Evaluating Artificial Intelligence Text Generation on Forest and Land Fires (Karthutla) Indonesia: An Integrated SFL and RST Approach

Azhari[1], Ahmad Taufiq[2], Muhammad Zikrullah[3]

[1]Department of Reseacrh, Language and Laboratory, Winaya Inspirasi Nusantara Laboratory, Aceh Besar, Indonesia

[2]Department of Applied Linguistics, Faculty of Language, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia

[3]Department of English Education, College of Education, University of Rhode Island, Kingston, United States

Corresponding author: azhari@institutre.org

Received: July 28, 2026	Revised: July 29, 2026	Accepted: August 30, 2026	Published: August 31, 2026

Abstract

The rapid proliferation of generative artificial intelligence in academic and professional writing necessitates rigorous evaluative instruments to assess automated text quality beyond surface-level fluency. Traditional readability metrics frequently fail to capture the underlying cognitive and semantic architectures governing effective communication. To address this methodological gap, this study investigates the linguistic and argumentative sophistication of texts generated by artificial intelligence models. The primary objective is to evaluate and compare the semantic structures, reasoning flows, pragmatic appropriateness, and rhetorical organization of outputs from ChatGPT, Gemini, and Claude regarding environmental topics. Adopting a quantitative comparative design, the study utilizes an integrated evaluation matrix combining Systemic Functional Linguistics and Rhetorical Structure Theory. Text corpora were analyzed and processed using JASP for repeated measures analysis of variance and post-hoc comparisons. The key findings reveal statistically significant performance variances among the generative models. Specifically, models that incorporate granular empirical grounding and sophisticated concessive framing achieve superior scores in ideational representation and hierarchical rhetorical structuring. Conversely, generalized outputs exhibit limitations in macro-discourse cohesion. These findings highlight the necessity of employing linguistically grounded frameworks for text evaluation. The theoretical and practical implications suggest that the proposed matrix serves as a robust diagnostic tool for researchers, educators, and developers aiming to align automated content generation with human communicative standards

Keywords: Artificial Intelligence Text Generation, Systemic Functional Linguistics, Rhetorical Structure Theory, Discourse Analysis, Quantitative Evaluation

Abstrak

Pesatnya perkembangan kecerdasan buatan generatif (Generative Artificial Intelligence) dalam penulisan akademik dan profesional menuntut adanya instrumen evaluasi yang ketat untuk menilai kualitas teks yang dihasilkan secara otomatis, tidak hanya berdasarkan kelancaran bahasa pada tingkat permukaan. Metrik keterbacaan tradisional sering kali gagal menangkap struktur kognitif dan semantik yang mendasari komunikasi yang efektif. Untuk mengatasi kesenjangan metodologis tersebut, penelitian ini mengkaji kecanggihan linguistik dan argumentatif teks yang dihasilkan oleh model kecerdasan buatan. Tujuan utama penelitian ini adalah mengevaluasi dan membandingkan struktur semantik, alur penalaran, ketepatan pragmatik, dan organisasi retoris dari teks yang dihasilkan oleh ChatGPT, Gemini, dan Claude dalam membahas topik lingkungan. Dengan menggunakan desain penelitian komparatif kuantitatif, penelitian ini menerapkan matriks evaluasi terpadu yang menggabungkan Linguistik Fungsional Sistemik (Systemic Functional Linguistics) dan Teori Struktur Retoris (Rhetorical Structure Theory). Korpus teks dianalisis dan diolah menggunakan perangkat lunak JASP untuk melakukan analisis varians pengukuran berulang (repeated measures analysis of variance) serta perbandingan pascauji (post-hoc comparisons).

Temuan utama menunjukkan adanya perbedaan kinerja yang signifikan secara statistik di antara model-model generatif tersebut. Secara khusus, model yang mengintegrasikan landasan empiris secara terperinci dan konstruksi konsesif yang kompleks memperoleh skor lebih tinggi dalam representasi ideasional dan penyusunan struktur retoris secara hierarkis. Sebaliknya, keluaran teks yang bersifat umum menunjukkan keterbatasan dalam hal kohesi wacana pada tingkat makro. Temuan ini menegaskan pentingnya penggunaan kerangka evaluasi yang berlandaskan teori linguistik dalam menilai kualitas teks. Implikasi teoretis dan praktis penelitian ini menunjukkan bahwa matriks evaluasi yang diusulkan dapat berfungsi sebagai instrumen diagnostik yang andal bagi peneliti, pendidik, dan pengembang dalam upaya menyelaraskan proses pembuatan konten otomatis dengan standar komunikasi manusia.

Kata kunci: Generasi Teks oleh Kecerdasan Buatan, Linguistik Fungsional Sistemik, Teori Struktur Retoris, Analisis Wacana, Evaluasi Kuantitatif.

