IKR Publishers

IKR Journal of Multidisciplinary Studies (IKRJMS)
Journal homepage: https://ikrpublishers.com/ikrjms/
Volume-2, Issue-6 (November-December) 2026
OPEN ACCESS
ISSN: 3107-3999 (Online)

Grammatical Categoricity and Contextual Variability: A Corpus-Based Study of AI-Generated and Human-Written English in Academic Prose

Pethias Siame*
Department of Literature and Languages, Kwame Nkrumah University, Zambia

DOI: *10.5281/zenodo.24447956*

ARTICLE INFO

Article history:
Received : 25-10-2026
Accepted : 07-11-2026
Available online : 23-11-2026

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial **4.0 International License (CC BY-NC 4.0),** which permits non-commercial use, sharing, and reproduction in any medium, provided the original author(s) and source are appropriately credited.

Citation: Siame, P. (2026). Grammatical Categoricity and Contextual Variability: A Corpus-Based Study of AI-Generated and Human-Written English in Academic Prose. *IKR Journal of Multidisciplinary Studies (IKRJMS), 2*(6), 19-30.

ABSTRACT **Original Research Article**

The rapid integration of generative artificial intelligence (AI) into academic writing has prompted urgent questions about the linguistic character of machine-produced text and its relationship to established norms of scholarly English. This study investigates whether AI-generated academic prose exhibits distinctive grammatical and lexical patterns when compared with human-written texts, and whether such patterns vary according to disciplinary context. Drawing on a corpus of 120 academic abstracts (60 generated by a large language model, 60 authored by human scholars) across three disciplines, linguistics, literary studies, and the natural sciences, the study employs a mixed-methods design informed by Systemic Functional Linguistics (SFL) and register theory. Quantitative analysis examines syntactic complexity, lexical diversity, nominalisation density, and modality distribution; qualitative analysis explores thematic progression and hedging strategies through close textual examination. Findings indicate that AI-generated texts display significantly higher nominalisation density and more uniform modality profiles, while human texts exhibit greater interpersonal variability and discipline-specific hedging. The study argues that AI-generated English constitutes a recognisable register rather than a neutral imitation of human academic style, with implications for academic integrity, writing pedagogy, and the sociology of scholarly communication. It concludes by proposing a descriptive framework for identifying register markers in machine-generated academic discourse.

Keywords: AI-Generated Text, Academic Register, Systemic Functional Linguistics, Corpus Linguistics, Nominalisation, Modality, Disciplinary Variation.

**Corresponding author: Pethias Siame*
Department of Literature and Languages, Kwame Nkrumah University, Zambia

Introduction

The emergence of large language models (LLMs) capable of producing fluent, extended prose has transformed the landscape of academic writing. Tools such as ChatGPT, Claude, and Gemini are now routinely deployed in drafting, editing, and summarising scholarly texts, prompting a recalibration of what constitutes authorship, originality, and linguistic authenticity in the academy (Kasozwe et al., 2025). Yet despite widespread commentary on the ethical and pedagogical implications of these technologies, comparatively little empirical attention has been paid to the *linguistic* character of AI-generated academic English, that is, to the specific grammatical, lexical, and rhetorical features that distinguish machine-produced prose from human-authored scholarship (Siame, 2026a; Siame et al., 2025; Siame, 2022).

This gap is significant for linguistics as a discipline. If AI-generated texts systematically differ from human texts in their grammatical and rhetorical organisation, then the assumptions underpinning corpus-based descriptions of academic register may require revision. Moreover, if such differences vary across disciplinary contexts, this would suggest that AI models reproduce not a single homogeneous academic style but a compressed, generalised approximation of multiple registers. The present study addresses these

questions through a comparative corpus analysis of AI-generated and human-written academic abstracts, framed within the theoretical apparatus of Systemic Functional Linguistics (SFL) and register theory.

The study is guided by three research questions:

1. What grammatical and lexical features distinguish AI-generated academic English from human-written academic English?
2. How does disciplinary context modulate these features?
3. To what extent can AI-generated academic prose be described as a distinct register rather than an imitation of existing registers?

The significance of these questions extends beyond purely descriptive linguistics. As generative AI becomes increasingly embedded in scholarly workflows, understanding the linguistic properties of machine-generated text has become essential for academic integrity, writing pedagogy, and our theoretical understanding of register variation itself. If AI-generated prose constitutes a recognisable register with identifiable linguistic markers, this has profound implications for how we conceptualise authorship, evaluate scholarly texts, and teach academic writing in an era of human-machine collaboration.

The present study makes three principal contributions to the field. First, it provides a systematic, theoretically grounded comparison of AI-generated and human-written academic prose across multiple disciplines, addressing a significant gap in the existing literature. Second, it develops a descriptive framework for identifying register markers in machine-generated text, offering practical tools for researchers and educators. Third, it advances theoretical understanding of how register is realised in machine-generated language, contributing to ongoing debates about the nature of linguistic variation in the age of artificial intelligence.

Literature Review

AI-Generated Text and Linguistic Description

Research on machine-generated text has historically focused on detection (e.g., watermarking, perplexity-based classifiers) rather than linguistic description (Siame, 2026b). The development of increasingly sophisticated language models has stimulated a corresponding growth in detection methodologies, with researchers developing techniques based on statistical properties of text, stylistic markers, and neural network classifiers (Jawahar et al., 2020; Solaiman et al., 2019). However, a growing body of work has begun to examine the stylistic and grammatical properties of LLM output.

Studies have noted tendencies toward lexical repetition, syntactic parallelism, and formulaic transitions in AI-generated prose (Herbold et al., 2023; Muñoz-Ortiz et al., 2023; Siame, 2026b). Herbold and colleagues, in their large-scale comparison of human-written versus ChatGPT-generated essays, found that AI-generated texts exhibited greater uniformity in sentence structure and vocabulary, with a tendency toward what they termed "pedagogical neutrality", a style that avoids marked registers and maintains a consistently formal tone. Similarly, Muñoz-Ortiz et al. (2023) identified distinctive patterns in LLM-generated news text, including higher frequencies of generic transitions and reduced lexical diversity compared to human-authored journalism.

