GenAIese - A Comprehensive Comparison of GPT-4o and DeepSeek-V3 for English-to-Chinese Academic Translation

Longhui Zou[1], Ke Li [2], Joshua Lamerton[3], Mehdi Mirzapour[3]

[1]University of Montana, [2]Guangdong University of Foreign Studies, [3]PropTexx USA

Correspondence: lzou4@kent.edu

Abstract

This study investigates the translation performance of two recent large language models—ChatGPT-4o and DeepSeek-V3—in translating English academic papers on language, culture, and literature into Chinese at the discourse level. Using a corpus of 11 academic texts totaling 3,498 sentences, we evaluated translation quality through reference-free automatic metrics (COMET-KIWI), lexical diversity indicators, and syntactic complexity measures. Our findings reveal an interesting contrast: while DeepSeek-V3 achieves higher overall quality scores, GPT-4o produces translations with consistently greater lexical richness (higher type-token ratio, standardized TTR, average sentence length, and word entropy) and syntactic complexity across all five measured metrics, such as Incomplete Dependency Theory Metric (IDT), Dependency Locality Theory Metric (DLT), Combined IDT+DLT Metric (IDT+DLT), Left-Embeddedness (LE), and Nested Nouns Distance (NND). Particularly notable are GPT-4o's higher scores in Left-Embeddedness and Nested Nouns Distance metrics, which are specifically relevant to Chinese linguistic patterns. The divergence between automatic quality estimation and linguistic complexity metrics highlights the multifaceted nature of machine translation quality assessment.

1 Introduction

Quality estimation (QE) of machine translation (MT) products has long been a key area of interest to both MT developers and translation scholars. A wide range of QE methods have been developed to provide information on improving and selecting MT systems. At the same time, specific features in machine-involved translation products, such as high levels of semantic and syntactic literality and unidiomatic target language expressions, have also attracted great research interest, offering implications for further understanding of MT systems and the design of more reliable and valid QE methods. For statistical and neural machine translation (NMT) systems, MT outputs tend to be easily influenced by source text structure, exhibiting a stronger structural shining-through effect, lower level of target-text normalization and reduced linguistic richness when compared to from-scratch human translation products (Bizzoni et al., 2020; Vanmassenhove et al., 2021). Those features, referred to as "machine translationese", will further bring an effect on downstream post-editors, leading to "post-editese" – features distinguishing post-editing products from from-scratch translation products, including lexical simplification and more salient syntactic influence from the source text (Toral, 2019).

The performance of large language models (LLMs) applied to translation tasks has been proven promising. LLMs outperformed commercial NMT models in various language pairs when it comes to document-level translation (Kocmi et al., 2024; Wang et al., 2023). Wang et al. (2023) found that the strength of GPT-powered translation is more salient when it comes to human evaluation, possibly due to its advantage in contextual coherence and naturalness of the target language. The better ability of context awareness in GPT-powered translation over NMT systems is proved by Castilho et al. (2023), who found the advantage in all tested language pairs except a low-resource pair (English-Irish). LLM-powered translation is also found to be less literal compared to NMT systems when translating out of English (Raunak et al., 2023), showing its potential in tackling issues of machine translationese. However, a comparison of translation error types between ChatGPT and NMT systems found more frequent over-translation and mistranslation

Proceedings of the Eleventh Workshop on Patent and Scientific Literature Translation (PSLT 2025), pages 1–12, June 24, 2025

errors in GPT-powered translation, suggesting the problem of hallucinations in LLM-powered translation (Jiao et al., 2023).

Despite the research effort in investigating LLM-powered translations, previous studies mostly followed quality assessment methods used in assessing MT outputs, either on a sentence level or on a document level of a short length. Such a short evaluation unit might prevent us from exploring linguistic issues that will probably bring negative effects to the users in real-life translation settings, for example, cohesion and coherence issues. This concern is particularly significant in the translation of academic papers, as these texts are typically lengthy, feature complex syntactic structures, and demand high accuracy in terminology. Rendering academic papers on topics such as language, culture and literature is even more challenging. These papers are often rich in cultural-loaded expressions, wordplays, and quotations of literary works, making them complex combinations of technical and creative texts. Therefore, translating such papers requires a high level of accuracy and creativity.

Previous research highlights that translators often face challenges in maintaining consistency, choosing precise terminology, and adapting to evolving technical language (Al-Smadi, 2022). Moreover, the complexity of subject matter and the need for adherence to proper academic style make the translation process time-consuming and meticulous (Paperpal, 2023). Ensuring high-quality translations requires not only linguistic proficiency but also careful proofreading and a deep understanding of the field's specialized knowledge. If LLMs can be effectively tested and proven capable of handling these challenges, they could significantly assist professional translators and streamline the translation process by improving consistency, reducing repetitive manual corrections, and assisting with complex syntactic structures.

Therefore, there is a pressing need to explore both LLM-powered MT capabilities and corresponding QE methods at the discourse level. This dual focus would not only advance the development of more sophisticated translation technologies but also ensure reliable quality assessment methods that can effectively evaluate long-form translations, considering factors such as terminology consistency, cross-reference accuracy, and overall document coherence.