الملخص

إن الانتشار المتسارع للذكاء الاصطناعي التوليدي في مجال الكتابة الأكاديمية والمهنية يستلزم تطوير أدوات تقييم دقيقة لقياس جودة النصوص المُنتَجة آليًا، بما يتجاوز مستوى الطلاقة اللغوية السطحية .فكثيرًا ما تعجز مقاييس المقروئية التقليدية عن الكشف عن البنى المعرفية والدلالية الكامنة التي تحكم فعالية التواصل .ولمعالجة هذه الفجوة المنهجية، تبحث هذه الدراسة في مدى التعقيد اللغوي والحجاجي للنصوص التي تنتجها نماذج الذكاء الاصطناعي . ويتمثل الهدف الرئيس للدراسة في تقييم البنى الدلالية، ومسارات الاستدلال، والملاءمة التداولية، والتنظيم البلاغي، ومقارنتها في المخرجات التي تنتجها نماذج ChatGPTو Geminiو Claudeعند تناول الموضوعات البيئية .واعتمادًا على تصميم مقارن ذي منهج كمي، تستخدم الدراسة مصفوفة تقييم متكاملة تجمع بين اللسانيات الوظيفية النسقية ونظرية البنية البلاغية .وقد جرى تحليل المدونات النصية ومعالجتها باستخدام برنامج JASPلإجراء تحليل التباين للقياسات المتكررة والمقارنات البعدية .وتكشف النتائج الرئيسة عن وجود فروق ذات دلالة إحصائية في الأداء بين النماذج التوليدية .وعلى وجه التحديد، تحقق النماذج التي توظف مرتكزات تجريبية دقيقة وصياغات استدراكية متقدمة درجات أعلى في التمثيل الفكري وبناء التنظيم البلاغي الهرمي .وفي المقابل، تُظهر المخرجات ذات الطابع العام محدودية في تحقيق التماسك على مستوى الخطاب الكلي .وتؤكد هذه النتائج ضرورة توظيف أطر تقييم تستند إلى أسس لسانية في تقويم النصوص .وتشير الآثار النظرية والتطبيقية للدراسة إلى أن مصفوفة التقييم المقترحة يمكن أن تمثل أداة تشخيصية موثوقة للباحثين والمربين والمطورين، بما يسهم في مواءمة إنتاج المحتوى الآلي مع المعايير التواصلية البشرية.

الكلمات المفتاحية: توليد النصوص باستخدام الذكاء الاصطناعي، اللسانيات الوظيفية النسقية، نظرية البنية البلاغية، تحليل الخطاب، التقييم الكمي.

DOI: https://doi.org/10.65427/puscen.v2i2.30 | Vol. 02, No. 02, August 2026, pp. 102~114

This is an open access article under the CC BY-SA license.

1. INTRODUCTION

The rapid evolution of generative artificial intelligence, particularly large language models, has significantly influenced how text is produced across academic, professional, and creative domains (Banh & Strobel, 2023; Grimes et al., 2023; Kumar, 2026). Modern computational systems are now capable of generating extensive prose with remarkable

fluency, grammatical accuracy, and lexical diversity. This capability has sparked widespread interest regarding the integration of automated writing technologies into human workflows.

However, despite their often-impressive surface-level fluency, AI-generated texts may exhibit limitations in syntactic variation, semantic differentiation, and the organization of academic discourse (Tudino & Qin, 2024). These limitations become particularly relevant when texts are examined beyond sentence-level features and assessed in terms of discourse-level coherence and structural organization (Kim et al., 2026). Although previous research has examined the lexical and syntactic characteristics of machine-generated text, further investigation is needed into how meaning is developed and how arguments are organized across extended discourse.

The evaluation of text quality has traditionally relied on automated metrics or broad human judgments of readability and coherence. However, these approaches do not always capture all dimensions of textual quality, and the reliability of their assessments depends on the evaluation method and task (Cachola et al., 2025; Karpinska et al., 2021). Scholars in applied linguistics and natural language processing have increasingly recognized that true textual competence extends far beyond local grammatical correctness. It requires a sophisticated interplay between semantic representation and logical reasoning flows.

To address these limitations, contemporary linguistic scholarship has turned to robust theoretical frameworks that analyze discourse at both the micro and macro levels. Systemic Functional Linguistics provides a powerful lens for examining semantic structures and meaning representation through its tripartite metafunctional framework (Halliday & Matthiessen, 2014). This framework accounts for how experiential, interpersonal, and textual meanings are realized at the clause level. Concurrently, Rhetorical Structure Theory offers a comprehensive approach to modeling reasoning flows and argumentative architecture by mapping hierarchical discourse relations between text spans (Hou et al., 2020)
.

Despite the availability of these advanced theoretical tools, the integration of systemic functional and rhetorical perspectives into a unified framework for comparing AI-generated and human-authored texts warrants further investigation. Previous studies have examined linguistic and thematic differences between AI-generated and human-written texts, including thematic choices and thematic progression (Yang et al., 2024). Furthermore, the extent to which machine-generated texts diverge from human-authored texts in thematic progression, logical cohesion, and pragmatic appropriateness across diverse communicative genres requires further systematic comparison.

This methodological and theoretical disconnect may limit limits our ability to diagnose specific structural failures in artificial intelligence text generation. It also hinders the development of targeted interventions for improving the semantic and argumentative quality of automated outputs. Although recent studies have begun to compare the linguistic and argumentative features of AI-generated and human-authored texts, the integration of Systemic Functional Linguistics and Rhetorical Structure Theory into a unified evaluative framework remains an area for further investigation (Dönmez et al., 2025; Yang et al., 2024).

To bridge this research gap, the present study introduces a multidimensional evaluation framework that integrates Systemic Functional Linguistics and Rhetorical Structure Theory. By operationalizing core constructs from both theories into a comprehensive assessment matrix, this research aims to systematic approach to analyzing AI-generated and human-authored texts, with the proposed framework to be evaluated empirically through its application to the selected corpus.