These observations align with the technical characteristics of autoregressive language models, which generate text probabilistically and therefore favour high-frequency constructions. The fundamental architecture of these models, predicting the next token based on preceding context, naturally produces text that gravitates toward statistically common patterns. This creates what some researchers have termed a "regression to the mean" effect, wherein the model's output converges on prototypical rather than marked linguistic choices (Holtzman et al., 2020).

Several studies have examined the syntactic properties of AI-generated text. Shibata et al. (2023) compared the syntactic complexity of texts generated by GPT-3 with human-authored texts, finding that AI-generated prose exhibited lower syntactic variation and a tendency toward simpler clause structures. However, they noted that this pattern varied across model versions, with more recent models showing increased syntactic sophistication. Similarly, Guo et al. (2023) analysed the discourse structure of AI-generated academic writing, observing that while local coherence was generally high, global rhetorical organisation often lacked the strategic variation characteristic of expert human writing.

Nevertheless, much of this literature remains preoccupied with detection accuracy rather than with the linguistic systems underlying observed patterns. There is consequently a need for theoretically grounded description that situates AI-generated text within established frameworks of register and genre analysis. As Biber and Conrad (2009) argue, register analysis requires attention to the systematic relationship between linguistic features and situational contexts, a dimension largely absent from existing AI detection research.

The theoretical implications of AI-generated text extend beyond descriptive linguistics. Debates about linguistic creativity, authorship, and the nature of language production itself are implicated in the study of machine-generated prose. If LLMs can produce text that is functionally indistinguishable from human writing in certain contexts, this raises fundamental questions about the relationship between linguistic competence and linguistic performance, between the abstract system of language and its situated use.

Academic Register and Systemic Functional Linguistics

Systemic Functional Linguistics, developed principally by Halliday (1978, 1994), treats language as a meaning-making

resource shaped by context. Within this framework, register is understood as the configuration of linguistic features associated with a particular situational context, described along three metafunctional dimensions: field (what is happening), tenor (who is participating), and mode (the role of language). These three dimensions, collectively termed the "context of situation", jointly determine the register of a text, which in turn constrains the linguistic choices available to speakers or writers.

The field of a text concerns its subject matter and the nature of the social activity being undertaken. In academic writing, field is characterised by the construction and transmission of specialised knowledge, typically involving technical vocabulary, abstract concepts, and the elaboration of theoretical frameworks. The tenor of academic prose involves the relationship between writer and reader, characteristically marked by formality, distance, and a particular configuration of epistemic authority and humility. The mode of academic writing is typically written, monologic, and carefully edited, allowing for complex syntactic structures and elaborate cohesion that would be difficult to sustain in speech.

Academic writing constitutes a well-documented register characterised by high lexical density, nominalisation, technical vocabulary, and complex clause structures (Halliday & Martin, 1993; Biber, 2006). Halliday and Martin's influential work on scientific discourse demonstrated how the development of modern science was accompanied by the emergence of a distinctive linguistic register, one that privileges nominalisation as a means of packaging complex processes into manageable entities, thereby facilitating the construction of technical taxonomies and the elaboration of theoretical arguments.

Biber's (1988, 2006) multidimensional approach to register variation has further demonstrated that academic prose occupies a distinctive position along dimensions of informational density and abstractness. Through factor analysis of a wide range of linguistic features across multiple registers, Biber identified dimensions along which registers systematically vary. Academic prose, along with official documents and technical reports, occupies the end of the dimension concerned with informational density and abstract content, a position achieved through features such as nominalisation, complex noun phrases, and reduced use of interactive features such as contractions and first-person pronouns.

However, these descriptions are based on corpora of human-authored texts. Whether AI-generated academic prose conforms to, exaggerates, or departs from these established patterns remains an open empirical question. The theoretical framework of SFL predicts that if AI models have not participated in the social processes that give rise to human academic registers, their outputs may exhibit systematic differences from human-authored texts, differences that should be identifiable through careful linguistic analysis.

The concept of "registerial hybridity" is relevant here. As Matthiessen (2015) has argued, registers are not fixed categories but dynamic configurations that evolve in response to changing communicative needs. The emergence of AI-generated text may represent a new development in this ongoing evolution, a novel configuration of linguistic features that, while drawing on existing registers, constitutes a distinctive new variety.

Disciplinary Variation in Academic Writing

A substantial body of literature documents disciplinary differences in academic discourse. Hyland (2004, 2005) has shown that hedging, boosting, and self-mention vary systematically across disciplines, reflecting different epistemological assumptions and community norms. In the hard sciences, where knowledge claims are presented as objective and verifiable, hedging tends to be relatively infrequent and is often limited to specific rhetorical functions such as acknowledging methodological limitations. In contrast, the soft sciences and humanities exhibit higher frequencies of hedging, reflecting the interpretative nature of knowledge claims and the importance of acknowledging alternative perspectives.

Hyland's work on metadiscourse has been particularly influential in demonstrating how academic writers use linguistic resources to engage with readers, signal stance, and organise discourse. His analyses reveal that disciplinary communities differ not only in the frequency of metadiscoursal features but also in their functions and distribution across different parts of texts. These findings underscore the importance of considering disciplinary context in any analysis of academic register.

Similarly, Becher and Trowler (2001) describe disciplinary "tribes" with distinct communicative cultures. Their sociological analysis of academic disciplines identifies characteristic patterns of interaction, publication, and knowledge production that vary systematically across fields. These patterns are reflected in distinctive linguistic practices, what they term "disciplinary discourses" that serve to mark membership in particular academic communities.

More recent work has extended these analyses to consider the role of genre and rhetorical structure in disciplinary variation. Swales (1990, 2004) has documented how the academic abstract, in particular, exhibits systematic variation across disciplines in terms of its rhetorical moves and linguistic realisations. His analyses of research article introductions, abstract structures, and other academic genres have provided detailed accounts of how disciplinary communities construct and communicate knowledge through conventionalised textual forms.