The present paper reports a study on the performance of LLM-powered MT products under the context of academic translation. Our focus is on comparing the performance of two non-reasoning models, namely, ChatGPT-4o and Deepseek-V3, in English-Chinese translation. Deepseek-V3 is a free-access model recently released in December 2024. Although user feedback on the use of Deekseek in translation tasks is generally positive, experts in the language industry remain skeptical of its translation capability, pointing out that it did not outperform other mainstream LLMs in some language pairs and domain-specific use cases (Slator, 2024). This calls for a systematic evaluation of Deepseek-powered translation products before reaching a solid conclusion on its translation performance. The significance of this study lies in its examination of Deepseek's translation capabilities since its recent launch. The investigation comes at a critical moment in the way that machine translation increasingly handles complex document-level tasks. This research addresses two notable gaps in the current literature: First, it provides early empirical evidence of Deepseek's translation performance, contributing to our understanding of emerging large language models; second, it offers a systematic comparison with ChatGPT at the discourse level, moving beyond the more common sentence-level evaluations. This comprehensive analysis at the discourse level is particularly valuable as it better reflects real-world translation scenarios and reveals how these models handle broader context and maintain consistency across longer texts.

2 Methodology

11 English open-access research papers published in linguistic journals were selected as source texts (STs) in this study. To facilitate the evaluation process and ensure consistent comparison across models, their titles, sub-titles, tables, figures, notes, bibliography, acknowledgments, and appendices were removed. This preprocessing step was necessary to focus our analysis on the continuous prose sections that form the core content of academic papers, eliminating structural elements that might be handled differently by LLM systems and potentially skewing the comparative results.

The STs totaled 3,498 sentences (93,865 tokens), covering a range of topics including language, culture, and literature. Examples appeared in some articles that involve languages other than English (1,044 tokens) were removed when calculating readability indices. Profiling of the STs is

reported in Table 1, in which the Flesch Reading Ease score (RDFRE), the Flesch-Kincaid Grade Level (RDFKGL), and the Coh-Metrix L2 Readability (RDL2), as indicators of text complexity, were calculated by the Coh-Metrix software (McNamara et al., 2014). Flesch Reading Ease score (RDFRE) measures the readability of a text based primarily on sentence length and word length (syllable count). Lower RDFRE scores indicate more complex text, while higher scores indicate more easy-to-read text. In our data, ST1 has the lowest RDFRE (18.206) and is described as the most complex.

Flesch-Kincaid Grade Level (RDFKGL) estimates the US school grade level required to understand the text. Higher scores indicate that more advanced education is needed to comprehend the text. For example, a score of 12 suggests a high school senior level, while 16+ suggests college graduate level. In our data, ST9 has the highest RDFKGL (18.157), suggesting it requires post-graduate level education to comprehend.

Coh-Metrix L2 Readability (RDL2) is specifically designed to assess text difficulty for second language learners. It incorporates additional linguistic features beyond sentence and word length, including cohesion, syntactic complexity, and lexical diversity. Lower RDL2 scores indicate text that would be more challenging for non-native speakers. In our dataset, ST11 has the lowest RDL2 (7.161), suggesting it would be particularly difficult for second language readers.

Overall, the readability scores indicate that the STs generally fall within the reading proficiency of college to graduate students who are native English speakers and are relatively difficult to comprehend by L2 English speakers (Flesch, 1979).

The STs were translated by ChatGPT-4o and Deepseek-V3 with the same prompt: "You are a professional translator working with academic texts. Translate this from English to Chinese: ST". We chose the models for two reasons. First, they are likely to be accepted by users in need of English-Chinese machine translation. In a recent survey among Chinese professional translators, 75.5% respondents reported use of ChatGPT as translation aid (Shi et al., 2024). Deepseek-V3 is reported to achieve "performance comparable to leading closed-source models, including ChatGPT-4o and Claude-3.5-Sonnet, on a series of standard and open-ended benchmark" and "surpasses these models in Chinese factual knowledge" (Liu et al., 2024), showing its potential in conducting English-to-Chinese translation tasks.

Second, they can generate the entire target text more effectively than newer reasoning models. In our pilot study, we tested the state-of-the-art reasoning model Deepseek-R1, released on January 20th. We observed that the R1 model tended to omit parts of sentences or entire sentences to a large extent in its translations. As shown in Table 2, with the same set of STs, the total target tokens generated by DeepSeek-R1 amount to only 59.87% of the mean total target tokens generated by ChatGPT-4o and DeepSeek-V3. The inclination of reasoning-capable language models to omit sentences and introduce creative elements during translation may arise from a fundamental tension between their reasoning abilities and the need for translation fidelity. As explored by He et al. (2024) in their study on human-like translation strategies, these models can emulate human translators by analyzing the ST and generating background knowledge to guide the translation process. This approach can lead to what He et al. (2024) describe as "creative reformulation," where the model restructures content based Liu et al. (2023) indicates that models with strong reasoning capabilities may produce "logical completions" during translation, potentially diverging from the original text in favor of outputs that the model deems more contextually appropriate or logically coherent. We plan to compare the translation performance of DeepSeek-R1 with OpenAI's recently released ChatGPT-4.5 model, which was introduced on February 27, 2025 (OpenAI, 2025). In this study, the responses of DeepSeek-R1 and DeepSeek-V3 were generated from its official website [1].

