Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Is AI truly proficient in academic writing? Evidence from a multidimensional analysis of academic Chinese

Changwei Hu
Northeast Agricultural University
Wanqin Zhou
Xiamen University
Yu Zhu
zhuyu@xmu.edu.cn

Xiamen University

Article

Keywords: Artificial intelligence, Register variation, Multidimensional analysis, Academic Chinese

Posted Date: April 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9070095/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

AI is generally regarded as having a good performance in the field of academic writing, but most studies are conducted from the perspective of general English academic writing. This study investigates the human-like of AI-generated texts by comparing them with human-authored writing across four disciplines. Utilizing Biber's multidimensional analysis, we constructed an AI–human parallel corpus of 8,000 texts, incorporating five advanced models and three prompting strategies.

Analysis across four linguistic dimensions: (1) Contextualized Situational Elaboration, (2) Loose vs. Compact Structure, (3) Modification-driven vs. Abstract Information, and (4) Explicit Epistemic Assertion —reveals systematic divergences. AI significantly overuses adverbial modifiers and fails to replicate the explicit epistemic assertions characteristic of human expertise, exhibiting a cautious, vague tone. By averaging semantic features, most models erase the sharp distinctions between disciplinary registers.

Critically, prompt engineering proves ineffective; multi-strategy prompting often exacerbates the stylistic gap by triggering simplistic academic stereotypes rather than genuine disciplinary reasoning. We conclude that AI awareness of register is confined to a generic academic tone while failing to navigate the intricate semantic structures of specific disciplines. These findings present patterns both similar to and distinct from those observed in academic English, offering new corpus-based evidence for AI–human comparison.

Introduction

Since the surge in AI advancements, particularly with Large Language Models (LLMs) like ChatGPT at the close of 2022, there has been a growing discourse surrounding the potential of AI-generated text to supplant human-authored content (Barrot,2023; Dwivedi et al., 2023; Shamsi et al., 2025). Within academic circles, AI has now become an integral component of the researcher's workflow (Batista et al., 2024; Freeman, 2025; Van Noorden & Perkel, 2023). The deep entrenchment of AI in research has alarmed publishers over academic integrity, fueling the rise of commercial detection tools. (Liang et al., 2024; Liu et al., 2025).

However, the current prevailing view suggests that AI can partially replace humans in performing language tasks, especially in some English-centric studies (Dilai & Dilai, 2025; Nguyen & Barrot, 2024; Shamsi et al., 2025). Although AI-generated and human-written texts differ in their expression of stance (Zhang & Zhang, 2025). In the Chinese context, current detection tools highlight the unreliable nature of AI-generated academic texts (Guo et al., 2025; Pratama, 2025). The limitations revealed by current tools highlight the pressing necessity of analyzing differences between AI-generated and human-authored academic texts in languages other than English.

The ability of AI-generated text to accurately represent register variation remains an open question that requires further investigation (Reinhart et al., 2025; Sardinha, 2024). While most studies concentrate on the detectability of AI content, they frequently neglect the mediating role of register (Doughman et al.,

2025) and are primarily confined to English-language corpora. Given the extremely low proportion of Chinese-language data in the training sets of LLMs the materials they produce are likely to exhibit significant bias (Guo et al., 2025). Consequently, it is essential to evaluate the performance of various models. Furthermore, prompting strategies can markedly influence model performance (Xia et al., 2026). Current detection and classification methods, including those based on word-frequency similarity and those using perplexity metrics derived from LLMs internal probabilities, fail to enable precise, in-depth comparisons of factors such as disciplinary register, prompting strategies, and model choice (Ahmed, 2026; Xia et al., 2026).

Accordingly, this study adopts Biber's (1988) classic multidimensional analysis framework, along with recent applications in AI as exemplified by Sardinha (2024), to assess the degree of human-like in AI-generated Chinese academic texts. We further investigate the impact of different disciplinary register, prompting strategies, and models on this human-like to probe the boundaries of AI's capabilities. By building a proprietary Chinese AI–human parallel corpus, this study reveals AI’s stylistic limitations and provides a basis for a linguistically informed framework to evaluate AI-generated academic writing.

Differences in Register, Prompting Strategies, and Models for AI-generated academic Chinese

The register perspective forms the foundational framework for modern corpus linguistics in understanding the functional distribution of language. As Biber (1988; Biber & Conrad, 2019) classically posited, a register is not merely a categorical label for a text's context. Instead, it is a systematic pattern of co-occurring linguistic features shaped by specific communicative pressures, cognitive demands, and rhetorical constraints. This perspective views language selection as a highly functional decision-making process intended to meet particular communicative goals. For example, the academic register tends to achieve information compression and objective argumentation through dense nominalization structures, while the spoken register relies on frequent pronouns and interactive features to handle immediate interpersonal communication (Hyland & Jiang, 2018). Through multidimensional analysis, Biber demonstrated that cross-register linguistic variation is often more significant than cross-linguistic variation. This finding provides a robust multidimensional reference system for assessing the degree of human-likeness of AI-generated texts in specific academic contexts (Demir & Egbert, 2025; Sardinha, 2024).

Nevertheless, the universality of register variation does not straightforwardly extend to Chinese academic discourse, where it manifests in markedly different ways. As a typical isolating language, Chinese lacks inflectional morphological changes, which places a more critical role on word order, co-occurrence of function words, and micro-lexical features in expressing academic meaning (Halliday & McDonald, 2004). Existing research points out that the modern Chinese academic style has a unique evolutionary path in terms of abstraction and information density (Feng, 1995; Liardet, 2013), often constructing rigorous logical systems through the high-frequency use of disyllabic words, complex discourse connectives, and specific discourse markers. Considering that LLMs face a severe shortage of

Chinese corpus during the training process, with English corpus accounting for as high as 92.1%, while Chinese only accounts for 0.16% (Brown et al., 2020). This highly imbalanced distribution has incited widespread scholarly skepticism about the quality of AI's acquisition of the Chinese language (Guo et al., 2025; Liang et al., 2023). Existing research has confirmed that while AI can mimic the superficial aspects of academic writing, it struggles with complex, specialized registers. Consequently, the texts it produces tend to be highly formulaic and lack nuance (Demir & Egbert, 2025; Casal & Kessler, 2023).

Furthermore, in the age of large language models, the nature of AI-generated text is typically not random but is instead the product of an interplay between the model, the prompting strategy, and the target register. Different models often display markedly different stylistic preferences when navigating complex register conventions (Hadi et al., 2024; Zhang & Zhang, 2025; Huang & Deng, 2025). Meanwhile, prompt engineering serves as an external intervention that can, through techniques like role-playing, chain-of-thought reasoning, and few-shot examples, steer models to emulate human expert styles across many tasks (Kim & Lu, 2024; Lugli & Strapparava, 2024). Nevertheless, this guidance can also lead to deviations in the AI's writing (Sadasivan et al., 2023). Additionally, an AI's performance can differ across various academic disciplines, as their core knowledge bases and language proficiencies vary (Demir & Egbert, 2025; Hendrycks et al., 2020).