The abstract, as a genre, is particularly interesting for comparative analysis because it condenses the essential features of academic writing into a relatively brief text. An abstract must convey the research aims, methods, findings,

and significance of a study while conforming to the conventions of its disciplinary community. As such, abstracts provide a concentrated sample of disciplinary linguistic practices, making them ideal for comparative corpus analysis.

These findings are pertinent to the present study because they raise the question of whether AI models, trained on multidisciplinary corpora, reproduce such variation or flatten it into a generalised academic style. Given that LLMs are trained on vast corpora encompassing multiple disciplines and registers, there is reason to suspect that they may produce text that approximates a "lowest common denominator" of academic style, one that draws on features common to many disciplines while under-representing the distinctive features that mark disciplinary membership.

Theoretical Perspectives on Machine-Generated Language

Beyond descriptive studies of AI-generated text, several theoretical perspectives offer insights into the linguistic character of machine-generated prose. The concept of "artificial intelligence" itself invites reflection on what it means to produce language, to communicate meaning, and to participate in discourse communities.

Bender et al. (2021) have argued that LLMs, despite their impressive performance on many tasks, fundamentally lack understanding of meaning. They term these models "stochastic parrots", systems that reproduce statistical patterns in their training data without any genuine comprehension of what the words mean. This perspective has important implications for understanding AI-generated text: if models lack semantic understanding, their outputs may exhibit subtle but systematic differences from human writing in terms of coherence, appropriateness, and contextual sensitivity.

However, other researchers have argued that the distinction between statistical pattern matching and genuine understanding may be less clear-cut than Bender et al. suggest. The distributional hypothesis in linguistics, the idea that words with similar meanings appear in similar contexts, implies that statistical patterns in language use encode semantic information. If this is correct, then LLMs may be capturing meaningful regularities in language even without explicit semantic representations (Boleda, 2020; Lenci, 2008).

Regardless of how this theoretical debate is resolved, the practical question remains: do AI-generated texts differ systematically from human-written texts in identifiable ways? And if so, what do these differences reveal about the nature of language, register, and discourse?

Research Gap

While detection studies and ethical commentaries abound, few studies have applied a systematic linguistic framework to compare AI-generated and human academic prose across disciplines. The existing literature on AI-generated text tends to focus on either detection (identifying whether a text was produced by a machine) or evaluation (assessing the quality of AI-generated outputs). What is largely absent is a descriptive analysis grounded in linguistic theory that characterises AI-generated text as a register in its own right.

The present study addresses this gap by combining corpus-based quantitative analysis with qualitative register analysis, thereby contributing both empirical findings and a descriptive framework. By situating AI-generated text within the theoretical framework of Systemic Functional Linguistics, the study aims to provide a principled account of how machine-generated prose differs from human-authored prose, and why these differences matter for our understanding of academic discourse.

Theoretical Framework

This study is anchored in Systemic Functional Linguistics (SFL**)**, specifically Halliday's metafunctional theory and the associated concept of register. SFL provides a principled account of how linguistic choices are motivated by contextual variables, making it particularly suited to comparing texts produced under different conditions of authorship.

The choice of SFL as the theoretical framework reflects its capacity to connect linguistic description with social context. Unlike formalist approaches that treat language as an autonomous system, SFL treats language as a resource for meaning-making—a system whose organisation is shaped by the functions it serves. This perspective is essential for understanding register variation, as it explains why different contexts of situation give rise to different configurations of linguistic features.

Metafunctional Theory

Halliday's metafunctional theory posits that language is organised around three simultaneous meanings: ideational, interpersonal, and textual. The ideational metafunction concerns the representation of experience, how language construes the world of entities, events, and relationships. The interpersonal metafunction concerns the enactment of social relationships, how language positions speakers and hearers in relation to each other. The textual metafunction concerns the organisation of discourse, how language creates coherence and texture.

Each metafunction is associated with particular systems of linguistic choice. The ideational metafunction is realised through transitivity structures (processes, participants, circumstances), the interpersonal metafunction through mood and modality, and the textual metafunction through theme and information structure. Register variation can be understood as systematic variation in the probabilities of choice within these systems, conditioned by features of the context of situation.

For the present study, the interpersonal metafunction is particularly relevant, as it encompasses hedging and modality, features that are expected to differ between AI-generated and human-written text. Hedging, in SFL terms, is a realisation of the interpersonal metafunction that positions the writer in relation to the truth value of propositions and the relationships with readers. The reduced hedging observed in AI-generated texts, if confirmed, would represent a shift in the interpersonal configuration of academic prose, a change in the tenor of the register.

Register Theory

Within this framework, register is understood as the configuration of linguistic features associated with a particular situational context, described along three metafunctional dimensions: field (what is happening), tenor (who is participating), and mode (the role of language). These dimensions jointly determine the register of a text, which can be defined as the variety of language associated with a particular configuration of field, tenor, and mode.

Academic writing constitutes a well-documented register characterised by high lexical density, nominalisation, technical vocabulary, and complex clause structures (Halliday & Martin, 1993; Biber, 2006). Within this framework, the field of academic abstracts concerns the presentation of research aims, methods, and findings; the tenor involves relations between authors and a specialised readership, typically marked by hedging and evaluative restraint; and the mode is written, monologic, and highly edited. Register theory predicts that these contextual features will be realised through specific linguistic choices: nominalisation, technical lexis, complex noun phrases, and modulated modality.

Explanatory Potential of SFL

The analytical value of SFL here lies in its capacity to explain *why* AI-generated texts might differ from human texts. If an LLM has not participated in the disciplinary socialisation that shapes human authors' tenor choices, its realisations of hedging and evaluation may be systematically different, not merely random. The model has not experienced the interpersonal negotiations, the peer review processes, the conference discussions, and the informal communications that shape human scholars' understanding of what claims can be made with what degree of confidence. Consequently, the model's output may lack the fine-grained interpersonal calibration that characterises expert human writing.

SFL thus supplies both descriptive categories and explanatory logic. It provides the conceptual vocabulary for describing how AI-generated texts differ from human texts (in terms of ideational, interpersonal, and textual meanings) and the theoretical rationale for understanding why these differences arise (because the model lacks the contextual experience that shapes human registerial choices).