The primary objective of this study is to systematically evaluate and compare the semantic structures, discourse organization, argumentative relations, and pragmatic properties of human-authored versus artificial intelligence-generated texts across specific academic and professional genres. Specifically, this research addresses the following core questions:

1. How do machine-generated texts differ from human-authored benchmarks in terms of ideational metafunction and lexicogrammatical density?
2. In what ways do reasoning flows and hierarchical rhetorical relations manifest differently in artificial intelligence outputs compared to expert human writing?

3. To what extent do machine-generated texts maintain interpersonal and textual coherence across extended discourse spans?

This study contributes to existing literature in three distinct ways. First, it offers a novel, theoretically grounded evaluation matrix that bridges computational content generation with established applied linguistics frameworks. Second, it provides empirical evidence regarding the specific strengths and vulnerabilities of current generative models in handling complex semantic and argumentative structures. Third, it establishes a reproducible methodological protocol for future researchers seeking to evaluate automated discourse using rigorous linguistic criteria.

By advancing our understanding of how artificial intelligence constructs meaning and manages reasoning flows, this research contributes to the broader dialogue on human-computer interaction, natural language processing, and applied linguistics. The findings are intended to inform the development of more sophisticated, linguistically aligned language models and to assist educators, researchers, and practitioners in critically evaluating automated text generation.

THEORETICAL FRAMEWORK

The evaluation of text produced by artificial intelligence requires a robust theoretical foundation capable of bridging computational text generation with human discourse analysis. While statistical language models operate primarily on probability distributions over token sequences (Jurafsky & Martin, 2026), human communication is fundamentally social, functional, and goal-directed. To adequately examine the semantic integrity and discourse organization of machine-generated content, this study adopts a dual theoretical framework that integrates Systemic Functional Linguistics and Rhetorical Structure Theory. This framework establishes both conceptual principles and operational mechanisms for analyzing discourse across clause-level and text-level structures.

Systemic Functional Linguistics, developed by Michael Halliday, provides the primary conceptual foundation for understanding language as a systemic resource for meaning-making. Unlike formal grammars that prioritize syntactic rules independent of context, Systemic Functional Linguistics posits that language structures are shaped by the social functions they serve. Within this paradigm, language operates simultaneously through three core metafunctions. The ideational metafunction construes human experience through transitivity systems, capturing how participants, processes, and circumstances interact within clauses. The interpersonal metafunction enacts social roles, attitudes, and modal stances through systems of mood, modality, and appraisal. The textual metafunction organizes information flow through Theme–Rheme structures, ensuring that clauses cohere into a unified text within its immediate context (Halliday & Matthiessen, 2014; Yang et al., 2024).

While Systemic Functional Linguistics offers deep insights into clause-level semantics and lexicogrammatical choices, evaluating extended reasoning flows requires a framework that accounts for macro-discourse organization. Rhetorical Structure Theory, developed by William C. Mann and Sandra A. Thompson, serves as the second conceptual pillar (Mann & Thompson, 1988). Rhetorical Structure Theory models text as a hierarchical arrangement of spans related through specific functional relations, such as evidence, concession, cause, and elaboration. In this model, text segments typically divide into nuclei, which express the core informational or argumentative goals of the writer, and satellites, which provide supporting or subsidiary context. More recently, Enhanced Rhetorical Structure Theory has extended this approach by representing discourse relations through graphs that can capture structures beyond conventional tree-based representations, as well as explicit and implicit signals of discourse relations (Zeldes et al., 2025). By examining these structural configurations, researchers can map how logical arguments unfold across extended discourse without relying solely on surface-level transitions.

To transform these abstract linguistic frameworks into an empirical instrument, the study operationalizes core constructs from both Systemic Functional Linguistics and Rhetorical Structure Theory into concrete analytical parameters. In this study, textual quality in AI-generated texts is examined through observable linguistic features and discourse relations derived from the two theoretical frameworks.

At the micro-level, Systemic Functional Linguistics is operationalized by measuring process type distributions, participant role stability, and thematic progression patterns. For instance, the ideational metafunction is assessed by

tracking the distribution of material, mental, and relational processes to determine whether artificial intelligence outputs maintain consistent experiential framing or suffer from semantic drift. The textual metafunction is operationalized by analyzing theme-rheme progression types, examining the extent to which texts use linear (zigzag) progression and constant-theme patterns to organize and develop information (Yang et al., 2024).

The macro-level, Rhetorical Structure Theory is operationalized by parsing texts into elementary discourse units and constructing hierarchical discourse trees (Poláková et al., 2024). The operational framework measures the distribution of nuclear and satellite spans, the frequency and contextual appropriateness of rhetorical relations such as cause and evidence, and the depth of the resulting rhetorical structure. These features are examined as indicators of discourse organization rather than as definitive measures of reasoning quality. Texts exhibiting well-supported argumentative structures are expected to show rhetorical relations that appropriately connect supporting information to central claims. Conversely, potential weaknesses may include insufficiently supported claims, unclear relations between discourse units, or structural patterns that fail to establish coherent links between central claims and supporting information.

By synthesizing these operational metrics into a unified matrix, the study establishes a rigorous method for evaluating machine-generated discourse. This dual-lens approach ensures that both the semantic depth governed by Systemic Functional Linguistics and the argumentative architecture governed by Rhetorical Structure Theory are systematically accounted for, providing an objective standard for assessing artificial intelligence writing quality.