3 Automatic Quality Estimation

In this paper, we use COMET-KIWI [2] as the automatic tool for QE, as it provides a more comprehensive and linguistically informed evaluation of MT outputs. Unlike BLEU(Papineni et al., 2002), which primarily measures lexical and syntactic similarity based on n-gram overlaps, COMET-KIWI captures deeper semantic relationships between the source and translation. This ability allows it to assess meaning more effectively, making it robust to variations in word choice and paraphrasing that

[1] https://chat.deepseek.com/
[2] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

ST index	Topics	RDFRE	RDFKGL	RDL2
ST1	Sociolinguistic scales	18.206	17.774	8.086
ST2	Food translation	35.401	14.666	11.857
ST3	Thought-language identification	33.758	14.407	12.575
ST4	Deceptive communication	38.201	14.584	10.488
ST5	Neurophenomenal space	39.854	12.810	9.662
ST6	Language-thought dependency	33.222	16.249	14.120
ST7	Pictorial assertion	48.425	12.644	17.876
ST8	Translation cognition	34.419	14.422	9.437
ST9	Translanguaging	23.568	18.157	10.062
ST10	Multilingualism & ethics	30.860	15.624	9.932
ST11	Poetic Technicity	37.566	13.524	7.161
Average		33.950	14.990	11.020

Table 1: Profiling of Source Texts

ST index	ST Segment count	ST Tokens	TT Tokens_DS R1	TT Tokens_DS V3	TT Tokens_GPT-4o
ST1	70	2049	1668	1911	2057
ST2	271	7936	4034	6961	7490
ST3	650	15814	7220	13966	15620
ST4	310	8709	5401	8376	9102
ST5	347	7383	4856	6655	7048
ST6	199	6424	3961	5932	6167
ST7	418	10457	5908	9174	9521
ST8	345	8862	4976	8097	8423
ST9	317	10998	4762	10048	10511
ST10	308	8447	5148	7486	7492
ST11	263	6786	4038	5386	6191
Total	3498	93865	51972	83992	89622

Table 2: Word Counts for different LLM models

traditional metrics often fail to recognize. Additionally, COMET-KIWI can function as a reference-free QE model, meaning it does not require a high-quality reference translation for comparison. This is particularly valuable in real-world applications and low-resource language settings where reference translations may not always be available. Research has shown that COMET-based models, including COMET-KIWI, correlate more strongly with human evaluations than BLEU and other traditional metrics such as TER and METEOR(Rei et al., 2022; Agarwal and Lavie, 2008). BLEU often fails to capture fluency and adequacy effectively, especially in cases where an NMT system produces highly fluent yet paraphrased translations. In contrast, COMET-KIWI's deep learning-based approach aligns more closely with how humans assess translation quality, making it a more reliable metric.

According to the results of overall COMET scores at text level, DeepSeek-V3 (DS-V3) demonstrates superior performance overall with an average COMET score of 0.7790, compared to ChatGPT-4o (GPT-4o)'s 0.7655. This 0.0135 point advantage, while seemingly modest, is consistent across nearly all texts (10 out of 11) and suggests a meaningful difference in translation quality.

As shown in Figure 1, DS-V3 shows more consistent performance across different STs. Its scores range from 0.6847 to 0.8171, whereas GPT-4o shows greater variability, with scores ranging from 0.6514 to 0.8074. This suggests DS-V3 may offer more reliable quality across diverse topics or text complexity levels.

The relationship between text complexity (readability) metrics and translation performance of the LLMs demonstrates inconsistent patterns. As illustrated in Table 3, the weak to moderate correlations between readability metrics and COMET scores suggest that traditional measures of text difficulty for human readers do not directly translate to difficulty for LLM-powered MT. The correlation between RDFRE and DS-V3 translation performance ($r = -0.31$, $p > 0.05$) suggests a moderate negative relationship, though not statistically significant given our relatively small sample size. For GPT-4o, this correlation is even weaker ($r = -0.14$, $p > 0.05$), suggesting its performance may be influenced by different factors altogether. Similarly, the correlations between performance and other readability metrics (RDFKGL: $r = 0.25$ for DS-V3, $r = 0.17$ for GPT-4o; RDL2: $r = 0.11$ for DS-V3, $r = 0.04$ for GPT-4o) fail to reach statistical significance, further suggesting that these models may be less sensitive to surface linguistic features and more influenced by content complexity and contextual factors.

The most revealing insights emerge from examining specific STs. ST3 (Thought-language identification) exhibits the largest performance gap (0.1292) between DS-V3 and GPT-4o despite having only moderate complexity scores (RDFRE: 33.76, RDFKGL: 14.41, RDL2: 12.58). GPT-4o's dramatic underperformance on this specialized linguistic content indicates a significant weakness in handling certain conceptual ambiguity.