The Concept of Registerial Reproduction

The study also engages with the concept of "registerial reproduction", the process by which a text reproduces features of existing registers without participating in the social practices that sustain them. An AI-generated academic abstract, on this view, is not a genuine instance of the academic register but a reproduction of its surface features, a kind of linguistic simulation that may approximate but not fully instantiate the register.

This concept has important implications for understanding AI-generated text. If AI models reproduce the surface features of academic writing without embodying the social practices that underlie the register, then their outputs may exhibit systematic differences from human writing in terms of the subtle, context-dependent features that characterise expert performance. Identifying these differences is a central task of the present study.

Methodology

Corpus Design

The study employed a purpose-built corpus comprising 120 academic abstracts of approximately 200 words each:

Sub-corpus	Linguistics	Literary Studies	Natural Sciences	Total
AI-generated	20	20	20	60
Human-written	20	20	20	60
Total	**40**	**40**	**40**	**120**

AI-generated abstracts were produced using a single LLM (GPT-4) with standardised prompts requesting an abstract on a specified topic in each discipline. The prompts specified the discipline and provided a brief topic description, but did not include specific instructions about style, length, or structure, in order to allow the model to produce text that reflected its default tendencies. All generations were conducted in a single session to control for temporal variation in model behaviour.

Human-written abstracts were sampled from peer-reviewed journals published between 2020 and 2024. The selection criteria were as follows: (i) the abstract was published in a leading journal in its field; (ii) the abstract was approximately 200 words in length; (iii) the abstract followed a standard structure (background, aims, methods, findings, implications). Abstracts were randomly selected from the available pool meeting these criteria (Okolie, 2025). For

linguistics, journals included *Language*, *Applied Linguistics*, and *Journal of Pragmatics*. For literary studies, journals included *Critical Inquiry*, *New Literary History*, and *Modern Language Quarterly*. For natural sciences, journals included *Nature*, *Science*, and *Proceedings of the National Academy of Sciences*.

The disciplines were selected to represent a range of epistemological orientations, from the interpretative emphasis of literary studies to the empirical focus of the natural sciences, allowing examination of whether disciplinary differences are reproduced or flattened in AI-generated text.

Analytical Procedures

Quantitative analysis measured four variables: (i) lexical density (content words per clause); (ii) nominalisation density (nominalisations per 100 words); (iii) type-token ratio (TTR) as a measure of lexical diversity; and (iv) modality distribution (frequency of modal verbs and hedges per 100 words). These variables were chosen because they are well-established indicators of registerial features in academic writing and because they represent the three metafunctional dimensions central to the theoretical framework.

Lexical density was calculated by dividing the number of content words (nouns, verbs, adjectives, adverbs) by the number of clauses in each text. This measure captures the informational density of the text, a key feature of the academic register according to previous research (Halliday & Martin, 1993).

Nominalisation density was calculated by identifying nominalised forms (nouns derived from verbs or adjectives) and counting their frequency per 100 words. A comprehensive list of nominalising suffixes (-tion, -ment, -ness, -ity, -ance, -ence, -ure, -al) was used to identify nominalisations, with manual verification to exclude false positives.

Type-token ratio (TTR) was calculated as the number of unique words divided by the total number of words in each text. This measure captures lexical diversity, with lower values indicating greater repetition and higher values indicating greater variety.

Modality distribution was analysed by identifying modal verbs (can, could, may, might, must, shall, should, will, would) and hedging devices (epistemic adverbs such as possibly, probably, arguably; epistemic phrases such as it seems that, it appears that; epistemic adjectives such as possible, likely, uncertain). Frequencies were calculated per 100 words.

Qualitative analysis examined thematic progression and hedging strategies in a subsample of twelve texts (two per sub-corpus per discipline) (Siame & Lubbungu, 2023; Siame et al., 2023). Thematic progression was analysed following the framework of Danes (1974), identifying patterns such as constant theme progression (where the theme of one clause becomes the theme of the next), linear progression (where the rheme of one clause becomes the theme of the next), and derived theme progression (where themes are derived from a "hypertheme").

Hedging strategies were analysed using Hyland's (1998) taxonomy, which distinguishes between content-oriented hedges (relating to the truth value of propositions) and reader-oriented hedges (relating to the relationship between writer and reader). This analysis aimed to identify qualitative differences in hedging strategies that might not be captured by frequency counts alone.

Statistical Analysis

Quantitative data were analysed using descriptive statistics (means and standard deviations) and inferential statistics (two-way ANOVA with text type and discipline as factors). Post-hoc comparisons were conducted using Tukey's HSD test. Effect sizes were calculated using Cohen's d. Statistical significance was set at $p < 0.05$.

The two-way ANOVA design allowed examination of both main effects (text type and discipline) and interaction effects (whether the difference between AI-generated and human-written texts varied across disciplines). This is important for the second research question, which concerns how disciplinary context modulates the observed differences.

Ethical Considerations

The study used publicly available published abstracts and AI-generated texts created for research purposes. No human participants were involved, and no personal data were processed. The use of published abstracts complies with fair use provisions for research purposes. AI-generated texts were created specifically for this study and do not reproduce copyrighted material.

Findings and Discussion

Quantitative Findings

Table 1 presents mean values across sub-corpora.

Table 1: Mean linguistic values by sub-corpus

Measure	AI (Ling.)	Human (Ling.)	AI (Lit.)	Human (Lit.)	AI (Sci.)	Human (Sci.)
Lexical density	6.8	6.2	6.5	5.9	7.4	7.1
Nominalisation density	14.2	10.8	12.9	9.4	16.1	13.7
TTR	0.42	0.51	0.44	0.54	0.39	0.47
Hedges per 100 words	2.1	3.8	1.9	4.2	1.4	2.9

Three patterns are notable in this dataset. First, AI-generated texts exhibit consistently higher nominalisation density and lower lexical diversity than human texts across all disciplines. The mean nominalisation density for AI texts ranged from 12.9 (literary studies) to 16.1 (natural sciences), compared with 9.4 to 13.7 for human texts. This difference was statistically significant ($F(1, 114) = 28.4$, $p < 0.001$) and consistent across disciplines (no significant interaction effect). The effect size was large (Cohen's $d = 0.89$), indicating a substantial difference between the two text types.

