Cohesion and texture. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), The handbook of discourse analysis (pp. 35–53). Blackwell.
[25] 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
[26] McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Using LSA to automatically assess coherence. In Proceedings of the Florida Artificial Intelligence Research Society Conference (pp. 637–642). AAAI Press.
[27] McCarthy, P. M., Renner, A. M., Duncan, M. R., Lightman, E. J., McNamara, D. S., & Graesser, A. C. (2007). Constructing a predictive measure of lexical sophistication. In Proceedings of the 29th Annual Conference of the Cognitive Science Society (pp. 1255–1260). Cognitive Science Society.
[28] McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press.

[29] McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47(4), 292–330. https://doi.org/10.1080/01638530902959943
[30] McNamara, D. S., Ozuru, Y., Graesser, A. C., & Louwerse, M. (2006). Validating Coh-Metrix. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 573–578). Cognitive Science Society.
[31] Millis, K. K., Graesser, A. C., & Haberlandt, K. (1993). The impact of connectives on the memory for expository texts. Applied Cognitive Psychology, 7(4), 317–339.
[32] Navrátilová, O., Dontcheva-Navrátilová, O., Jančaříková, R., Miššíková, G., & Povolná, R. (Eds.). (2017). Coherence and cohesion in English discourse. Masaryk University Press.
[33] Qiu, Z., Duan, X., & Cai, Z. G. (2023). Does ChatGPT resemble humans in processing implicatures? In Proceedings of the 4th Natural Logic Meets Machine Learning Workshop (pp. 25–34). Association for Computational Linguistics. https://aclanthology.org/2023.naloma-1.3
[34] Revell, T., Yeadon, W., Cahilly-Bretzin, G., & Inyang, O. O. A. (2024). ChatGPT versus human essayists: An exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. International Journal for Educational Integrity, 20, Article 18. https://doi.org/10.1007/s40979-024-00161-8
[35] Ripoll Y Schmitz, L. M., & Sonnleitner, P. (2025). Evaluating AI-generated vs. human-written reading comprehension passages: An expert SWOT analysis and comparative study for an educational large-scale assessment. Large-scale Assessments in Education, 13, Article 20. https://doi.org/10.1186/s40536-025-00255-w
[36] Spooren, W., & Sanders, T. (2008). The acquisition order of coherence relations: On cognitive complexity in discourse. Journal of Pragmatics, 40(12), 2003–2026. https://doi.org/10.1016/j.pragma.2008.04.021
[37] Trueswell, J. C., & Tanenhaus, M. K. (1991). Tense, temporal context, and syntactic ambiguity resolution. Language and Cognitive Processes, 6(4), 303–338. https://doi.org/10.1080/01690969108406961
[38] van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. Academic Press.
[39] Widdowson, H. G. (2004). Text, context, pretext: Critical issues in discourse analysis. Blackwell Publishing. https://doi.org/10.1002/9780470758427
[40] xAI. (n.d.). Grok. https://x.ai/grok
[41] Zhang, X., Diaz, A., Chen, Z., Wu, Q., Qian, K., Voss, E., & Yu, Z. (2024). DECOR: Improving coherence in L2 English writing with a novel benchmark for incoherence detection, reasoning, and rewriting. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 11436–11458). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.639
[42] Zhou, T., Cao, S., Zhou, S., Zhang, Y., & He, A. (2023). Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing. System, 118, Article 103141. https://doi.org/10.1016/j.system.2023.103141
[43] Zwaan, R. A. (1996). Processing narrative time shifts. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(5), 1196–1207. https://doi.org/10.1037/0278-7393.22.5.1196
[44] Zwaan, R. A. (2024). Comprehension: From clause to conspiracy narrative. Discourse Processes, 61(4-5), 166–179. https://doi.org/10.1080/0163853X.2024.2327271
[45] Zwaan, R. A. (2025). From words to worlds: Twenty-five years of advances in situation model research. Current Directions in Psychological Science. Advance online publication. https://doi.org/10.1177/09637214251326812
[46] Zwaan, R. A., Madden, C., & Stanfield, R. (2001). Time in narrative comprehension: A cognitive perspective. In D. Schram & G. Steen (Eds.), The psychology and sociology of literature: In honor of Erlund Ibsch (pp. 71–86). John Benjamins Publishing Company.
[47] Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. Psychological Bulletin, 123(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162
[48] Zwaan, R. A., & Rapp, D. N. (2006). Discourse comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), Handbook of psycholinguistics (2nd ed., pp. 725–764). Academic Press.