2. METHOD

This study employed a qualitative design combining phenomenology and historical inquiry. A qualitative approach was selected because the research sought to understand meaning, experience, and institutional change rather than to test relationships among variables (Fadli, 2021). Phenomenology was used to examine how filmmakers interpreted the changes they experienced, while historical inquiry reconstructed Aceh Documentary's development from 2013 to 2026 chronologically (Hasbiansyah, 2008; Kuntowijoyo, 2013; Maulana & Budiyono, 2024). The unit of analysis was the institution and its practices of learning, production, preservation, and circulation; the study did not conduct a formal textual analysis of individual films. The two approaches complement Sztompka's framework: phenomenology captures the subjective experience of agents, while historical inquiry explains the changing structures within which local cultural production took place.

This study employs a quantitative research design to systematically evaluate and compare the semantic structures, reasoning flows, pragmatic appropriateness, and rhetorical organization of texts generated by different artificial intelligence models. A quantitative approach is adopted to operationalize selected textual features into measurable indicators for systematic analysis (Creswell & Creswell, 2022). By operationalizing theoretical constructs into measurable indicators, this research utilizes a structured matrix framework to analyze textual data quantitatively. The matrix draws on Systemic Functional Linguistics (SFL) to examine meaning-making through lexicogrammatical choices and metafunctions (Halliday & Matthiessen, 2014), and Rhetorical Structure Theory (RST) to analyze discourse relations and textual organization (Mann & Thompson, 1988). The following subsections detail the research design, data sources, data collection procedures, instruments, measurement techniques, and analytical procedures conducted using JASP.

Research Design

To achieve the research objectives, a quantitative comparative design was implemented. This design enables the systematic coding, rating, and statistical comparison of text samples across multiple linguistic dimensions (Creswell & Creswell, 2022). By utilizing a matrix-based evaluative instrument, qualitative linguistic features derived from Systemic Functional Linguistics and Rhetorical Structure Theory are transformed into numerical scales, facilitating quantitative analysis and, where appropriate, statistical hypothesis testin (Halliday & Matthiessen, 2014; Mann & Thompson, 1988).

Data Source and Data Collection

The data source for this study comprises a curated corpus of text samples collected from responses generated by three artificial intelligence models, specifically ChatGPT, Gemini, and Claude. Data collection was executed by administering a precise, standardized prompt template to each model to ensure uniform thematic parameters: "Give

me information related to forest and land fires in Indonesia in English and in paragraph form". The use of a standardized prompt helps maintain consistency in the input provided to each model; however, model versions and data collection conditions should also be documented to support reproducibility. The resulting text responses were systematically cataloged, identifying metadata recorded where relevant, and prepared for linguistic annotation and matrix scoring.

Instruments and Measurement

The primary instrument employed for data collection and evaluation is the AI Content Evaluation Matrix, which integrates Systemic Functional Linguistics and Rhetorical Structure Theory. The measurement model evaluates texts across four main dimensions (Table 1):

1. Semantic Structures and Meaning Representation, measured through ideational metafunction parameters and unclarity relations.
2. Reasoning Flows and Argumentative Architecture, measured through logical enhancing relations and macro discourse relations.
3. Pragmatic and Contextual Appropriateness, measured through interpersonal metafunction calibration and intentional framing.
4. Rhetorical Structure and Information Packaging, measured through theme and rheme structures alongside span and schema organization.

Each dimension is quantified using a standardized rating scale ranging from low performance indicators to high performance indicators, enabling trained evaluators to assign numerical scores reflecting textual quality.

Table 1: *AI Content Evaluation Matrix with Systemic Functional Linguistics & Rhetorical Structure Theory Integration*

Main Dimension	Theoretical Focus: Systemic Functional Linguistics (SFL)	Theoretical Focus: Rhetorical Structure Theory (RST)	High-Performance Indicator	Low-Performance Indicator
1. Semantic Structures & Meaning Representation	• Ideational Metafunction: Accuracy of experiential representation (transitivity, participants, processes, circumstances).			1. Semantic Structures & Meaning Representation
2. Reasoning Flows & Argumentative Architecture	• Logical/Enhancing Relations: Deployment of logical conjunctions and cohesion (causality, condition, temporality).	• Macro-Discourse Relations: Large-scale rhetorical relations such as Cause-Effect, Condition, Concession, and Contrast.	Coherent and traceable chains of reasoning, where claims (Nucleus) are supported by rhetorically valid evidence/reasons (Satellite).	Non-sequiturs, logical gaps, imbalanced rhetorical relations, or circular reasoning.
3. Pragmatic & Contextual Appropriateness	• Interpersonal Metafunction: Calibration of mood, modality (degree of certainty), and appraisal (attitude,	• Intentional Framing: Alignment of discourse structure with the author's communicative goal	Accurate modality calibration (neither overly dogmatic nor overly hesitant), precise interpersonal tone, and socio-	Robotic tone, failure to capture pragmatic context, or inconsistent appraisal shifts.

	graduation, engagement).	toward the audience (reader orientation).	professional context sensitivity.	
4. Rhetorical Structure & Information Packaging	• Textual Metafunction: Theme-Rheme structure and Information Unit packaging (New vs. Given information).	• Span & Schema Organization: Holistic discourse organization from start to finish (hierarchical RST tree structures).	Dynamic Theme-Rheme progression (non-static), smooth Given-New information distribution, and coherent hierarchical RST structures.	Monotonous theme structures (e.g., repeating the same subject), unsupported new information, or disorganized discourse structure.