Second, the hedging gap is most pronounced in literary studies, where human authors hedge substantially more than AI. Human literary abstracts contained a mean of 4.2 hedges per 100 words, compared with 1.9 for AI-generated literary abstracts, a difference of 2.3 hedges per 100 words. In linguistics, the gap was 1.7 hedges per 100 words (3.8 vs. 2.1), and in the natural sciences, it was 1.5 hedges per 100 words (2.9 vs. 1.4). The interaction between text type and discipline was significant ($F(2, 114) = 4.7$, $p < 0.05$), indicating that the magnitude of the hedging gap varied across disciplines.

Third, lexical density is higher in AI texts, suggesting a compression of information that exceeds even the already dense norms of academic prose. The mean lexical density for AI texts ranged from 6.5 (literary studies) to 7.4 (natural sciences), compared with 5.9 to 7.1 for human texts. This difference was significant ($F(1, 114) = 12.8$, $p < 0.01$), though the effect size was moderate (Cohen's $d = 0.52$).

The type-token ratio (TTR) results show that AI-generated texts consistently exhibit lower lexical diversity than human texts. The mean TTR for AI texts ranged from 0.39 (natural sciences) to 0.44 (literary studies), compared with 0.47 to 0.54 for human texts. This difference was significant ($F(1, 114) = 35.2$, $p < 0.001$) and consistent across disciplines. The lower TTR in AI texts indicates greater repetition of vocabulary, a pattern consistent with the probabilistic tendencies of autoregressive models, which tend to favour high-frequency words and phrases.

Qualitative Findings: Hedging and Thematic Progression

Close analysis revealed that human authors employed discipline-specific hedging strategies, whereas AI texts favoured a narrower repertoire. The following excerpts illustrate this contrast.

Excerpt 1 (Human, Literary Studies):

> *It might be argued that the novel's temporal structure somewhat complicates prevailing accounts of narrative closure, although this reading remains open to contestation.*

Excerpt 2 (AI, Literary Studies):

> *The novel's temporal structure complicates prevailing accounts of narrative closure, demonstrating the significance of narrative experimentation.*

The human excerpt deploys multiple hedges such as "might," "somewhat," "although," and "open to contestation", realising the tenor of scholarly tentativeness. The AI excerpt asserts its claim more categorically, with only one evaluative item, "demonstrating". This pattern was consistent across the subsample: human texts exhibited a wider range of hedging devices, including modal verbs, epistemic adverbs, and metadiscursive hedging phrases, while AI texts tended to rely on a limited set of hedges, primarily modal verbs such as "may" and "can."

Further analysis revealed that human authors used hedging not only to qualify claims but also to position themselves in relation to disciplinary debates and to signal membership in interpretative communities. For example, the phrase "open to contestation" in Excerpt 1 does not merely hedge the claim; it invokes a disciplinary norm according to which interpretations are always provisional and subject to challenge. AI-generated texts, by contrast, tended to use hedging primarily as a syntactic softener rather than as a marker of disciplinary identity.

Excerpt 3 (Human, Linguistics):

> *The findings suggest that input frequency may play a role in the acquisition of grammatical constructions, though the correlational design limits any causal interpretation.*

Excerpt 4 (AI, Linguistics):

> *The findings demonstrate that input frequency plays a crucial role in the acquisition of grammatical constructions, with significant implications for language pedagogy.*

Again, the human excerpt employs multiple hedging strategies: the epistemic verb "suggest," the modal "may," the concessive "though," and the explicit acknowledgement of methodological limitations. The AI excerpt presents its findings more assertively, with the strong claim that input frequency "plays a crucial role" and "significant implications." While this excerpt does contain some hedging-like features, the term "significant" could be interpreted as a statistical rather than substantive claim; the overall tenor is more confident than the human excerpt.

Excerpt 5 (Human, Natural Sciences):

> *These results appear to indicate that the protein may function as a regulator of cell proliferation, although further experiments will be required to confirm this hypothesis.*

Excerpt 6 (AI, Natural Sciences):

> *These results demonstrate that the protein functions as a regulator of cell proliferation, providing new insights into the mechanisms of cell cycle control.*

The human excerpt is notably more hedged than might be expected for a natural science abstract, with three distinct hedging devices ("appear," "may," "although") and an explicit

statement of future research needs. The AI excerpt makes a more direct claim, with no explicit hedging. Interestingly, the human excerpt's hedging is not necessarily typical of all natural science abstracts; it may reflect the particular stage of the research or the specific journal's conventions. This highlights the importance of considering disciplinary sub-genres and rhetorical contexts when analysing hedging patterns.

Thematic progression analysis further showed that AI texts tended to maintain constant thematic chains, whereas human texts exhibited more varied progression patterns, including derived and split themes. This suggests that AI-generated prose, while coherent, is rhetorically more uniform.

For example, in a typical AI-generated abstract, the theme of each sentence tended to repeat or closely paraphrase the theme of the preceding sentence:

> *The study examines the relationship between X and Y. The study uses a mixed-methods approach. The findings indicate a significant relationship. The findings have implications for practice.*

In contrast, human-authored abstracts exhibited more varied progression patterns:

> *Recent research has highlighted the importance of X in Y contexts. However, existing studies have tended to neglect the role of Z. To address this gap, the present study employs a mixed-methods approach. Our findings suggest that Z plays a crucial role in mediating the relationship between X and Y.*

Here, the theme of the second sentence "existing studies" is derived from the theme of the first sentence, "recent research", while the theme of the third sentence "the present study" introduces a new but related element. This kind of thematic variation creates a more dynamic and rhetorically effective flow of information.