Conversely, ST1 (Sociolinguistic scales), the formally most complex text (lowest RDFRE: 18.206, second lowest RDL2: 8.086), shows strong and nearly identical performance from both models (DS-V3: 0.8171, GPT-4o: 0.8041). A likely explanation is that ST1's content domain may contain terminology and concepts that are well-represented in the training data of both models. Even though the text is structurally complex, the semantic content may be more accessible to these models compared to other domains. Alternatively, the text may be complex but internally consistent in its terminology and logical reasoning patterns, making it more manageable for the LLMs to translate despite its high readability scores.

The correlation between performance gap and readability measures is minimal (RDFRE: $r = -0.11$; RDFKGL: $r = 0.005$; RDL2: $r = 0.05$; all $p > 0.05$), reinforcing that domain-specific knowledge rather than general readability differentiates these models' translation performance. While COMET-KIWI and similar QE metrics provide a general assessment of MT quality of both LLMs, their performance in terms of lexical diversity and syntactic complexity is worth further investigation as part of their translation quality evaluation (Yu, 2024).

4 Comparison of Lexical diversity between DS-V3 and GPT-4o

Lexical diversity refers to the variety and richness of vocabulary used in a text, representing a crucial factor in assessing translation quality (Kim, 2020). It is typically measured through metrics such as type-token ratio (TTR), moving-average TTR (MATTR), and measure of textual lexical diversity (MTLD) (McCarthy, 2005; Koizumi, 2012). Previous studies have indicated that MT systems

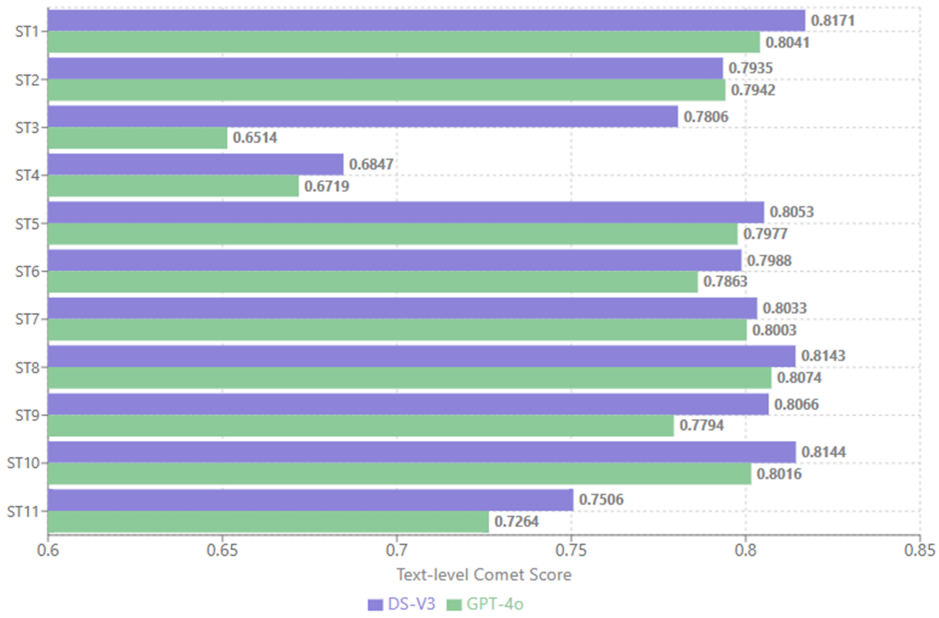

Figure 1: Comparison of Text-level Comet Score between DeepSeek-V3 (DS-V3) and ChatGPT-4o (GPT-4o)

Readability Metric	DS-V3 Score	GPT-4o Score	Performance Gap
RDFRE	-0.31 ($p > 0.05$)	-0.14 ($p > 0.05$)	-0.11 ($p > 0.05$)
RDFKGL	0.25 ($p > 0.05$)	0.17 ($p > 0.05$)	0.01 ($p > 0.05$)
RDL2	0.11 ($p > 0.05$)	0.04 ($p > 0.05$)	0.05 ($p > 0.05$)

Table 3: Correlation Between Readability Metrics and LLM Translation Quality

often produce outputs with lower lexical diversity compared to human translations, exhibiting tendencies toward vocabulary simplification and repetition(Fu and Nederhof, 2021). This phenomenon is partly due to MT systems favoring the most probable translations based on their training data, which can lead to the reduction of alternative expressions and a conformity to well-documented modes of expression. These patterns have been documented across various language pairs and domains, with earlier NMT systems particularly struggling to maintain lexical richness when translating stylistically complex texts (Brglez and Vintar, 2022).

While advancements in LLMs promise improved translation quality, their ability to maintain lexical diversity remains an area of active investigation. In this paper, we compare the lexical diversity of translation outputs produced by GPT-4o and DS-V3 at discourse level using our dataset of academic papers on topics such as language, culture, and literature. The translation of such papers demands both precision and creativity, making them ideal test sets for evaluating advanced LLMs' ability to maintain lexical richness while preserving semantic accuracy.