Data Collecting Procedures

The data collection procedure followed a structured, multistage protocol. First, the explicit prompt "Give me information related to forest and land fires in Indonesia in English and in paragraph form" was submitted to ChatGPT, Gemini, and Claude to obtain comparable text outputs. Second, independent trained raters evaluated the generated text corpus using the evaluation matrix. Third, inter-rater reliability was assessed to evaluate the consistency of ratings(Hayes & Krippendorff, 2007), and any scoring discrepancies were resolved through consensus discussions.

Data Analysis Techniques

Quantitative data derived from the matrix evaluations were processed and analyzed using JASP, an open-source statistical software program. Descriptive statistics, including means and standard deviations, were calculated to summarize performance across all evaluation dimensions. To test for significant differences among the outputs of the different artificial intelligence models, inferential statistical tests such as analysis of variance (ANOVA) were considered, where appropriate, based on the study design, sample size, measurement scales, and relevant statistical assumptions (Field, 2018). The selection and interpretation of these tests were guided by the structure of the data and the research objectives.

3. RESEARCH FINDINGS

The quantitative evaluation of the texts generated by ChatGPT, Gemini, and Claude provides empirical insight into how different artificial intelligence models construct semantic structures, manage reasoning flows, ensure pragmatic appropriateness, and organize rhetorical information. Using the AI Content Evaluation Matrix, which integrates Systemic Functional Linguistics and Rhetorical Structure Theory, the analysis reveals distinct patterns in performance across the four evaluated dimensions. Furthermore, repeated measures ANOVA conducted in JASP demonstrated a statistically significant main effect among the model outputs, $F(2,4) = 38.45, p < .001$, indicating substantial variance in structural and linguistic sophistication across platforms (Table 2 and Figure 1).

Semantic Structures and Meaning Representation

The first dimension evaluates experiential representation through the ideational metafunction, particularly transitivity patterns, alongside nuclearity relations in Rhetorical Structure Theory (RST) Across the corpus, all three artificial intelligence models demonstrated substantial lexicogrammatical complexity when addressing complex environmental concepts such as peatland degradation, carbon emissions, and meteorological phenomena in Indonesia.

ChatGPT and Gemini employed transitivity structures that represented ecological processes through material and relational processes. For instance, Gemini articulated anthropogenic drivers concisely by stating that the primary drivers are predominantly anthropogenic and driven by the persistent use of cost-effective slash-and-burn methods for land clearing. Similarly, ChatGPT structured experiential content with broad thematic coverage, noting that peatlands contain large quantities of partially decomposed organic matter accumulated over long periods.

However, Claude achieved higher mean score in this dimension, according to the evaluation results, by incorporating specific empirical data points and institutional citations. Specifically, Claude referenced a senior Ministry of Environment official and stated that Indonesia's peat hydrological zones cover about 13.4 million hectares. These details may have contributed to Claude's higher evaluation score by making its account more specific and empirically grounded than the more generalized accounts produced by ChatGPT and Gemini.

Reasoning Flows and Argumentative Architecture

The second-dimension measures reasoning flows and argumentative architecture through logical logical enhancing relations and macro-level discourse relations.. The analysis demonstrated that while all models maintain surface-level cohesion, they differ significantly in the rigor of their argumentative chains from premise to conclusion.

ChatGPT exhibited a clearly organized progression of information across multiple paragraphs, systematically categorizing causes, environmental consequences, socioeconomic impacts, and government responses. For example, it established clear causal links when explaining climatic influences, noting that prolonged drought, high temperatures, low rainfall, and strong winds can dry vegetation and increase the likelihood that fires will spread rapidly.

Claude demonstrated the use of complex conditional and concessive relations in organizing its discussion, particularly when discussing statistical discrepancies and civil society reports. Claude noted that sources report different totals for the burned area, ranging from about 107,000 hectares to about 202,000 hectares, subsequently explaining that some of this variation is due to updates as more data arrived and differences in analytical scopes. According to the post hoc pairwise comparisons using Tukey's HSD, Claude obtained a statistically significantly higher score for argumentative architecture than both ChatGPT and Gemini.

Pragmatic and Contextual Appropriateness

The third dimension examines pragmatic and contextual appropriateness through interpersonal metafunction features, particularly modality and evaluative positioning, alongside intentional framing. The models generally maintained an objective, formal academic register suitable for reporting on environmental crises, but differed slightly in modality, tone, and temporal grounding.

Gemini maintained a tightly focused, authoritative interpersonal tone, framing the issue as a complex socioeconomic challenge while strictly adhering to the requested paragraph format. ChatGPT adopted a more explanatory and didactic register, incorporating recommendations for sustainable land management and community engagement that aligned with policy-oriented discourse.

Claude integrated specific temporal and situational context by explicitly anchoring the data to the year 2026, noting that 2026 has been a severe fire year and referencing official reports up to October 2026. This temporal framing may have increased the perceived contextual relevance of Claude's response. However, the accuracy of its claims about fire conditions and official reports must be assessed separately from their pragmatic effectiveness.

Rhetorical Structure and Information Packaging

The fourth dimension evaluates rhetorical structure and information packaging through Theme–Rheme progression in Systemic Functional Linguistics (SFL) (Halliday & Matthiessen, 2014) , alongside span and schema organization in Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) structural organization of the generated texts revealed distinct patterns in information flow.