Present research tends to examine single variables in isolation and prioritizes detection over explanation (Birhane et al., 2023; Sadasivan et al., 2023). Although existing research has started to incorporate cross-linguistic comparisons, there remains a notable lack of studies focused on Chinese academic writing concerning different model architectures, prompting strategies, and disciplinary registers. This study seeks to build an AI-Human multilingual academic Chinese comparable corpus and employ multidimensional analysis to contrast the stylistic variations of AI-generated text in Chinese academic writing, thereby investigating the following research questions:

(1) To what extent do AI-generated texts diverge from human-authored texts in the context of academic Chinese?

(2) How do these divergences differ according to the disciplinary register?

(3) Do the choice of model and the prompting strategy affect the degree of human-like in AI-generated text?

Methods

3.1 Corpus construction

3.1.1 Human-authored texts design

The human-authored texts for this study were obtained from authoritative, core journals across various academic disciplines, as indexed by the China National Knowledge Infrastructure (CNKI). To specifically

target the "high-order academic Chinese" style, the selection was limited to the most prestigious journals in each field. Drawing from Biglan's (1973) classification of academic domains, we selected four representative fields: Natural Sciences, Engineering, Humanities, and Social Sciences.

Regarding sample size, we selected 125 papers from each of the four disciplinary registers, resulting in a corpus of 500 papers with a total of 599,098 words. All papers were randomly selected from publications dating to 2022 or earlier. This pre-AI boom period was intentionally chosen to guarantee that the language had not been edited or enhanced by artificial intelligence. The focus was placed on the "Introduction" sections of the papers, as this section is a universal component of academic writing, which is advantageous for comparative analysis across disciplines. In addition, the typical length of an introduction aligns well with the output constraints of modern generative AI models, making it an appropriate segment for evaluating academic Chinese text generation.

3.1.2 AI-generated text design

The AI-generated text originates from the APIs of various AI models that are widely used globally. From these, we have selected five AI models that are extensively used worldwide and demonstrate outstanding performance in Chinese tasks. We selected five models: GPT-5.2, Grok 4, Gemini 3 Flash, Qwen3-Max, and DeepSeek-V3.2. Among these, Qwen and DeepSeek are developed in China, and their training corpora contain a higher proportion of Chinese content compared to the previous three models, which helps to reveal the influence of regional corpus bias on output style. The versions of these models are the most advanced and most currently used versions prior to the writing of this paper, in order to fully explore the boundaries of AI-generated advanced Chinese academic text.

In terms of prompting, the complete title of each human-authored paper serves as a core element of the prompt. To simulate various user application scenarios and assess the effectiveness of prompt engineering in controlling output style, we employ three distinct prompting strategies (Liu et al., 2023):

(1) Zero-shot: The instruction is limited to "Please write an academic paper introduction based on the following title: [Title]";

(2) Few-shot: This builds upon the Zero-shot instruction by adding 2–3 human-authored introductions from the same field as contextual examples. These examples are sourced from a holdout set not used in training to prevent data leakage;

(3) Multi-strategy Prompting: This approach integrates multiple guiding mechanisms, such as style constraints, self-checking instructions, and statistical alignment.

Each of these prompting strategies was applied to all 500 titles, with five distinct models generating an introduction for each title. This process yielded a total of 7,500 AI-generated texts, comprising 7,191,926 characters in aggregate. A detailed description of the corpus can be found in Appendix A, while examples of the prompts used are provided in Appendix B.

3.2 Language feature extraction

Drawing upon Biber's (1988) multidimensional analysis framework and adapting it to accommodate Chinese grammatical properties, we developed an annotation scheme comprising 105 fine-grained linguistic features. The feature design was primarily informed by Xu et al. (2025), with modifications to suit our technical constraints. These features were selected based on their established capacity in the literature to distinguish registers and capture core dimensions of Chinese linguistic variation.

Specifically, the 105 features were drawn from three sources: (1) English linguistic features from Biber (1988, 2006) that are also applicable to Chinese; (2) shared features from Chinese MDA studies by Song et al. (2021) and Zhang (2012); and (3) lexical features and grammatical features identified in prior research (e.g., Feng et al., 2008; Huang & Li, 2017). These features span categories including Pronouns & Deictics, Verbs & Modality, Nouns & Lexical Semantics, Adverbs & Hedges, Adjectives, Stylistic Features, Adpositions & Connectives, Particles & Numerals, Syntax & Clause Types, and Metrics & Punctuation, ensuring comprehensive coverage of Chinese linguistic characteristics. For the complete feature list, see Appendix C.

We then employed THULAC (Language and Lexical Analysis for Chinese) (Sun et al., 2016) for Chinese word segmentation and Part-of-Speech (POS) tagging. Subsequently, regular expression matching was utilized in conjunction with a purpose-built dictionary and syntactic sequence targets to quantify the frequency of each linguistic feature. In line with Biber's methodology, these feature frequencies were standardized to occurrences per thousand characters to normalize for text length and ensure comparability.

3.3 Statistical analysis

The suitability of the 105 original features for factor analysis is first assessed through factorability tests. If the data meets the criteria for factor analysis, the process continues; otherwise, direct comparisons of specific language features will be considered. The subsequent step involves conducting an Exploratory Factor Analysis (EFA) to extract dimensions, retaining only those components with eigenvalues exceeding 1. Prior to dimension aggregation, language features are filtered by removing items with communality values below 0.3 or correlation coefficients above 0.95. Factor extraction is then performed using Unweighted Least Squares (ULS), followed by a Varimax rotation to optimize factor interpretability.

For *RQ1*, the complete set of 8,000 texts was initially categorized into two groups according to origin: AI-generated and human-authored. For each factor score, we calculated the group mean (*M*) and standard deviation (*SD*). Levene's test was conducted to determine the equality of variances; in cases of unequal variance, a Welch's adjusted independent samples *t*-test was utilized, whereas the standard *t*-test was applied when variances were equal. Boxplots were generated for each dimension to facilitate visual comparison. Additionally, Cohen's *d* values were computed for each dimension to illustrate the effect size between the AI and human groups, thereby accounting for the considerable difference in their respective sample sizes.

For *RQ2*, the data were partitioned according to four broad academic disciplines. For each stylistic factor, a two-way Type II ANOVA was employed to assess the main effects of Source and Disciplinary Registers, along with their interaction effect. A significant interaction ($p < 0.05$) would suggest that the linguistic disparity between AI-generated and human text depends on depending on the academic domain. Subsequently, to measure the fidelity of LLMs to human writing, a proximity matrix was constructed by computing the mean absolute deviation between AI and human factor scores within each discipline. Finally, to understand stylistic distinctions across disciplines, a Tukey HSD post-hoc test was performed on the entire dataset to compare linguistic styles on a pairwise basis, thereby uncovering the inherent stylistic heterogeneity among the disciplines.

For *RQ3*, the comprehensive deviation (Global Deviation, *GD*) for each model-prompt strategy combination was calculated within the pure AI samples. Using the overall mean of human text on each factor as the baseline, a lower *GD* indicates a closer approximation of human writing style (Pillutla et al., 2021). This data was used to construct a Model–Prompt Deviation Matrix for the identification of optimal generation configurations. Following this, a two-way Type II ANOVA was conducted for each factor, with model type and prompt strategy as the independent variables. For significant main effects, Tukey HSD post-hoc tests were performed on both model categories and prompt strategies to pinpoint specific sources of significant stylistic deviation, and to assess the presence of any interaction effects between the two factors.