We first calculate traditional lexical diversity metrics using the WordSmith tool to analyze surface linguistic features in the translated texts by both GPT-4o and DS-V3. These metrics include type-token ratio (TTR), which is the ratio of unique words to total words; standardized type-token ratio (STTR), which is TTR calculated per 1,000-word segments to control for text length effects; and average sentence length (ASL), measured by the word count of each target text segment corresponding to a source text sentence.

To provide a more comprehensive assessment of lexical diversity, we also incorporate word entropy (WE) as an additional indicator. While TTR and STTR measure vocabulary diversity based on the proportion of unique words relative to total words, they do not capture the randomness and unpredictability of word usage within a text. Entropy, derived from information theory (Shannon, 1948), quantifies the degree of uncertainty in word distribution, offering a more refined understanding of lexical variation in the translated text. A higher

WE value indicates greater word unpredictability and diversity, suggesting a richer and more varied vocabulary, though potentially increasing reading difficulty (Liu et al., 2022).

Our results of lexical diversity metrics, including TTR, STTR, ASL, and WE show that GPT-4o consistently outperforms DS-V3 across all four lexical diversity metrics when translating English academic papers on language, culture, and literature into Chinese. GPT-4o shows a notably higher TTR (approximately 21.80% vs 19.99% for DS-V3), which suggests its translations contain a richer vocabulary variety. The average 9.06% higher TTR indicates GPT-4o produces less repetitive translations than DS-V3 in our dataset.

In terms of STTR, GPT-4o maintains an advantage (approximately 42.24 vs 39.08 for DS-V3). The 8.08% higher STTR confirms that GPT-4o's higher lexical diversity is consistent even when controlling for text length.

For ASL, GPT-4o produces consistently longer sentences (around 26.5 words compared to 22.3 words for DS-V3). This nearly 19% difference in average sentence length aligns with our findings in Table 2, which shows that when translating the same set of STs, GPT-4o generates 89,622 target tokens, while DS-V3 produces 83,922 tokens—representing 6.79% more target tokens overall. This consistent pattern across both sentence-level and document-level measurements reinforces the observation that GPT-4o tends to create more elaborated translations.

Regarding WE, GPT-4o demonstrates slightly higher word entropy (approximately 5.85 vs 5.74 for DS-V3). The 1.78% higher WE indicates GPT-4o translations have a more balanced distribution of word frequencies, leading to a richer and more varied vocabulary. This suggests that GPT-4o tends to produce more information-rich content with less predictable word choices.

As illustrated in the radar chart of Figure 2, the differences between these two models are not uniform across all metrics. The most pronounced differences appear in ASL and TTR, while WE shows the smallest difference. This pattern not only suggests that GPT-4o employs a broader vocabulary range, but may also indicate that its translation approach preserves more complex sentence structures of the original texts than DS-V3. We will further examine this possibility through syntactic complexity metrics in the following section.

5 Comparison of syntactic complexity between translations of DS-V3 and GPT-4o

In this study, we computed five syntactic metrics for each of the 3,498 segments in our dataset, including the Incomplete Dependency Theory Metric (IDT), Dependency Locality Theory Metric (DLT), Combined IDT+DLT Metric (IDT+DLT), Left-Embeddedness (LE), and Nested Nouns Distance (NND). The IDT, DLT, and IDT+DLT metrics are based on linguistic complexity theories derived from Gibson's Incomplete Dependency Theory (IDT) and Dependency Locality Theory (DLT) (Gibson, 1998; Gibson et al., 2000). The LE metric is adapted with slight modifications from CohMetrix analysis (Graesser et al., 2011), while NND was introduced by Zou et al. (2021). LE and NND were selected due to their relevance in capturing syntactic differences between English and Chinese (Fang, 2020).

Unlike previous studies that utilize categorical proof nets (Moot and Retoré, 2012) for syntactic representation, these metrics adopt universal dependencies (De Marneffe et al., 2021). This framework provides a consistent approach to annotating grammar, encompassing part-of-speech tagging, morphological features, and syntactic dependencies. Additionally, Blache's reformulation of Incremental Dependency Theory (IDT) and Dependency Length Theory (DLT) (Blache, 2011a,b) is applied for analyzing dependency relations. To parse segments in our dataset, Stanford Stanza (Qi et al., 2020) is employed to generate dependency trees. The decision to use dependency tree parsing over categorical proof nets is primarily driven by the availability of high-quality, scalable parsers like Stanza, which support a wide range of languages. Furthermore, previous research has demonstrated that dependency trees yield reliable and interpretable results, reinforcing their suitability for this study.

The definition and implementation of these metrics are detailed in Zou (2024); Zou et al. (2024). Developing these metrics allows us to assess whether GPT-4o retains more syntactic complexity of the STs compared to DS-V3. Additionally, it enables us to examine whether the syntactic complexity of the STs has a greater influence on translations generated by GPT-4o versus DS-V3 in the context of English-to-Chinese academic translations of papers on language, culture, and literature.