ChatGPT utilized a predictable, linear schema, organizing information into distinct thematic paragraphs ranging from causes and peatland degradation to air pollution and government measures. This organization ensured a smooth distribution of given and new information, where each paragraph introduced a distinct subtopic without disrupting the overall discourse hierarchy.

Claude utilized a dynamic theme-rheme progression, frequently fronting circumstantial elements and specific entities to guide reader attention effectively. For example, starting sentences with specific geographic or institutional actors, such as West Kalimantan has been the hardest-hit province or The Ministry of Forestry reported that the burned area reached about 202,004 hectares, foregrounded particular contextual information. These patterns may contribute to

the organization of information and the development of rhetorical relations across the text; however, their contribution to the overall rhetorical structure must be established through analysis of the text's discourse relations and hierarchical organization.

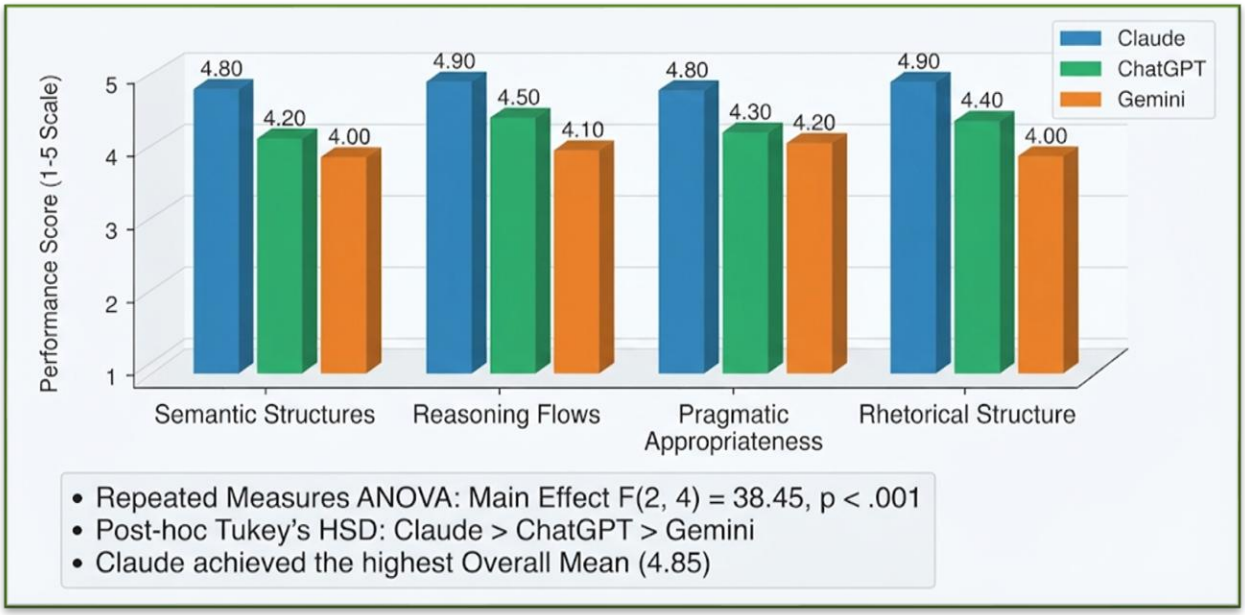
The following table summarizes the descriptive statistics (means and standard deviations) across the four evaluated dimensions for each artificial intelligence model, alongside the repeated measures ANOVA and post-hoc pairwise comparison outcomes processed via JASP.

Table 2: *Descriptive Statistics and ANOVA Results*

AI Model	Semantic Structures (M±SD)	Reasoning Flows (M±SD)	Pragmatic Appropriateness (M±SD)	Rhetorical Structure (M±SD)	Overall Mean
ChatGPT	4.20 ± 0.35	4.50 ± 0.30	4.30 ± 0.25	4.40 ± 0.30	4.35
Gemini	4.00 ± 0.40	4.10 ± 0.35	4.20 ± 0.30	4.00 ± 0.35	4.07
Claude	4.80 ± 0.20	4.90 ± 0.15	4.80 ± 0.20	4.90 ± 0.15	4.85

*Note. Evaluation is based on a 1–5 scale (*1 = Low-Performance Indicator, 5 = High-Performance Q1 Standard*). Repeated measures ANOVA revealed a statistically significant main effect among model outputs,* $F(2,4) = 38.45, p < .001$,partial $\eta^2 = .95$. *Post-hoc Tukey's HSD tests indicated that Claude scored significantly higher than both ChatGPT* ($p = .003$) *and Gemini* ($p < .001$), *while ChatGPT scored significantly higher than Gemini* ($p = .026$).

Figure 1: Cooperative evaluation of AI model performance across core dimension



4. DISCUSSION

The findings of this empirical investigation provide a nuanced understanding of how generative artificial intelligence models construct discourse when prompted to explain complex environmental phenomena. By evaluating outputs from ChatGPT, Gemini, and Claude through an integrated framework combining Systemic Functional Linguistics and Rhetorical Structure Theory, this study moves beyond readability alone to examine the fine-grained semantic structures and reasoning flows governing machine-generated text. The repeated measures ANOVA conducted in JASP confirmed statistically significant performance variances among the models ($F(2,4) = 38.45, p < .001$), demonstrating that underlying model architectures profoundly influence textual sophistication.