Results

4.1. Factor Analysis and Dimension Establishment

A preliminary assessment confirmed the suitability of the data for factor analysis, as indicated by a *KMO* measure of sampling adequacy of 0.748 and a statistically significant result from Bartlett's test of sphericity ($p < 0.001$), thus meeting the necessary statistical assumptions for this analysis (Kaiser, 1974). Following the screening of linguistic features, a final set of 17 features was selected. These were extracted four principal dimensions, which collectively accounted for 62.47% of the total variance. These four dimensions constitute the primary framework for the subsequent comparative analysis between AI-generated text and human writing, and for investigating the influence of disciplinary register, model type, and prompting strategies. The specific names and operational definitions for each dimension are detailed in Table 1, where the values in parentheses following each linguistic feature represent the corresponding loadings; note that a larger loading indicates greater feature importance.

Table 1
Dimension and Linguistic Features.

Dimension	Description	Linguistic Features
D1: Contextualized Situational Elaboration	This dimension reflects the use of language to describe specific, immediate situations and interactive contexts.	Positive features Onomatopoeia (0.91); Interrogative pronouns (0.88); Average word length (0.87); Locative nouns (0.75); Prepositions (0.60); Numerals (0.60) Negative features None
D2: Loose Structure vs. Compact Expression	This dimension contrasts a more fragmented, repetitive style with a densely packed, formal linguistic structure.	Positive features Monosyllabic words (0.65); Structural particle "De1" (□) (0.56) Negative features High-frequency verbs (-0.84); Coupled disyllabic words (-0.58)
D3: Modification-driven vs. Abstract information	This dimension contrasts a style that relies on frequent adverbial markers with one that depends on dense nominal content, particularly high-frequency and abstract nouns.	Positive features High-frequency adverbs (0.58); Attitudinal adverbs (0.53); Time adverbs (0.53) Negative features High-frequency nouns (-0.72);

Dimension	Description	Linguistic Features
		Abstract nouns (-0.45)
D4: Explicit Epistemic Assertion	This dimension focuses on the direct expression of certainty and the use of copular structures to establish facts.	Positive features Adverbs of possibility (0.88); Copula "shi" (是) (0.78) Negative features None

4.2. Systematic differences between AI-Generated texts and Human-authored texts

Following the construction of the analytical dimensions, a Levene's test was performed to assess variance homogeneity between the AI-generated and human-authored text corpora. The analysis revealed significant heteroscedasticity for D1, D2, and D4, which necessitated the application of Welch's *t*-test for these comparisons. Conversely, D3 satisfied the assumption of homoscedasticity, warranting the use of the standard independent samples *t*-test. A summary of these t-test results is provided in Table 2; comprehensive statistical details are available in Appendix D.

Table 2
A summary table of AI and human scores.

Dimension	AI Mean (*SD*)	Human Mean (*SD*)	Mean Difference	Cohen's *d*	*p*
D1	0.005 (0.46)	−0.082 (0.46)	0.088	0.19	< .001
D2	0.003 (0.30)	−0.041 (0.30)	0.043	0.14	.002
D3	0.008 (0.28)	−0.119 (0.22)	0.127	0.50	< .001
D4	−0.006 (0.22)	0.089 (0.19)	−0.095	−0.46	< .001

As indicated by the table, the *t*-test yields statistically significant results across all four linguistic dimensions. A closer look at Cohen's *d* values shows that the effect sizes for Contextualized Situational Elaboration (0.19) and Loose Structure vs. Compact Expression (0.14) are small. In contrast, the effect sizes for Modification-driven vs. Abstract information (0.50) and Explicit Epistemic Assertion (-0.46) are substantial. Additional distinctions between the groups are illustrated in Fig. 1.

The AI group's mean score on the Contextualized Situational Elaboration dimension is markedly higher than the human group's, though the effect size is small (Cohen's *d*). The AI group's scores are more tightly clustered around 0, whereas the human group's scores trend toward the negative end of the scale. This suggests a deliberate, albeit not pronounced, tendency in human academic writing to avoid and

downplay immediate situational descriptions. Similarly, on the Loose Structure vs. Compact Expression dimension, the effect size is also small. Here, the AI group's scores are slightly positive, indicating a modest tendency toward looser organization and showing a centralized score distribution similar to that observed on the Contextualized Situational Elaboration dimension.

The dimension of Modification-driven vs. Abstract information shows the most pronounced disparity between the AI and human groups, with a Cohen's *d* of 0.502, positioning it as a key differentiator. The AI group's significantly high scores indicate a strong inclination to guide the reader toward a subjective viewpoint, often relying on adverbial cohesion, particularly the most common high-frequency adverbs. Conversely, the human group's scores lean negative, demonstrating that human authors are adept at using concrete, entity-focused nouns to build abstract academic frameworks.

Explicit Epistemic Assertion is the sole dimension where human writers significantly outperformed the AI model, with a substantial effect size of -0.46. This dimension assesses the frequency of use for adverbs of possibility and the copular verb "shi." The human group's higher scores reflect a greater confidence in making definitive factual claims during academic arguments. The AI group's negative scores, by contrast, reveal a cautious and vague approach, indicating a relative weakness in establishing strong, logical assertions and an inability to construct factual claims as directly as human authors.

4.3. The differences between AI-generated texts and Human-authored text in Disciplinary Registers

This study further investigates the stylistic discrepancies between AI-generated and human-authored texts across four distinct linguistic dimensions, exploring whether these differences vary systematically across Disciplinary Registers. A summary of these results is presented in Table 3; comprehensive statistical details from the two-way Type II ANOVA and subsequent post-hoc tests are available in Appendix E.

Table 3
A summary table of scores across different registers.

Dimension	**Register Ranking (High → Low)**	**AI vs. Human Difference**	**Interaction (Source × Domain)**
D1	Social Sci > Science ≈ Engineering > Humanities	$p < .001$	$p < .001$
D2	Engineering > Humanities > Science > Social Sci	$p = .002$	$p < .001$
D3	Humanities > Science > Engineering > Social Sci	$p < .001$	$p < .001$
D4	Humanities > Engineering > Science > Social Sci	$p < .001$	$p = .208$

The scores for each group, as presented in the table, indicate the following: In the Contextualized Situational Elaboration category, the social sciences group achieved significantly higher scores than the science, engineering, and humanities groups. For Loose Structure vs. Compact Expression, the engineering group scored highest, followed in descending order by the social sciences. In Modification-driven vs. Abstract information, the humanities group exhibited the strongest performance, followed by science, with engineering and social sciences showing comparatively lower scores. Similarly, for Explicit Epistemic Assertion, the humanities group achieved the highest scores, with engineering scoring slightly higher, and science and social sciences scoring lower. These cross-disciplinary discrepancies are further illustrated in the box plots in Fig. 2.

