ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 15, No. 12, pp. 4017-4026, December 2025
DOI: https://doi.org/10.17507/tpls.1512.21

# A BLEU-Based Evaluation of ChatGPT's Chinese-to-English Translation

Linli He
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

Mozhgan Ghassemiazghandi*
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

***Abstract*—Political text translation presents unique challenges requiring precise ideological expression, cultural sensitivity, and terminological consistency—aspects that extend beyond conventional linguistic accuracy. While ChatGPT demonstrates growing capabilities in machine translation tasks, its performance in specialized political discourse remains underexplored. This study evaluates ChatGPT's Chinese-to-English translation quality using the 2023 Chinese Government Work Report, employing both BLEU metrics and human assessment across three criteria: syntax and grammar, cultural and ideological accuracy, and fluency and coherence. Three experienced translators evaluated ChatGPT's translations using a 6-point scale, while BLEU scores provided automated evaluation. Results reveal a significant contradiction: while BLEU scores remained low (0.31-0.37), human evaluation showed moderate performance with notable variations across criteria. ChatGPT achieved the highest scores in fluency and coherence (5.53 average) but struggled significantly with cultural and ideological accuracy (4.43 average), particularly in preserving political terminology precision and contextual appropriateness. Critical issues include generic translations of politically specific terms and inadequate handling of culturally embedded expressions. The study's key finding demonstrates that BLEU evaluation alone is fundamentally insufficient for assessing political text translation quality due to single-reference constraints and inability to capture ideological nuances. Our findings highlight the limitations of BLEU in evaluating politically nuanced texts and underscore the necessity of human evaluation for meaningful assessment of specialized domain translation. This research contributes to understanding AI translation capabilities in political discourse and provides evidence-based recommendations for developing more appropriate evaluation frameworks for specialized translation domains.***

***Index Terms*—BLEU, ChatGPT, machine translation, machine translation evaluation**

## I. Introduction

ChatGPT, a conversational Large Language Model (LLM), has demonstrated remarkable capabilities in machine translation (MT) tasks despite not being specifically designed for translation purposes (Jiao et al., 2023). While empirical evidence indicates that LLMs can achieve translation quality comparable to or surpassing dedicated systems such as Google Translate and DeepL (Lee et al., 2023), significant research gaps remain in understanding their performance across specialized domains.

Currently, most MT research focuses on assessing English-to-Chinese translation (Liu & Zhu, 2023), with limited systematic examination of ChatGPT's translation quality in the reverse direction. More critically, there exists a notable scarcity of research comprehensively analyzing translation quality for Chinese-English political document translations (Liu & Zhu, 2023; Wu, 2023). This research gap is particularly significant because political texts present unique challenges that extend beyond linguistic accuracy to encompass ideological precision, cultural sensitivity, and terminological consistency (Du, 2025). Political documents, such as Government Work Reports, contain specialized terminology, complex syntactic structures, and culturally embedded expressions that test the boundaries of AI translation capabilities. Understanding how conversational AI models handle such texts is crucial for both theoretical advancement in MT research and practical applications in cross-cultural political communication.

Research Questions

RQ1: What are the performance characteristics, strengths, and limitations of ChatGPT in translating Chinese political texts to English?

RQ2: What are the relative strengths and weaknesses of BLEU evaluation versus human assessment in evaluating ChatGPT's translations of politically nuanced documents?

This study employs a dual evaluation approach combining BLEU metrics with human assessment, particularly suited for political domain translation evaluation. BLEU provides quantitative, replicable measurements essential for systematic comparison and benchmarking in MT research (Papineni et al., 2002). However, political texts demand evaluation beyond surface-level lexical similarity. The political domain requires precise ideological expression, cultural appropriateness, and

* Corresponding Author.

contextual accuracy—aspects that automated metrics cannot adequately capture. Human evaluation becomes indispensable for assessing semantic nuances, pragmatic coherence, and cultural sensitivity inherent in political discourse (Jiang et al., 2024). By comparing the insights provided by both methodologies, this study examines their respective capabilities and limitations in political text translation evaluation, rather than relying solely on either approach.

Taking the translation of the 2023 Chinese Government Work Report (CGWR) as a case study, this research evaluates ChatGPT's translation quality in political texts through this integrated assessment approach. The study's significance lies in advancing our understanding of AI translation capabilities in specialized domains and informing the development of more effective human-AI collaboration in translation processes.

The rest of the manuscript is organized as follows: Section 2 provides a comprehensive literature review examining ChatGPT's capabilities, BLEU evaluation metrics, and existing research on ChatGPT's translation performance. Section 3 provides the proposed methodology by taking into consideration the factors of automatic evaluation and human evaluation. Section 4 discusses the obtained results and their implications for both research questions, with a detailed analysis of ChatGPT's performance across different evaluation criteria. Finally, section 5 presents conclusions and highlights the future work perspective.