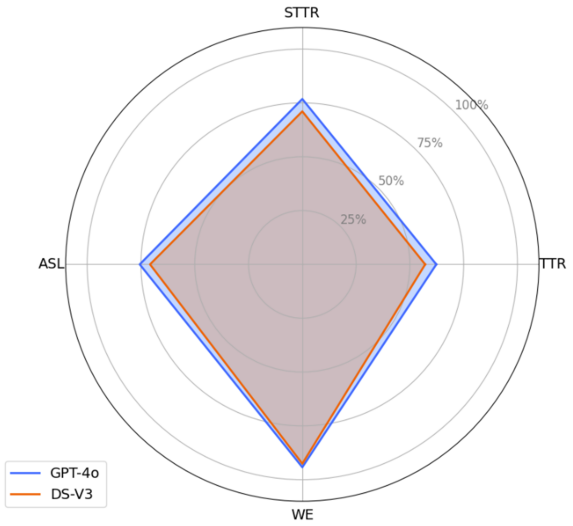

Figure 2: Comparison of Lexical Diversity Between GPT-4o and DS-V3

The results shown in Figure 3 reveal that GPT-4o consistently exhibits higher syntactic complexity than DS-V3 across all five measured metrics. These differences are statistically significant ($p < 0.05$) for all metrics, suggesting a systematic difference in the linguistic structures produced by the two models. These differences may stem from the models' underlying architectures and training data.

Incomplete Dependency Theory (IDT) counts the number of incomplete dependencies between tokens. GPT-4o shows significantly higher average IDT values (136.30) compared to DS-V3 (122.58), representing a 10.07% difference. The higher IDT values suggest that GPT-4o's outputs contain more complex syntactic structures with greater numbers of incomplete dependencies that span across the target texts. This may manifest as more sophisticated sentence constructions with multiple embedded clauses or modifying phrases.

The Dependency Locality Theory Metric (DLT) metric, which counts discourse referents (nouns, proper nouns, and verbs) between a head token and its longest leftmost dependent, is 5.28% higher in GPT-4o (13.12) compared to DS-V3 (12.43). The higher DLT values indicate GPT-4o produces target texts with longer dependency distances containing more nouns and verbs between head words and their dependents, potentially creating greater processing demands on target readers.

Not surprisingly, the combined metric shows the values of GPT-4o (149.42) are on average 9.65% higher than DS-V3 (135.01). This comprehensive measure reinforces that GPT-4o's target texts contain both more incomplete dependencies and more discourse referents within those dependencies. The significant higher values in this combined metric suggest GPT-4o's tendency toward greater overall syntactic complexity.

The Left-Embeddedness (LE) metric, which counts non-verb tokens before the main verb, shows a significant 9.89% higher value in GPT-4o (22.07) compared to DS-V3 (19.89). This difference is particularly meaningful for translations for the English-to-Chinese language pair because Chinese syntax fundamentally relies on left-embedded structures where substantial information is placed before the main verb. Topicalization in Chinese requires placing important contextual elements at the beginning of sentences, and temporal, locative, and adverbial information naturally precedes the predicate. GPT-4o's significantly higher LE score might indicate it better captures this essential characteristic of Chinese syntax, producing target text that follows the natural information flow patterns expected by the audience of native Chinese speakers.

The Nested Nouns Distance (NND) metric shows the largest percentage difference (13.20%) between the two models, with GPT-4o scoring 3.54 compared to DS-V3's 3.08. This metric is particularly relevant for English-to-Chinese translation because Chinese noun phrases follow different structural patterns than English, with modifiers strictly preceding the head noun. Chinese permits complex nested nominal structures with multiple embedded modifiers, and the distance relationships between