On the Contextualized Situational Elaboration dimension, the AI group's scores fluctuated around zero across several dimensions. In the three registers of social science, humanities, and engineering, the AI group's scores were slightly higher than the human group's, similar to the results in Fig. 1. However, in science, there was a significant difference; the human group's score was markedly higher than the AI group's, and the human group's score in science was also higher than in other registers. In the human group, science required detailed descriptions of experiments and related details, a point that the AI did not capture effectively.

On the Loose Structure vs. Compact Expression dimension, science also exhibited a similar performance to the Contextualized Situational Elaboration dimension. Additionally, the engineering register showed a phenomenon where the human group's score was slightly higher than the AI group's. This indicates that in the human group, the language in science and engineering is more concise and to the point.

In the Modification-driven vs. Abstract information dimension, the AI group's scores were significantly higher than the human group's. Notably, the AI group's scores in science and engineering were both higher than the human group's, and in engineering, the human group's score was the lowest while the AI group's score was the highest. This difference reflects that the AI failed to adequately imitate the high abstract information density of human writing and used too many modifying components, a phenomenon particularly evident in science and engineering.

In the Explicit Epistemic Assertion dimension, the disciplinary register did not play a significant distinguishing role; in all four registers, the AI failed to make clear conclusions and judgments effectively.

4.4 Effects of Various Models and Prompting Strategies

A summary of the overall stylistic profiles of AI-generated and human-authored texts is provided in Table 4. For a detailed breakdown of the two-way Type II ANOVA and subsequent post-hoc tests, refer to Appendix F.

Table 4
Model–Prompt Deviation Matrix.

DeepSeek-V3.2	zero-shot	few-shot	multi
	0.0798	**0.0939**	**0.1233**
Gemini 3 Flash	0.0824	0.0634	0.1071
GPT-5.2	0.0709	0.0872	0.2016
Grok 4	0.1093	0.1241	0.1558
Qwen3-Max	0.0684	0.1064	0.1678

Note that for this metric, smaller values indicate an output closer to a human-like style. As the table illustrates, all models registered higher GD values under the multi-strategy condition compared to both zero-shot and few-shot settings. Specifically, GPT-5.2 displayed the greatest deviation under the multi-strategy approach at 0.2016, whereas Qwen3-Max achieved the lowest score under zero-shot at 0.0684, and Gemini 3 Flash also performed remarkably well under few-shot with a score of 0.0634.

A two-way Type II ANOVA confirmed that the main effects for both model and prompt, along with their interaction effect, were highly significant ($p < 0.001$) across all four dimensions. This indicates the presence of systematic stylistic differences among the models, suggesting that prompting strategies exert a strong regulatory influence and that these two factors are not separable. Tukey's HSD post-hoc tests further revealed that, overall, Qwen3-Max and Gemini 3 Flash were the top performers. In contrast, Deepseek-V3.2 demonstrated the most consistent style, exhibiting minimal variation across the three prompting strategies while consistently performing at a mid-to-low level. Conversely, Grok 4 and GPT-5.2 exhibited weaker capabilities, characterized by substantial performance variations across different dimensions. These discrepancies are further illustrated in the box plots in Fig. 3.

Specifically, on the Contextualized Situational Elaboration metric, the median scores for Grok 4 and GPT-5.2 under the multi-strategy condition were significantly and negatively biased, far below the central range of the human group; in contrast, the distributions for Qwen3-Max and DeepSeek-3.2 under the zero-shot condition were closer to the human group. At the same time, the differences between the models were significant: Grok 4 was significantly higher than Gemini 3 Flash and DeepSeek-V3.2, indicating its strongest situational generation capability, but it also deviated the most from the human restrained style.

In terms of Loose Structure vs. Compact Expression, Grok 4 also shows the highest median across all strategies, whereas the distributions for Gemini 3 Flash and DeepSeek-V3.2 are closer to the human preference for compact expression. Both few-shot and multi-strategy cause the output to deviate more from the human group on these two dimensions, with the deviation degree of multi-strategy being greater than that of few-shot.

The differences in Modification-driven vs. Abstract information are the most prominent. The medians for GPT-5.2 and Qwen3-Max under multi-strategy shift significantly higher, markedly surpassing the negative

central tendency of the human group; meanwhile, Gemini 3 Flash under all conditions is close to or below the human median, exhibiting the most convergent style. Statistically, the prompt effect peaks at this point (= 384.79), and the mean difference for multi-strategy relative to zero-shot is as high as + 0.1781 ($p < 0.001$).

Regarding Explicit Epistemic Assertion, DeepSeek-V3.2 and Qwen3-Max under multi-strategy similarly demonstrate the strongest assertion tendency, with medians significantly higher than other models. Similar to other dimensions, the deviation degree of multi-strategy is also significantly higher than that of zero-shot and few-shot. The difference is that on this dimension, few-shot is superior to zero-shot, reflecting that different prompting methods have varying effects on different dimensions.

Discussion

Current AI-generated texts exhibit systematic differences from human-authored Chinese academic papers, with particularly pronounced distinctions in the Modification-driven vs. Abstract information and Explicit Epistemic Assertion dimensions. This finding differs from Zhang and Zhang's (2025) analysis of English abstracts, which concluded that AI-generated texts perform well in epistemic stance but poorly in affective stance. In contrast, AI performance in academic Chinese is also limited in epistemic stance. This significant divergence is not merely an incidental choice but rather arises from a fundamental conflict between the probabilistic nature of LLMs and the cognitive requirements of academic writing (Bender et al., 2021; Caruccio et al., 2024). Additionally, these differences are influenced by disciplinary-specific conventions, and variations in generation conditions can likewise affect the degree of similarity between AI-generated texts and human-authored Chinese academic papers.

Although AI can generate grammatically correct and logically coherent academic texts that assist writers in structuring their work, its language style consistently diverges from the nuanced characteristics of human-authored Chinese academic writing. This indicates that the model fails to capture the intricate semantic structures of such texts. At its core, academic writing is a highly register-sensitive activity where authors must construct discipline-specific academic discourse (Hyland, 2004). AI does not adequately differentiate between registers, as its representations across them are highly similar. This issue is not simply a matter of training data proportions but rather results from the model's averaging of semantic features within its vector space (Holtzman et al., 2019). To maximize output probability, models often converge toward a generic, conventional academic tone, erasing the sharp distinctions between engineering and social sciences. This reflects that while LLM outputs are fluent and coherent, they exhibit heightened stylistic uniformity and tight clustering across models (O’Sullivan, 2025).

This finding aligns with previous research confirming AI's strong capacity for generating academic texts (Dwivedi et al., 2023), yet it is regrettable that AI's awareness of registers is limited to the broad academic category and does not extend to specific disciplines. AI fails to recognize that the linguistic characteristics of academic Chinese texts differ significantly across individual disciplinary registers (Williams, 2024).