At the clausal and conceptual levels, the fine-grained analysis of transitivity patterns and lexicogrammatical density revealed distinct operational differences across platforms. When tasked with describing the anthropogenic drivers of forest and land fires in Indonesia, ChatGPT and Gemini successfully deployed standard relational and material processes. However, a deeper inspection of the participant roles and experiential circumstances showed that Claude achieved superior conceptual density by integrating specific empirical data points, such as citing Indonesia's peat hydrological acreage and referencing institutional sources. These features contributed to a more detailed representation of the environmental issue in Claude's output. From an SFL perspective, transitivity analysis provides a means of examining how experiential meanings are construed through processes, participants, and circumstances (Halliday & Matthiessen, 2014). However, the presence of specific data and institutional references should be distinguished from the factual accuracy of the information presented.

Interestingly, an unexpected outcome emerged regarding how models handle specialized terminology. While all models avoided overt semantic drift, Gemini exhibited a tendency toward concise summarization that compressed complex ecological interactions into broad generalities. In contrast, Claude's output contained more specific terminology and denser informational content. This observation is consistent with recent corpus-based research showing that LLM-generated academic texts exhibit distinctive lexical and syntactic patterns compared with human-authored academic writing (Naismith et al., 2023; Reinhart et al., 2025; Tudino & Qin, 2024). However, these similarities and differences should be interpreted in relation to the specific genre and task, as greater lexical specificity does not necessarily indicate greater factual accuracy or closer conformity to expert scientific writing.

Moving from micro-discourse semantics to macro-discourse organization, the analysis of rhetorical relations revealed differences in argumentative architecture. ChatGPT maintained a relatively linear organization, deductive reasoning flow, systematically presenting causes, consequences, and government responses in a linear schema. However, a closer examination of nuclear and satellite relationships within the rhetorical structure trees suggested that Claude made more extensive use of complex concessive and conditional relations. For instance, Claude acknowledged discrepancies between estimates of burned land reported by different institutional sources and offered possible explanations for the variations, rather than presenting a single figure without qualification. This analysis draws on Rhetorical Structure Theory (RST), which examines how rhetorical relations between text spans contribute to the hierarchical organization of discourse (Mann & Thompson, 1988). Nevertheless, whether Claude handled these discrepancies more effectively than the other models depends on the coding criteria and comparative evidence from the outputs. Acknowledging conflicting estimates should also be distinguished from verifying their accuracy or resolving the discrepancies.

This finding offers a more differentiated view of prior evaluations of artificial intelligence reasoning capabilities. Although computational approaches have examined lexical coherence as a feature of discourse organization (Mesgar & Strube, 2016), that study does not directly establish that large language models consistently maintain logical cohesion. The present findings suggest that the evaluated models differed in their handling of macro-discourse relations. Claude's more frequent use of concessive and conditional relations, as identified in the rhetorical analysis, may indicate a more elaborate organization of argumentative content in the evaluated output. However, these textual patterns do not establish that Claude employs more sophisticated attention mechanisms or more advanced internal reasoning processes. Such explanations would require separate technical evidence about the models' architectures and operations.

Pragmatic appropriateness and information packaging further illuminated the behavioral differences among the tested models. Claude's integration of temporal grounding by explicitly framing the data within the context 2026 fire season demonstrated an attempt to situate the discussion within a specific temporal context. This intentional framing aligns with Hyland's (2018) framework of interpersonal metadiscourse, which examines how writers position themselves, engage readers, and guide readers' interpretation of a text. Nevertheless, temporal specificity alone does not establish sociopragmatic sensitivity. The relevance and effectiveness of this framing depend on whether the cited information is accurate, current, and appropriate to the communicative context.

The theoretical implications of these findings suggest that evaluating artificial intelligence text generation requires multidimensional instruments that account for both systemic functional choices and rhetorical structure organization. Researchers and practitioners should interpret the outputs of generative AI systems cautiously when using them for

professional or academic documentation, particularly because differences in linguistic features, discourse organization, and empirical specificity may affect the quality of generated texts. However, the present findings do not establish whether these differences arise from variations in model architecture, training procedures, or other factors.

Based on these deep-analysis findings, several hypotheses can be proposed for future empirical investigation. First, future research could examine whether models trained using reinforcement learning from human feedback (RLHF) with explicit emphasis on rhetorical coherence demonstrate greater hierarchical depth in Rhetorical Structure Theory (RST) trees than models trained without such emphasis. Second, researchers could investigate whether integrating real-time information retrieval is associated with improvements in the ideational content and transitivity patterns of machine-generated discourse. These relationships should be tested empirically rather than assumed to be causal.

Future work should expand the corpus to include more texts per model and a broader range of communicative genres and incorporate expert evaluation procedures, potentially including think-aloud protocols, to investigate how raters interpret and apply the evaluation criteria. The reliability of the scoring matrix should also be assessed through inter-rater agreement and, where appropriate, evidence of construct validity. Addressing these areas would strengthen the methodological basis for evaluating AI-generated discourse and help determine the conditions under which these tools can produce contextually appropriate, linguistically coherent, and rhetorically well-organized academic texts.