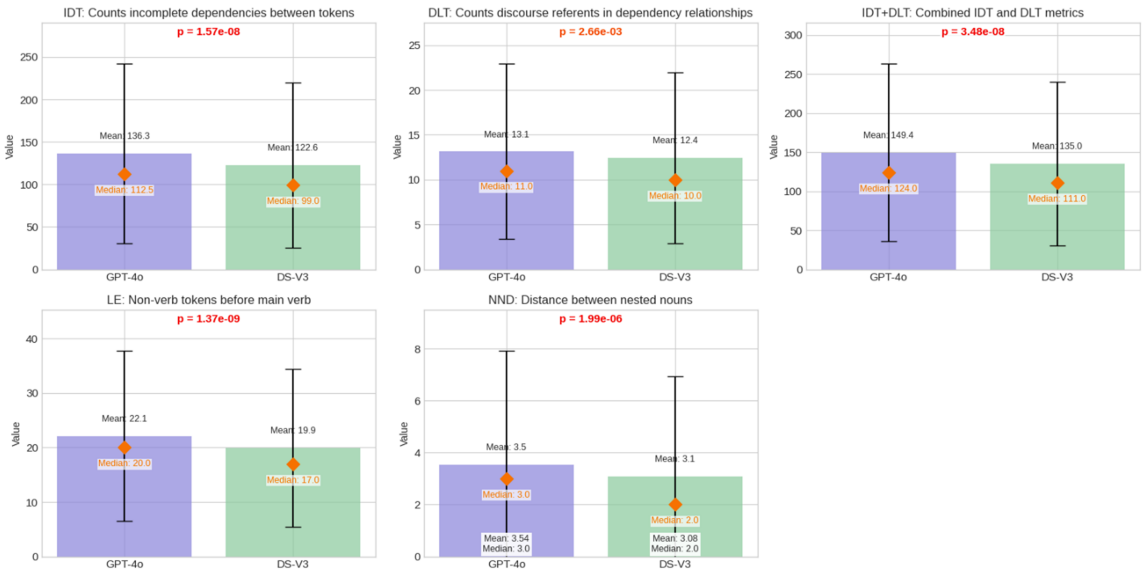

Figure 3: Comparison of Syntactic Cmplexity metrics Between Two Models

nested nouns follow language-specific conventions. GPT-4o's significantly higher NND suggests that it might better replicates the characteristic distances and relationships between nested nominal elements in Chinese. This may contribute to more authentic-sounding modifier structures and more natural nominal phrases that align with native Chinese linguistic expectations.

6 Discussion and Conclusion

This study has examined the performance of ChatGPT-4o and DeepSeek-V3 in translating English academic papers on language, literature, and culture to Chinese at the discourse level. Our analysis reveals that DeepSeek-V3 demonstrates better overall translation quality than ChatGPT-4o according to automatic quality estimation results from COMET-KIWI. DeepSeek-V3 achieved a higher average score (0.7790 versus 0.7655) and showed more consistent performance across different source texts, suggesting it may offer more reliable quality across diverse topics and complexity levels.

Interestingly, despite its lower COMET-KIWI scores, GPT-4o exhibits greater lexical richness according to all four indicators we measured. GPT-4o consistently produced translations with higher type-token ratio (21.80% versus 19.99%), standardized type-token ratio (42.24 versus 39.08), average sentence length (26.5 versus 22.3 words), and word entropy (5.85 versus 5.74). This suggests GPT-4o translations contain a more varied vocabulary and less repetitive language patterns than DeepSeek-V3.

We also found that GPT-4o's translations have greater syntactic complexity across all five segment-level metrics we examined. The higher values in Incomplete Dependency Theory (IDT), Dependency Locality Theory (DLT), combined IDT+DLT, Left-Embeddedness (LE), and Nested Nouns Distance (NND) metrics indicate GPT-4o produces more complex syntactic structures. Particularly notable are the higher scores in LE and NND, which were specifically selected as English-Chinese pair-specific metrics.

The higher syntactic complexity in LE and NND metrics might correspond to more authentic Chinese patterns, but could alternatively reflect unnecessarily complicated structures that native speakers would find awkward or unnatural. Whether this increased complexity actually results in more natural-sounding Chinese would require human evaluation by native speakers of simplified Chinese, as syntactic complexity metrics alone cannot definitively determine naturalness or fluency in the target language.

Our findings contribute to the growing body of research on what we term "GenAIese" - the distinctive linguistic characteristics of text generated by large language models. Just as previous research identified "translationese" in statistical and neural MT outputs, our study suggests that different LLM architectures may produce systematically different translation patterns that could potentially be identi-

fied using the metrics we have developed.

The divergent performance patterns between DeepSeek-V3 and GPT-4o reveal that distinct architectural foundations and training methodologies produce complementary translation strengths. This finding suggests significant opportunities for developing specialized LLM translation agents tailored to specific academic domains and communication goals. Rather than pursuing a universal translation approach, future systems could strategically leverage different architectural choices to optimize for domain-appropriate quality dimensions, whether terminological precision, syntactic naturalness, or stylistic richness. Such purpose-built translation agents could allow researchers and publishers to select models that align with disciplinary conventions, potentially transforming academic translation workflows by offering configurable balance between content fidelity and linguistic sophistication based on contextual requirements.

While our study focused on English-to-Chinese academic translation, the evaluation framework combining automatic quality assessment with lexical and syntactic complexity metrics provides a methodological foundation that can be applied to evaluate other language pairs, domains and systems. This multidimensional assessment approach addresses the limitations of relying solely on automatic metrics like COMET, which may not fully capture the linguistic qualities that contribute to translation effectiveness in specialized contexts.

For practitioners using LLMs in academic translation, our findings suggest that careful selection of models based on text characteristics is crucial. DeepSeek-V3 may be preferable for texts requiring consistent overall quality, while GPT-4o might offer advantages for texts where syntactic complexity and lexical richness are valued. Our results also suggest potential benefits in combining the strengths of different models. Practical workflows could leverage DeepSeek-V3's overall quality while selectively incorporating GPT-4o's syntactic capabilities for specific text types or sections. Research into effective human-AI collaboration protocols for academic translation could maximize the strengths of both human translators and various LLMs.

However, this study has several limitations that should be acknowledged. The sample size for discourse-level assessment is limited to 11 texts, which might explain why some of the discourse-level assessments did not reach statistical significance. Additionally, our analysis focused on non-reasoning models, which may not represent the full capabilities of the latest LLM-powered systems.