What is more concerning is that prompt engineering, a technique widely and effectively used in other fields to boost model performance, has proven largely ineffective for academic Chinese writing. Far from bringing the AI's writing style closer to that of a human, these methods often exacerbate the gap between AI-generated and human-authored text. This anomalous behavior suggests a fundamental mismatch between the instructions given and the model's internal representation. When we instruct an AI to "write like a scholar," it activates simplistic stereotypes about academia from its training data, rather than engaging in genuine disciplinary reasoning. Moreover, performance across models is inconsistent; even general-purpose models like Gemini sometimes rival or surpass Chinese-specialized models like DeepSeek. This pattern points to a broader, systemic difficulty among LLMs in capturing the stylistic and rhetorical conventions of academic Chinese, rather than issues confined to specific architectures or prompts.

Consider GPT-5.2 as an example: when prompted with a multi-strategy approach, its GD metric spikes to 0.2016. This jump is primarily driven by an extremely high score in the Modification-driven vs. Abstract information dimension. While the goal was to encourage more human-like writing, the outcome was that the AI overused modifying adverbs across all academic disciplines. This can be seen as a form of semantic dilution. Human authors increase academic rigor by elevating the abstraction of their ideas, whereas the AI attempts to mimic complexity by adding superfluous modifiers (Sadasivan et al., 2023). In contrast, models like Qwen and Gemini exhibit lower deviation in a zero-shot setting, suggesting a more conservative generation strategy that avoids the indiscriminate blurring of stylistic registers. Just as Huang & Cheung (2025) discovered, not all more powerful models can perform better in translation tasks. This is closely related to factors such as model training methods and data structures.

The core issue is that AI fails to grasp the subtle balance between restraint and assertion in academic Chinese. Authentic academic writing in Chinese does not depend on frequent hedging adverbs; rather, it establishes facts directly using the copula shi, based on stable concepts created through nominalization. AI's deficiency in this domain is twofold: it shies away from definitive claims due to a pervasive ambiguity in its training data (Ji et al., 2025), and simultaneously, its lack of conceptual grounding compels a reliance on adverbial connectors. This results in a discourse that is neither resolute nor incisive, remaining in a perpetual state of "float"—ostensibly objective but ultimately devoid of a substantive position (Santurkar et al., 2023). This represents a risk-averse generative strategy, where models are trained to minimize hallucinations, thereby favoring obfuscated assertions. This tendency is deeply embedded within the alignment mechanisms of Reinforcement Learning from Human Feedback (RLHF): during safety training, a vague truth is often prioritized over a precise error, compelling the model to default to flatter, more probabilistic expression when confronted with academically rigorous claims that entail intellectual risk (Ouyang et al., 2022). In the context of academic discourse, however, precisely calibrated assertiveness is a critical rhetorical tool for building authority and facilitating the advancement of knowledge (Hyland, 2010).

Consequently, while AI's cautious ambiguity may lower its error rate, it simultaneously erodes the epistemic courage essential for scholarly argumentation.

Conclusion

This research supplements the prevailing optimism in AI writing research (Sardinha, 2024) by demonstrating that model adaptation in academic Chinese contexts remains superficial. When faced with specific disciplinary registers, current LLMs exhibit a marked tendency toward stylistic convergence rather than nuanced register discrimination. In particular, the dilution of certainty and excessive use of adverbs reflect the alignment constraints of RLHF protocols, which often prioritize safety over substantive academic precision (Ouyang et al., 2022).

Furthermore, our analysis reveals that neither enhanced prompt engineering nor superior general language proficiency leads to consistently higher personification in academic writing. This suggests that academic discourse involves a complex integration of multiple competencies, such that no single model or prompt reliably produces more personified text across contexts.

Looking forward, this study calls for a transition from surface-level language optimization toward a deep integration of disciplinary characteristics. True progress lies in evolving AI from a generative tool into a discerning, discipline-aware writing partner (Amershi et al., 2019). Achieving this requires a collaborative synergy between computer scientists, linguists, and subject matter experts to refine the sophisticated linguistic decision-making necessary for specialized academic domains.

Declarations

Author Contribution

C.H. and W.Z. contributed equally to this work and should be considered co-first authors. W.Z. performed the experiments, conducted the data analysis, and wrote the original draft of the manuscript. C.H. acquired funding for the project and reviewed and edited the manuscript. Y.Z. provided resources including equipment and materials, and reviewed and edited the manuscript. All authors reviewed and approved the final version of the manuscript.

Acknowledgement

We thank Ying Fu for valuable contributions to the data collection.

Data Availability

All data supporting the findings of this study, including all research corpora and related supplementary materials, are available within the paper and its Supplementary Information.

References

1. Ahmed M, Knockel J, Greenstadt R (2025) An analysis of Chinese censorship bias in LLMs. Proc Priv Enh Technol 2025(4):112–129. https://doi.org/10.56553/popets-2025-0122
2. Amershi S, Weld D, Vorvoreanu M et al (2019) Guidelines for human-AI interaction. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, pp 1–13. https://doi.org/10.1145/3290605.3300233
3. Barrot JS (2023) Using ChatGPT for second language writing: pitfalls and potentials. Assess Writ 57:100745. https://doi.org/10.1016/j.asw.2023.100745
4. Batista J, Mesquita A, Carnaz G (2024) Generative AI and higher education: trends, challenges, and future directions from a systematic literature review. Information 15(11):676. https://doi.org/10.3390/info15110676
5. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, pp 610–623. https://doi.org/10.1145/3442188.3445922
6. Biber D (1988) Variation across speech and writing. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511621024
7. Biber D (2006) A corpus-based study of spoken and written registers. John Benjamins Publishing Company, Amsterdam. https://doi.org/10.1075/scl.23
8. Biber D, Conrad S (2019) Registers, genres, and styles: fundamentals of the triple perspective. Register, genre, and style, 2nd edn. Cambridge University Press, Cambridge, pp 1–35
9. Biglan A (1973) The characteristics of subject matter in different academic areas. J Appl Psychol 57(3):195. https://doi.org/10.1037/h0034701
10. Birhane A, Kasirzadeh A, Leslie D, Wachter S (2023) Science in the age of large language models. Nat Rev Phys 5(5):277–280. https://doi.org/10.1038/s42254-023-00581-4
11. Bommasani R, Liang P, Lee T (2023) Holistic evaluation of language models. Ann N Y Acad Sci 1525(1):140–146. https://doi.org/10.1111/nyas.15007
12. Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
13. Caruccio L, Cirillo S, Polese G et al (2024) Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. Expert Syst Appl 235:121186. https://doi.org/10.1016/j.eswa.2023.121186
14. Casal JE, Kessler M (2023) Can linguists distinguish between ChatGPT/AI and human writing? a study of research ethics and academic publishing. Res Methods Appl Linguist 2(3):100068. https://doi.org/10.1016/j.rmal.2023.100068
15. Demir NY, Egbert J (2026) Register alignment of ChatGPT-generated academic texts. Appl Corpus Linguist 6(1):100174. https://doi.org/10.1016/j.acorp.2025.100174