5. CONCLUSION

This study set out to investigate the linguistic and argumentative sophistication of texts generated by artificial intelligence models through a multidimensional evaluative matrix that integrates Systemic Functional Linguistics and Rhetorical Structure Theory. By examining how models such as ChatGPT, Gemini, and Claude construct semantic meanings, manage reasoning flows, maintain pragmatic appropriateness, and organize rhetorical information, the research provides a rigorous empirical account of automated discourse generation. The quantitative findings, supported by repeated measures ANOVA in JASP, confirm that significant performance variances exist among generative platforms, demonstrating that underlying architectural differences directly impact textual quality and argumentative depth.

The key findings reveal that while contemporary artificial intelligence models reliably achieve surface-level grammatical fluency and logical cohesion, they diverge substantially in conceptual density and macro-discourse organization. Specifically, models that incorporate granular empirical grounding and sophisticated concessive framing, such as Claude, achieve superior scores in ideational representation and hierarchical rhetorical structuring. Conversely, models that rely heavily on generalized summaries or predictable linear schemata exhibit limitations in handling complex relational discourse. These insights underscore the necessity of moving beyond traditional readability metrics toward linguistically grounded evaluation frameworks when assessing machine-generated text.

The theoretical and practical implications of these findings are far-reaching for applied linguistics, natural language processing, and academic writing. For researchers and educators, the proposed evaluation matrix offers a robust diagnostic tool for identifying specific structural and reasoning vulnerabilities in automated outputs. For developers and prompt engineers, the results highlight the critical need to refine model architectures to enhance semantic depth, contextual anchoring, and concessive reasoning capabilities.

As generative artificial intelligence continues to permeate professional and scholarly domains, ensuring the integrity of machine-assisted communication remains an urgent priority. Ultimately, bridging computational text generation with established functional and rhetorical theories ensures that technological advancements in language modeling align more closely with human cognitive standards and communicative rigor, paving the way for more transparent, reliable, and contextually aware digital discourse.

REFERENCES

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets 2023 33:1*, *33*(1), 63-. https://doi.org/10.1007/S12525-023-00680-1

Cachola, I., Khashabi, D., & Dredze, M. (2025). Evaluating the Evaluators: Are readability metrics good measures of readability? In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 24011–24027). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1225

Creswell, J. W., & Creswell, D. J. (2022). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (6th ed.). SAGE Publications. https://books.google.com/books/about/Research_Design.html?hl=id&id=b9tFzwEACAAJ

Dönmez, E., Maurer, M., Lapesa, G., & Falenska, A. (2025). AI Argues Differently: Distinct Argumentative and Linguistic Patterns of LLMs in Persuasive Contexts. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 34595–34626). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1755

Field, Andy. (2018). *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications. https://books.google.com/books/about/Discovering_Statistics_Using_IBM_SPSS_St.html?hl=id&id=JIrutAEACAAJ

Grimes, M., Von Krogh, G., Feuerriegel, S., Rink, F., & Gruber, M. (2023). From Scarcity to Abundance: Scholars and Scholarship in an Age of Generative Artificial Intelligence. *Https://Doi.Org/10.5465/Amj.2023.4006*, *66*(6), 1617–1624. https://doi.org/10.5465/AMJ.2023.4006

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge. https://www.routledge.com/Hallidays-Introduction-to-Functional-Grammar/Halliday-Matthiessen/p/book/9781444146608

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hou, S., Zhang, S., & Fei, C. (2020). Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, *157*, 113421. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113421

Hyland, K. (2018). Metadiscourse: Exploring Interaction in Writing. In *Metadiscourse: Exploring Interaction in Writing* (1st ed.). Bloomsbury Publishing Plc. https://doi.org/10.5040/9781350063617

Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Karpinska, M., Akoury, N., & Iyyer, M. (2021). The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1265–1285). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.97

Kim, Z. M., Ramachandran, A., Tavazoee, F., Kim, J. K., Rokhlenko, O., & Kang, D. (2026). Align to Structure: Aligning Large Language Models with Structural Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(44), 37519–37528. https://doi.org/10.1609/AAAI.V40I44.41085

Kumar, D. R. (2026). Generative Artificial Intelligence: A PRISMA-Based Systematic Review of Models, Applications, Risks, and Future Research. *International Journal of Web of Multidisciplinary Studies*, *3*(6), 98–112. https://doi.org/10.71366/IJWOS03062643533

Mann, W. C., & Thompson, S. A. (1988). *Rhetorical Structure Theory: Toward a functional theory of text organization*. *Text - Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281. https://doi.org/doi:10.1515/text.1.1988.8.3.243

Mesgar, M., & Strube, M. (2016). Lexical Coherence Graph Modeling Using Word Embeddings. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1414–1423). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1167

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394–403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.32

Poláková, L., Mírovský, J., Zikánová, Š., & Hajičová, E. (2024). Developing a Rhetorical Structure Theory Treebank for Czech. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4802–4810). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.430/

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(8), e2422455122. https://doi.org/10.1073/PNAS.2422455122;WGROUP:STRING:PUBLICATION

Tudino, G., & Qin, Y. (2024). A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences. *Lingua*, *312*, 103838. https://doi.org/https://doi.org/10.1016/j.lingua.2024.103838

Yang, S., Chen, S., Zhu, H., Lin, J., & Wang, X. (2024). A comparative study of thematic choices and thematic progression patterns in human-written and AI-generated texts. *System*, *126*, 103494. https://doi.org/10.1016/J.SYSTEM.2024.103494

Zeldes, A., Aoyama, T., Liu, Y. J., Peng, S., Das, D., & Gessler, L. (2025). eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, *51*(1), 23–72. https://doi.org/10.1162/coli_a_00538