Discussion

These findings support the argument that AI-generated academic English constitutes a recognisable register rather than a neutral imitation. Its heightened nominalisation and reduced lexical diversity reflect the probabilistic tendencies of autoregressive models, which favour high-frequency constructions and compact informational packaging. The reduced hedging in AI texts is particularly consequential: hedging is not merely stylistic but encodes the epistemic stance of a disciplinary community.

Nominalisation and Informational Compression

The higher nominalisation density observed in AI-generated texts aligns with the technical characteristics of autoregressive language models. As noted above, these models generate text probabilistically, favouring high-frequency constructions. Nominalisation, as a grammatical process that packages processes and qualities into entities, is one of the most productive resources for creating informationally dense prose, a feature that characterises academic writing across disciplines (Halliday & Martin, 1993; Biber, 2006).

However, the present findings suggest that AI models may over-apply this resource, producing nominalisation densities that exceed even the already dense norms of human academic prose. This may reflect the models' tendency to approximate the "prototype" of academic writing, a style that is maximally informational, rather than reproducing the more varied and contextually calibrated uses of nominalisation found in human writing.

This interpretation is supported by the work of Biber (2006), who argued that different registers vary systematically in their use of nominalisation and other informational features. Academic writing is generally high in these features, but there is considerable internal variation; human authors calibrate their use of nominalisation to the specific demands of their rhetorical situation, the expectations of their readers, and their own stylistic preferences. AI models, lacking this contextual calibration, may default to a high-nominalisation style as a kind of "average" of the academic corpus.

The implications of this finding extend beyond descriptive linguistics. If AI-generated text over-applies nominalisation, this may affect its accessibility and readability. High nominalisation density is associated with increased processing difficulty (Halliday & Martin, 1993), and texts that exceed human norms in this regard may be more challenging for readers, even expert readers, to process. This has practical implications for the use of AI in academic writing, suggesting that AI-generated drafts may require substantial editing to achieve appropriate readability.

Lexical Diversity and Repetition

The lower lexical diversity observed in AI-generated texts is consistent with the probabilistic nature of autoregressive models. These models generate text by predicting the most likely next word given the preceding context. This process naturally favours high-frequency words, which have higher probabilities of occurrence. As a result, AI-generated texts tend to exhibit greater repetition of vocabulary than human texts, where authors actively vary their word choices to avoid repetition and to create stylistic effects.

This finding aligns with the observations of Herbold et al. (2023) and Muñoz-Ortiz et al. (2023), who noted tendencies toward lexical repetition and formulaic transitions in AI-generated prose. It also connects to broader theoretical debates about the nature of linguistic creativity. If AI models are fundamentally pattern-matching systems that reproduce statistical regularities in their training data, then their outputs will inevitably be less lexically varied than human writing, which is shaped by intentional choices about word selection, rhetorical effect, and stylistic variation.

However, the reduced lexical diversity in AI-generated texts is not merely a quantitative difference; it has qualitative implications for the texture of the prose. Human academic writing is characterised not only by higher lexical diversity but also by strategic variation, the deliberate selection of words to achieve particular effects, to mark emphasis, to create cohesion, or to signal disciplinary membership. AI-generated texts, with their more uniform vocabulary, may lack this strategic dimension, producing prose that is coherent but stylistically flat.

Hedging and Interpersonal Meaning

The reduced hedging in AI-generated texts is arguably the most significant finding of the study. Hedging is a key resource for realising the interpersonal metafunction in academic writing; it encodes the writer's stance toward the truth value of claims, acknowledges alternative perspectives, and signals appropriate epistemic modesty (Hyland, 1998, 2005).

The disciplinary variation in hedging observed in the present study, with human literary studies abstracts showing the highest frequencies, followed by linguistics and natural sciences, aligns with previous research. Hyland (2004) documented similar patterns, attributing them to differences in epistemological assumptions: interpretative disciplines, where knowledge claims are inherently provisional and contested, exhibit more hedging than empirical disciplines, where claims are presented as objective and verifiable.

The AI-generated texts, by contrast, exhibited consistently lower hedging across all disciplines, with the gap being widest in literary studies. This suggests that AI models may under-represent the interpersonal variability that distinguishes disciplines. Rather than calibrating hedging to the expectations of specific disciplinary communities, AI models appear to produce a more uniform, less hedged style, one that may reflect an "average" of academic writing across disciplines.

This finding has important implications for the sociology of academic communication. Hedging is not merely a stylistic feature; it is a marker of disciplinary membership and a resource for enacting the social relations that characterise academic communities. The attenuation of hedging in AI-generated text may signal a shift in the tenor of academic discourse, a shift toward more categorical, less dialogic modes of claim-making.

If AI-generated text becomes more prevalent in academic publishing, this could have cascading effects on disciplinary norms. Readers may become accustomed to less hedged prose, and authors may feel pressure to adopt more assertive styles to compete. Alternatively, the distinctiveness of human hedging practices may become a marker of human authorship, a linguistic signal that distinguishes human from machine writing.

The finding that the hedging gap was widest in literary studies is particularly significant. Literary studies is a discipline in which interpretative humility is central to the rhetorical economy; claims are expected to be presented as provisional, open to contestation, and situated within ongoing scholarly debates. The AI-generated literary abstracts in this study exhibited a more assertive style, presenting interpretations as if they were established findings rather than contestable readings. This represents a fundamental misalignment with the epistemological norms of the discipline, a kind of rhetorical "category error" that may signal the limitations of AI models' understanding of disciplinary contexts.

Thematic Progression and Discourse Structure

The qualitative analysis of thematic progression revealed further differences between AI-generated and human-written texts. Human authors exhibited more varied progression patterns, including linear, derived, and split themes, creating a more dynamic and rhetorically effective flow of information. AI-generated texts, by contrast, tended to maintain constant thematic chains, producing prose that was coherent but rhetorically uniform.

This finding aligns with the work of Guo et al. (2023), who observed that while AI-generated academic writing achieves high local coherence, it often lacks the global rhetorical organisation characteristic of expert human writing. Thematic progression is one dimension of this global organisation: it reflects the writer's strategic decisions about how to sequence information, how to create emphasis, and how to guide the reader through the argument.