16. Dilai M, Dilai I (2024) AI vs humans: a corpus-based study of English academic writing. Lang Technol Humanit Soc 5(0):11: 145–154
17. Doughman J, Mohammed Afzal O, Toyin HO et al (2025) Exploring the limitations of detecting machine-generated text. In: Rambow O, Wanner L, Apidianaki M (eds) *Proceedings of the 31st International Conference on Computational Linguistics*. *Association for Computational Linguistics*, pp 4274–4281
18. Dwivedi YK, Kshetri N, Hughes L et al (2023) Opinion paper: So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manag 71:102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
19. Feng S (1995) Prosodic structure and prosodically constrained syntax in Chinese. University of Pennsylvania, Philadelphia
20. Feng SL, Wang J, Huang M (2008) Hanyu shumian yuti zhuangyadu de zidong celiang [An automatic feature checking algorithm for degree of formalities in written Chinese]. Linguist Sci 7(2):113–126 (in Chinese)
21. Freeman J (2025) Student generative AI survey 2025. Higher Education Policy Institute, London
22. Guo Y, Conia S, Zhou Z et al (2025) Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp 3823–3838. https://doi.org/10.18653/v1/2025.acl-long.193
23. Hadi MU, Al Tashi Q, Qureshi R et al (2023) Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Prepr 1(3):1–26
24. Halliday MA, McDonald E (2006) Metafunctional profile of the grammar of Chinese. Language typology: a functional perspective. John Benjamins Publishing Company, Amsterdam, pp 305–396. https://doi.org/10.1075/cilt.253.08hal
25. Hendrycks D, Burns C, Basart S et al (2020) Measuring massive multitask language understanding. *arXiv:2009.03300*. https://doi.org/10.48550/arXiv.2009.03300
26. Holtzman A, Buys J, Du L et al (2019) The curious case of neural text degeneration. *arXiv:1904.09751*. https://doi.org/10.48550/arXiv.1904.09751
27. Huang BR, Liao XD (2017) *Xiandai Hanyu* [Modern Chinese language], 6th edn. Higher Education Press, Beijing. (in Chinese)
28. Huang L, Deng J (2025) This dissertation intricately explores… ChatGPT's shell noun use in rephrasing dissertation abstracts. *System* 129:103578. https://doi.org/10.1016/j.system.2024.103578
29. Huang Y, Cheung AKF (2026) Exploring AI's performance in literary autobiography translation: how closely do AI models match human translation. Humanit Soc Sci Commun. https://doi.org/10.1057/s41599-026-06630-4

30. Hyland K (2004) Writing in the disciplines. In: *Disciplinary discourses: social interactions in academic writing*, Michigan Classics edn. University of Michigan Press, Ann Arbor, pp 3–25
31. Hyland K (2010) Metadiscourse: mapping interactions in academic writing. Nordic J Engl Stud 9(S2):125–143. https://doi.org/10.35360/njes.220
32. Hyland K, Jiang F (2018) Academic lexical bundles: how are they changing? Int J Corpus Linguist 23(4):383–407. https://doi.org/10.1075/ijcl.17080.hyl
33. Ji J, Qiu T, Chen B et al (2025) AI alignment: a contemporary survey. ACM Comput Surv 58(5):132. https://doi.org/10.1145/3770749
34. Ji L, Yao L, Xu W et al (2025) The impact of prompt language and representation on LLM reasoning: a multilingual empirical study. IEEE Access 14:2249–2267. https://doi.org/10.1109/ACCESS.2025.3649535
35. Kaiser HF (1974) An index of factorial simplicity. Psychometrika 39(1):31–36. https://doi.org/10.1007/BF02291575
36. Kim M, Lu X (2024) Exploring the potential of using ChatGPT for rhetorical move-step analysis: the impact of prompt refinement, few-shot learning, and fine-tuning. J Engl Acad Purp 71:101422. https://doi.org/10.1016/j.jeap.2024.101422
37. Koponen M, Salmi L, Nikulin M (2019) A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. Mach Transl 33(1):61–90
38. Krings H (1986) Translation problems and translation strategies of advanced German learners of French (L2). In: House J, Blum-Kulka S (eds) Interlingual and intercultural communication. Gunter Narr, Tübingen, pp 263–275
39. Krüger R (2020) Explicitiation in neural machine translation. Lang Cult 21(2):195–216
40. Kuang S, Xiong D, Luo W et al (2017) Modeling coherence for neural machine translation with dynamic and topic caches. *arXiv:1711.11221*
41. Kuzman T, Vintar Š, Arcan M (2019) Neural machine translation of literary texts from English to Slovene. In: *Proc Qualities Lit Machine Transl*, pp 1–9
42. Lee T (2023) Artificial intelligence and posthumanist translation: ChatGPT versus the translator. Appl Linguist Rev. https://doi.org/10.1515/applirev-2023-0122
43. Li CN, Thompson SA (1979) Third-person pronouns and zero-anaphora in Chinese discourse. Discourse Syntax. Brill, Leiden, pp 311–335
44. Li H, Graesser AC, Cai Z (2014) Comparison of Google translation with human translation. In: *The 27th Int FLAIRS Conf*
45. Li J (2011) Xinghe vs. yihe in English-Chinese translation—analysis from the perspective of syntactic construction. Perspect 19(3):189–203
46. Li J, Gao Y (2025) Variability of cohesion and coherence in Chinese-to-English translation: measuring the effect of translation variety and register divergence. Hum Soc Sci Commun 12(1):1–18

47. Liang W, Yuksekgonul M, Mao Y et al (2023) GPT detectors are biased against non-native English writers. Patterns 4(7). https://doi.org/10.1016/j.patter.2023.100779
48. Liardét CL (2013) An exploration of Chinese EFL learner's deployment of grammatical metaphor: learning to make academically valued meanings. J Second Lang Writ 22(2):161–178. https://doi.org/10.1016/j.jslw.2013.03.008
49. Liu P, Yuan W, Fu J et al (2023) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 55(9):1–35. https://doi.org/10.1145/3560815
50. Liu X, Li Y, Li K (2025) Enhancing the robustness of AI-generated text detectors: a survey. Mathematics 13(13):2145. https://doi.org/10.3390/math13132145
51. Lugli S, Strapparava C (2024) Multimodal chain-of-thought prompting for metaphor generation. In: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, vol 3878, pp 523–530
52. Nguyen L, Barrot JS (2024) Detecting and assessing AI-generated and human-produced texts: the case of second language writing teachers. Assess Writ 62:100899. https://doi.org/10.1016/j.asw.2024.100899
53. O'Sullivan J (2025) Stylometric comparisons of human versus AI-generated creative writing. Humanit Soc Sci Commun 12:1708. https://doi.org/10.1057/s41599-025-05986-3
54. Ouyang L, Wu J, Jiang X et al (2022) Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 35:27730–27744. https://doi.org/10.48550/arXiv.2203.02155
55. Pillutla K, Swayamdipta S, Zellers R et al (2021) Mauve: measuring the gap between neural text and human text using divergence frontiers. Adv Neural Inf Process Syst 34:4816–4828
56. Pratama AR (2025) The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. PeerJ Comput Sci 11:e2953. https://doi.org/10.7717/peerj-cs.2953
57. Reinhart A, Markey B, Laudenbach M et al (2025) Do LLMs write like humans? Variation in grammatical and rhetorical styles. Proc Natl Acad Sci USA 122(8):e2422455122. https://doi.org/10.1073/pnas.242245512
58. Sadasivan VS, Kumar A, Balasubramanian S et al (2023) Can AI-generated text be reliably detected? *arXiv:2303.11156*. https://doi.org/10.48550/arXiv.2303.11156
59. Santurkar S, Durmus E, Ladhak F et al (2023) Whose opinions do language models reflect? In: *International Conference on Machine Learning*. PMLR, pp 29971–30004
60. Sardinha TB (2024) AI-generated vs human-authored texts: a multidimensional comparison. Appl Corpus Linguist 4(1):100083. https://doi.org/10.1016/j.acorp.2023.100083
61. Shamsi A, Wang T, Amraei M, Raju NV (2026) Evaluating AI text detection tools for distinguishing human-written from AI-generated abstracts in Persian-language journals of library and information science. Acta Inf Pragensia 2026(1):126–134. https://doi.org/10.18267/j.aip.293