In future studies, we plan to increase the number of texts analyzed to strengthen the statistical power of our discourse-level assessments. We also intend to further investigate reasoning-capable models such as DeepSeek-R1 and ChatGPT-4.5, which may demonstrate different translation strategies and capabilities. Further investigation into how reasoning capabilities affect translation fidelity and creativity could inform model selection and development for different translation needs. Moreover, incorporating human evaluation by native speakers of simplified Chinese would provide valuable insights into the perceived naturalness and acceptability of translations with different lexical diversity and syntactic complexity profiles.

In conclusion, this study reveals an interesting integration of automatic quality estimation scores and linguistic complexity metrics in LLM-powered translation. While DeepSeek-V3 achieves higher COMET scores, GPT-4o produces translations with greater lexical diversity and syntactic complexity. These findings suggest that different evaluation methods may capture different aspects of translation quality, highlighting the need for comprehensive assessment approaches that combine automatic metrics, linguistic analysis, and human evaluation to effectively leverage LLMs for specialized translation tasks. This extends beyond the specific models evaluated, which also applies to quality assessment of translations generated by other LLMs and commercial neural machine translation systems.

References

Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118.

Hadeel M Al-Smadi. 2022. Challenges in translating scientific texts: Problems and reasons. *Journal of language teaching and research*, 13(3):550–560.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International conference on spoken language translation*, pages 280–290.

Philippe Blache. 2011a. A computational model for linguistic complexity. In *Biology, Computation and Linguistics*, pages 155–167. IOS Press.

Philippe Blache. 2011b. Evaluating language complexity in context: New parameters for a constraint-based model. In *CSLP-11, workshop on constraint solving and language processing*. Citeseer.

Mojca Brglez and Špela Vintar. 2022. Lexical diversity in statistical and neural machine translation. *Information*, 13(2):93.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? what about a GPT model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 393–417.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Jing Fang. 2020. Pause in sight translation: a pilot study. *Translation Education: A Tribute to the Establishment of World Interpreter and Translator Training Association (WITTA)*, pages 173–192.

Rudolf Flesch. 1979. How to write plain english. *University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.[Retrieved 5 February 2016]*.

Yingxue Fu and Mark-Jan Nederhof. 2021. Automatic classification of human translation and machine translation: A study from the perspective of lexical diversity. *arXiv preprint arXiv:2105.04616*.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Hoonmil Kim. 2020. The effects of lexical diversity and lexical sophistication of english on korean-english translation. *The Journal of Translation Studies ()*, 21(2):43–65.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Findings of the wmt24 general machine translation shared task: the llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Rie Koizumi. 2012. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1):60–69.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023. Crystal: Introspective reasoners reinforced with self-feedback. *arXiv preprint arXiv:2310.04921*.

Kanglong Liu, Zhongzhu Liu, and Lei Lei. 2022. Simplification in translated chinese: An entropy-based approach. *Lingua*, 275:103364.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Richard Moot and Christian Retoré. 2012. *The logic of categorial grammars: a deductive account of natural language syntax and semantics*, volume 6850. Springer.

OpenAI. 2025. Introducing GPT-4.5. OpenAI Website. Accessed: April 26, 2025.

Paperpal. 2023. How to make translating academic papers less challenging. Paperpal Blog. Accessed: April 26, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do gpts produce less literal translations? *arXiv preprint arXiv:2305.16806*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Y Shi, H Xu, HL Kwok, and K Liu. 2024. Chatgpt in professional translation: A double-edged sword—insights from chinese translators on capabilities, concerns, and future prospects. *Translation and Interpreting in the Age of Artificial Intelligence. London/New York: Routledge*, page 2.

Slator. 2024. Experts weigh in on deepseek ai translation quality. Slator. Accessed: April 26, 2025.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. *arXiv preprint arXiv:1907.00900*.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *arXiv preprint arXiv:2102.00287*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Lei Yu. 2024. ChatGPT [lexical diversity and syntactic complexity in ChatGPT translation]. *Foreign Language Teaching and Research ()*, 56(2):297–307.

Longhui Zou. 2024. *Cognitive Processes in Human-ChatGPT Interaction during Machine Translation Post-editing*. Ph.D. thesis, Kent State University.

Longhui Zou, Michael Carl, Mehdi Mirzapour, Hélène Jacquenet, and Lucas Nunes Vieira. 2021. Ai-based syntactic complexity metrics and sight interpreting performance. In *International Conference on Intelligent Human Computer Interaction*, pages 534–547. Springer.

Longhui Zou, Michael Carl, Shaghayegh Momtaz, and Mehdi Mirzapour. 2024. Impact of syntactic complexity on the processes and performance of large language models-leveraged post-editing. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 259–260.

Sustainability Statement

CO2 Emission Related to Experiments

Experiments were conducted using Google Cloud Platform in region us-east1, which has a carbon efficiency of 0.37 $kgCO_2eq/kWh$. A cumulative of 120 hours of computation was performed on hardware of type A100 PCIe 40/80GB (TDP of 250W).

Total emissions are estimated to be 11.1 $kgCO_2eq$ of which 100 percents were directly offset by the cloud provider.

Estimations were conducted using the Machine-Learning Impact calculator presented in (Lacoste et al., 2019).