The uniformity of thematic progression in AI-generated texts may reflect the models' reliance on local coherence strategies, techniques that ensure each sentence connects smoothly to the previous one, rather than global rhetorical planning. Human expert writers, by contrast, typically begin with a global rhetorical plan and then execute it through local linguistic choices, creating thematic patterns that serve the overall communicative purpose of the text.

AI-Generated Text as a Distinct Register

Taken together, these findings support the conclusion that AI-generated academic English constitutes a recognisable register rather than a neutral imitation of human academic style. The distinctive configuration of features observed in AI-generated texts- higher nominalisation density, lower lexical diversity, reduced hedging, and more uniform thematic progression- represents a novel configuration of linguistic resources that does not simply reproduce any existing human register.

This conclusion has theoretical implications for register theory. If AI-generated text constitutes a distinct register, then the descriptive apparatus of register analysis must be extended to accommodate machine-authored texts as a

legitimate object of analysis. The concept of register, as developed in SFL and related frameworks, was formulated with reference to human language use, language produced by speakers and writers who participate in social practices and are shaped by their experiences as members of discourse communities. The emergence of AI-generated text challenges us to consider whether and how these concepts can be extended to machine-generated language.

One possibility is that AI-generated text represents a kind of "registerial hybrid", a configuration of features that draws on multiple existing registers but does not fully instantiate any of them. On this view, AI models synthesise features from their training data, producing text that approximates but does not reproduce the registers they are imitating. The resulting configuration is novel because it reflects the statistical properties of the training corpus rather than the situated choices of human authors.

Another possibility is that AI-generated text represents a new register in its own right, one that is emerging as AI systems become more prevalent in academic and professional contexts. Just as new registers have emerged historically in response to new communicative technologies (e.g., the scientific register in the seventeenth century, the digital register in the late twentieth century), the AI-generated register may represent the latest stage in the evolution of academic discourse.

Implications for Academic Integrity

The findings of this study have practical implications for academic integrity. While much current effort focuses on developing detection tools, algorithms that can identify whether a text was produced by AI, the present study suggests that linguistic description may offer a complementary approach. By identifying the distinctive features of AI-generated text, it may be possible to develop more robust and interpretable criteria for distinguishing human from machine writing.

However, the findings also caution against simplistic approaches to detection. The differences between AI-generated and human texts, while statistically significant, are not absolute. There is considerable overlap between the two categories, and some AI-generated texts may exhibit features that are more typically associated with human writing (and vice versa). Moreover, as AI models continue to evolve, their outputs may become increasingly indistinguishable from human writing, a development that would challenge detection approaches based on surface linguistic features.

This underscores the importance of complementing detection approaches with a deeper understanding of the linguistic and rhetorical characteristics of AI-generated text. Rather than simply asking "Is this text AI-generated?" we might ask "What kind of text is this? What register does it instantiate? What rhetorical purposes does it serve?" These questions require the kind of theoretically grounded description that the present study aims to provide.

Implications for Writing Pedagogy

The findings also have implications for writing pedagogy. The distinctive features of AI-generated text- higher nominalisation, lower lexical diversity, reduced hedging highlight the value of interpersonal and discipline-specific rhetorical strategies that AI models under-represent. These are precisely the features that expert human writers deploy to create effective academic prose: calibrating claims to the expectations of disciplinary communities, acknowledging uncertainty and alternative perspectives, and creating varied and engaging discourse structures.

If AI tools are used in writing instruction, it is important that students learn not only to use these tools effectively but also to recognise their limitations. AI-generated text can be a useful starting point for drafting, but it typically requires substantial revision to achieve the rhetorical sophistication of expert human writing. Students need to develop the critical awareness to identify where AI-generated text falls short and the linguistic resources to revise it appropriately.

The findings also suggest that writing pedagogy should continue to emphasise the interpersonal dimension of academic writing, how writers position themselves in relation to their readers, their disciplines, and their knowledge claims. These are precisely the dimensions that AI models struggle to reproduce, and they are essential for effective scholarly communication.

Limitations and Future Research

Several limitations of the study should be acknowledged. First, the corpus is relatively small (120 abstracts) and limited to a single genre (the academic abstract). Future research should extend the analysis to other genres (e.g., research articles, conference presentations, grant proposals) and larger corpora to assess the generalisability of the findings.

Second, the study used a single LLM (GPT-4) and a single set of prompts. Different models, different prompting strategies, and different generation parameters (e.g., temperature, top-p) might produce different linguistic outputs. Future research should compare the linguistic characteristics of text generated by different models and under different conditions.

Third, the study focused on a limited set of quantitative variables (lexical density, nominalisation density, TTR, hedging frequency) and qualitative analyses (thematic progression, hedging strategies). Other linguistic features, such as syntactic complexity, cohesion, stance, and engagement, might reveal further differences between AI-generated and human-written texts.

Fourth, the human-written abstracts were sampled from published journals, while the AI-generated abstracts were

produced in response to prompts. This introduces a potential confound: the human abstracts may have undergone editorial revision and may reflect the conventions of specific journals, while the AI-generated abstracts reflect the model's default output. Future research should attempt to control for these factors by comparing AI-generated and human-written texts produced under more comparable conditions.

Finally, the study is cross-sectional, examining texts produced at a single point in time. As AI models continue to evolve, their linguistic outputs may change, potentially becoming more similar to human writing (or diverging in new ways). Longitudinal research is needed to track these changes and to assess their implications for our understanding of AI-generated text as a register.

Conclusion

This study has demonstrated that AI-generated academic English differs systematically from human-written prose in nominalisation density, lexical diversity, and hedging behaviour, and that these differences are modulated by disciplinary context. Framed within SFL and register theory, the findings support the conclusion that AI-generated text constitutes a distinct register, one characterised by informational compression and reduced interpersonal modulation.