62. Song J, Qu Y, Zhu X et al (2021) A multi-dimensional approach to register variations in Mandarin Chinese. Glottometrics 51:39–69. https://doi.org/10.53482/2021_51_393
63. Sun M, Chen X, Zhang K et al (2016) Thulac: an efficient lexical analyzer for Chinese. Retrieved Jan 10, 2022
64. Van Noorden R, Perkel JM (2023) AI and science: what 1,600 researchers think. Nature 621(7980):672–675. https://doi.org/10.1038/d41586-023-02980-0
65. Williams A (2024) Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. Int J Educ Technol High Educ 21(1):52. https://doi.org/10.1186/s41239-024-00485-y
66. Xia Y, Stańczak K, Roth B (2026) Explaining generalization of AI-generated text detectors through linguistic analysis. *arXiv:2601.07974*. https://doi.org/10.48550/arXiv.2601.07974
67. Xu Q, Tabata T, Zhu Y (2025) Multi-dimensional analysis of Japanese native speakers' L2 Chinese writing. J Second Lang Writ 68:101204. https://doi.org/10.1016/j.jslw.2025.101204
68. Zhang M, Zhang J (2025) Human-written vs. ChatGPT-generated texts: stance in English research article abstracts. System 134:103842. https://doi.org/10.1016/j.system.2025.103842
69. Zhang ZS (2012) A corpus study of variation in written Chinese. Corpus Linguist Linguist Theory 8(1):209–240. https://doi.org/10.1017/S1351324911000279

Figures

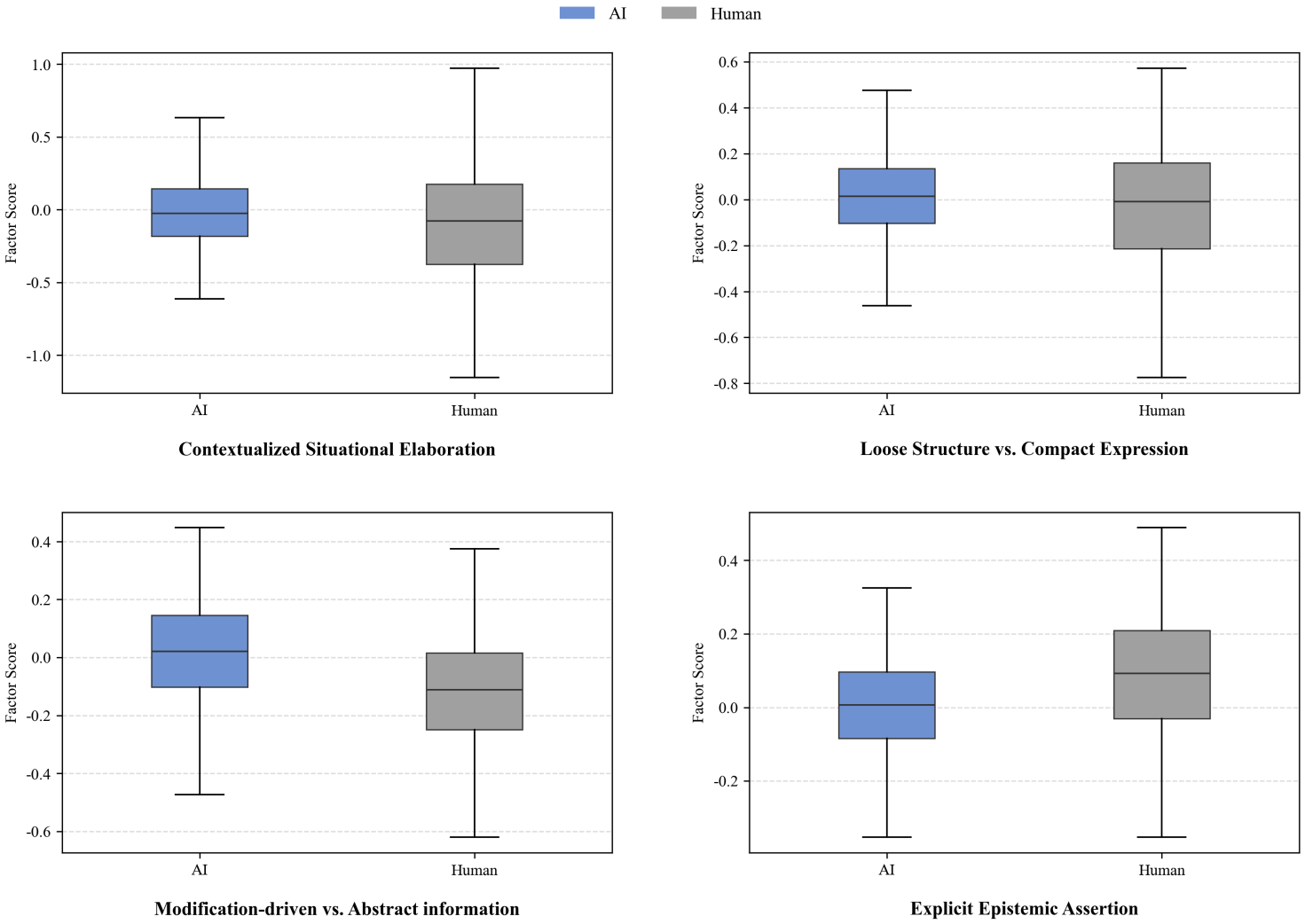

Figure 1

Boxplots of AI and human scores.

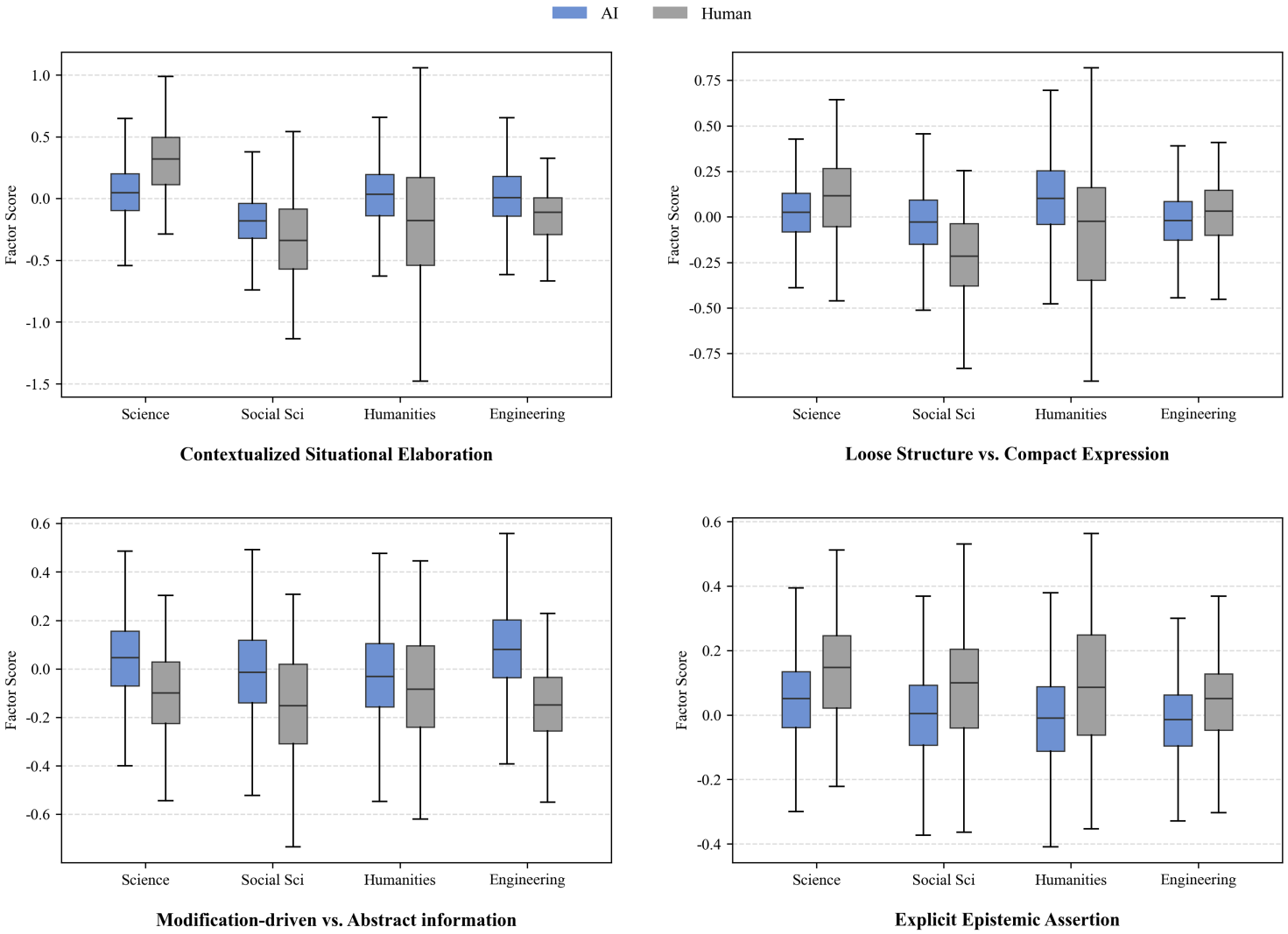

Figure 2

Boxplots of AI and human scores across different registers.

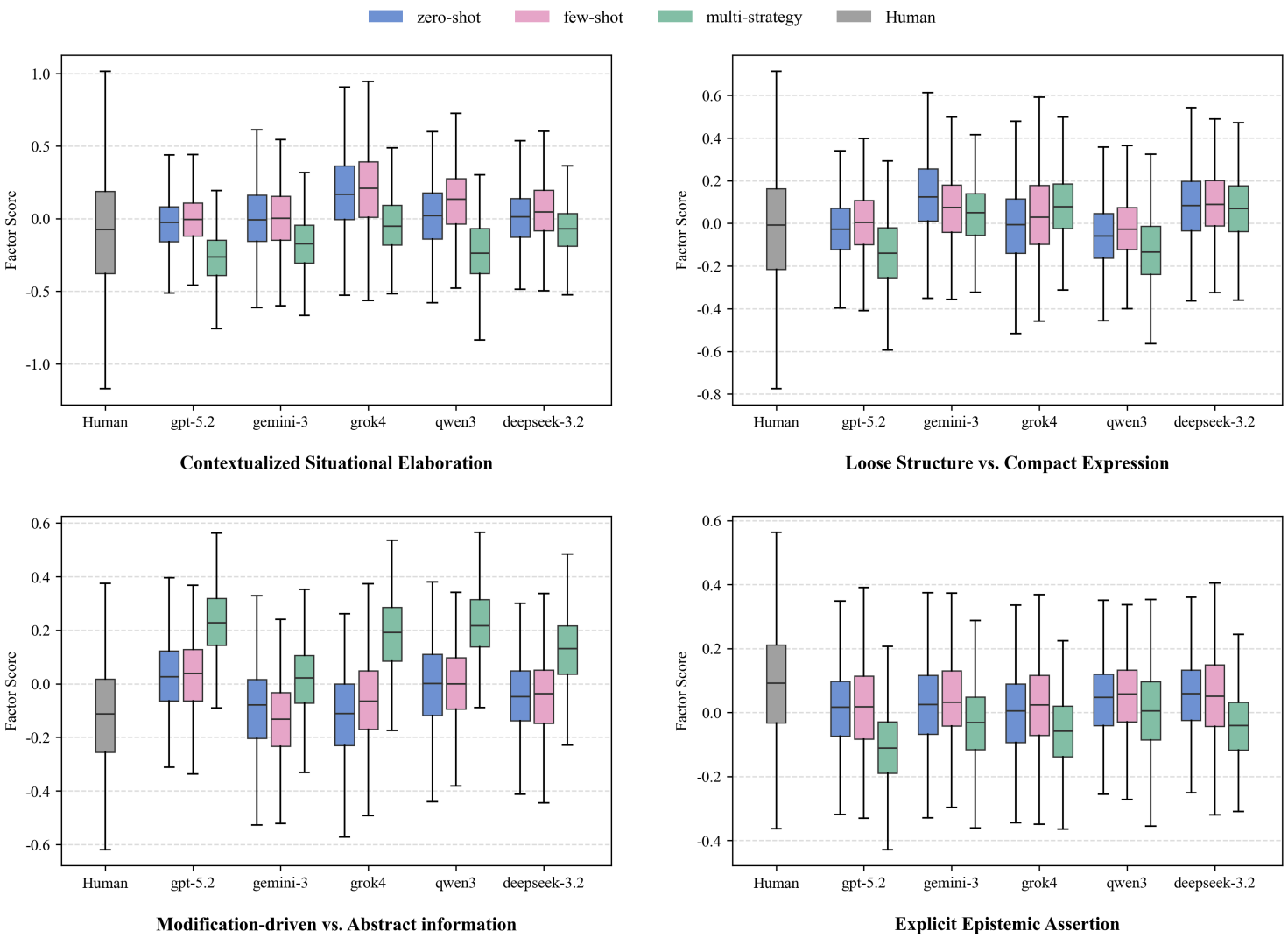

Figure 3

Boxplots of AI and human scores for different models and prompt strategies.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx
- corpus.json