The theoretical contributions of the study are threefold. First, it extends register theory to encompass machine-generated text as a legitimate object of analysis, demonstrating that the descriptive apparatus of SFL can be productively applied to AI-generated prose. Second, it proposes a framework for identifying register markers in machine-generated academic discourse, a framework that identifies nominalisation density, lexical diversity, hedging frequency, and thematic progression as key dimensions along which AI-generated and human-written texts differ. Third, it contributes to theoretical debates about the nature of register, suggesting that AI-generated text may represent a novel configuration of linguistic resources rather than a simple imitation of existing registers.

The practical implications are equally significant. For academic integrity, linguistic description offers a complementary approach to detection, providing interpretable criteria for distinguishing human from machine writing. For writing pedagogy, the findings highlight the value of interpersonal and discipline-specific rhetorical strategies that AI models under-represent, suggesting that these should remain central to writing instruction. For linguistic theory, the study suggests that register descriptions must now account for machine-authored texts as a legitimate object of analysis.

The broader significance of the study lies in its contribution to understanding the linguistic character of AI-generated text, an increasingly important dimension of contemporary academic communication. As AI tools become more embedded in scholarly workflows, understanding their linguistic properties becomes essential for researchers, educators, and administrators alike. The present study provides a theoretically grounded, empirically rigorous foundation for this understanding, and it is hoped that it will stimulate further research in this rapidly evolving field.

Future research should extend the corpus to other genres, languages, and models, and should examine longitudinal change as AI systems evolve. Comparative studies across different LLMs would help to establish whether the patterns observed here are specific to GPT-4 or reflect more general tendencies of autoregressive language models. Research on AI-generated text in languages other than English would help to assess the cross-linguistic generalisability of the findings. And longitudinal research tracking changes in AI output over time would help to understand how the AI-generated register is evolving and what implications these changes have for academic communication.

The study also raises philosophical questions about the nature of authorship and linguistic creativity in the age of AI. If AI-generated texts constitute a distinct register with identifiable linguistic features, what does this mean for our understanding of what it is to write, to author, and to communicate? These questions are beyond the scope of the present study, but they are questions that linguists, philosophers, and scholars across the disciplines will need to address as AI systems become increasingly sophisticated and pervasive in scholarly communication.

References

1. Becher, T., & Trowler, P. (2001). *Academic tribes and territories*. Buckingham: University Press.
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.
3. Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.
4. Biber, D. (2006) *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
5. Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
6. Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6, pp. 213–234.
7. Danes, F. (1974). Functional sentence perspective and the organisation of the text. In F. Danes (ed.), *Papers on Functional Sentence Perspective*. Prague: Academia, pp. 106–128.
8. 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.

9. Halliday, M. A. K. (1978). *Language as social semiotic*. London: Edward Arnold.
10. Halliday, M. A. K. (1994). *An introduction to functional grammar*. 2nd edn. London: Edward Arnold.
11. Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. London: Falmer Press.
12. 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, 18617.
13. Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text Degeneration. *Proceedings of the 8th International Conference on Learning Representations*.
14. Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.
15. Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbour: University of Michigan Press.
16. Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.
17. Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. S. (2020). Automatic detection of machine-generated text: A critical survey. *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2296–2309.
18. Kasozwe, A., Kapau, H.M., Siame, P., Kangwa, K.N., Amoakohene, B., & Mambwe, K. (2025). Recalibrating Zambia through linguistic choices: A Faircloughian microanalysis of Hakainde Hichilema's inaugural speech. *Scholars International Journal of Linguistics and Literature*, 8(2), 14-28.
19. Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), pp. 1–31.
20. Matthiessen, C. M. I. M. (2015). Register in the round: Registerial cartography. *Functional Linguistics*, 2(1), pp. 1–48.
21. Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2023). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 56, pp. 1–28.
22. Okolie, U.C., Lubbungu, J., Osuji, C.U., Idike, I.M., Ugwuda, S.O., Ehiobuche, C., & Siame, P. (2025). Career-related teacher support and decision to choose teaching as a career: A three-wave study of Nigerian and Zambian university students. *International Journal for Educational and Vocational Guidance*, 25 (2), 633-661
23. Shibata, T., Kondo, M., & Matsumoto, T. (2023). Syntactic complexity of ChatGPT-generated texts: A comparative analysis. *Journal of Natural Language Processing*, 30(2), pp. 456–478.
24. Siame, P. (2026a). Integrating AI literacy and critical discourse analysis into the 21st century University Curriculum. *International Journal of Multilingualism and Languages for Specific Purposes*, 8(01), 124-143. https://revue.univ-oran2.dz/Revue/IJMLSP/index.php/IJMLSP/article/view/146/103
25. Siame, P. (2026b). The algorithmic appeal: How AI constructs the eco-conscious consumer sociolinguistic discourses in digital retail spaces. *Global Journal of Multidisciplinary Studies*,1(5), 67-83. https://gjmshk.org/papers/the-algorithmic-appeal-how-ai-constructs-the-eco-conscious-consumer-sociolinguistic-discourses-in-digital-retail-spaces/
26. Siame, P., Amoakohene, B., & Kapau, H.M. (2023). Phonological Processes Governing Borrowing from English to Lungu, Mambwe, and Namwanga Languages. *Daengku: Journal of Humanities and Social Sciences Innovation*, 3 (5), 821-830. https://doi.org/10.35877/454RI.daengku1765
27. Siame, P. & Lubbungu, J. (2023). Decolonisation: Enhancing equity in the management of language education in Secondary Schools in Kabwe District. *Kwame Nkrumah University Multi-Disciplinary Journal - Zambia*, 1(1), 101-116. https://accessjournal.nkrumah.edu.zm/index.php/knuj/article/view/16
28. Siame, P. (2022). Potholes in the teaching of Zambian languages in secondary schools: A case of Bemba language. *EduLine: Journal of Education and Learning Innovation*, 2 (4), 548-560
29. Siame, P., Chisenga, R.P.A., Kangwa, K.N., Kapau, H.M., & Amoakohene, B. (2025). Constructing legitimacy through language: A critical discourse analysis of Zambian Parliamentary debates. *IKR Journal of Arts, Humanities and Social Sciences (IKRJAHSS)*, 1 (4), 232-243
30. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv preprint*, arXiv:1908.09203.
31. Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
32. Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.