## II. Literature Review

### *A. ChatGPT Overview*

AI adeptly manipulates electromagnetism's characteristics for text processing, achieving remarkable success and frequently yielding results that closely resemble those achievable by human beings (Floridi, 2023). This observation highlights the potential of AI models such as ChatGPT to generate translations that closely resemble those produced by humans. However, the question remains whether this general linguistic capability translates effectively to specialized domains, particularly politically nuanced texts. The AI chatbot application ChatGPT, developed by OpenAI, an American AI laboratory, has ignited extensive scholarly discourse. While preceding Large Language Models (LLMs) demonstrated proficiency in diverse NLP tasks, ChatGPT distinguishes itself from those models by functioning as an AI chatbot designed explicitly for dialog, exhibiting notable competence in engaging human interlocutors through conversation emulation (De Angelis et al., 2023). This conversational foundation raises important questions about its suitability for formal translation tasks, particularly in contexts requiring precision and cultural sensitivity.

According to Mattas (2023), as an AI conversational model, ChatGPT, distinguished by its substantial and expansive advantages, constitutes a notable milestone in AI language processing. Nonetheless, these advancements prompt significant inquiries concerning ChatGPT's ability to address domain-specific complexities, particularly those inherent in Chinese Government Work Reports (CGWRs). ChatGPT has solidified its position as a preeminent language model on a global scale because of its noteworthy proficiency across diverse language-related tasks and benchmark assessments (Chen et al., 2021). ChatGPT employs a transformer-driven design, enabling it to grasp various languages' nuances and grammatical frameworks. This architectural foundation provides the basis for its translation capabilities, yet also reveals potential limitations when processing specialized political discourse.

The significance of ChatGPT extends beyond technical capabilities. Zhu and Wang (2023) point out that ChatGPT represents the third revolutionary product to humanity, following the internet and smartphones, potentially initiating a "Cognitive Revolution" that could reshape various industries. In contrast to conventional rule-based approaches, ChatGPT can acquire knowledge from an extensive collection of textual sources such as books, articles, and websites, producing contextually appropriate results and adhering to grammatical rules (Zhu, 2023). ChatGPT has demonstrated versatility across various domains, including education, medicine, and customer support, underscoring its broad applicability in NLP tasks (Alawida et al., 2023; Antaki et al., 2023; Ray, 2023). Despite its impressive capabilities, it is imperative to acknowledge that ChatGPT remains fundamentally an AI tool, with inherent limitations that necessitate human intervention and revision (Evans, 2023; Khoshafah, 2023).

### *B. BLEU in Machine Translation Evaluation*

The performance of MT systems has conventionally been evaluated using quantitative metrics, with the BLEU (Bilingual Evaluation Understudy) metric being one of the most widely utilized. Proposed by Papineni et al. (2002), BLEU assesses translation quality by measuring the degree of n-gram overlap between candidate translations and reference translations, while incorporating a brevity penalty to address the issue of excessively concise outputs. The BLEU metric evaluates adequacy by assessing word-level precision and fluency through the calculation of n-gram precision for n = 1, 2, 3, 4 (Lavie, 2011). According to Marie et al. (2021), BLEU scores were reported in 98.8% of MT studies, with 74.3% of these studies relying exclusively on BLEU as the sole evaluation metric for assessing MT system performance. Its replicability and simplicity have solidified its position as a benchmark in the field. However, BLEU exhibits significant limitations, primarily due to its emphasis on surface-level lexical similarity while overlooking deeper syntactic and semantic subtleties (Lee et al., 2023), which necessitates the integration of human evaluation approaches.

BLEU is categorized as a reference-based n-gram metric, similar to NIST (Doddington, 2002) and METEOR (Banerjee & Lavie, 2005), all of which emphasize surface-level lexical similarity. Although these metrics are efficient and widely used, their dependence on exact word or phrase matching limits their ability to capture deeper syntactic and semantic features in translation (Han & Lu, 2025). The implications of these limitations are particularly significant for politically

and culturally sensitive texts, where translation quality extends beyond lexical accuracy to encompass contextual appropriateness, ideological precision, and cultural sensitivity. In such contexts, context, tone, and implicit meaning are critical that surface-level metrics cannot adequately assess. To address this, Human evaluations offer essential insights into contextual and semantic accuracy, pragmatic coherence, and stylistic appropriateness, serving as a valuable complement to the quantitative analysis provided by BLEU metrics. This integrated methodology facilitates a more nuanced analysis, especially in the context of politically sensitive texts, where reliance solely on lexical alignment proves inadequate.

### *C. ChatGPT's Performance in Translation*

In the field of language translation, ChatGPT can accurately translate one language into another in real time, enabling seamless communication and exchange between different languages (Kalla & Smith, 2023). Users can input the text or voice they need to translate, and ChatGPT, a general NLP model, can automatically translate it into another language (Larroyed, 2023). This capability is particularly noteworthy given that ChatGPT was not specifically designed as a translation tool, raising important questions about the transferability of general linguistic competence to specialized translation tasks.

Empirical studies have documented particularly strong performance in several key areas. Khoshafah (2023) demonstrated that GPT-4 significantly improved ChatGPT's translation performance, rivaling commercial tools even for distant languages. ChatGPT has excelled in translation in specific fields, especially Chinese-English translation of technical, economic, and medical texts (Feng, 2024). Compared to Google Translate, ChatGPT generally produces more accurate Arabic–English translations with fewer revisions (AlAfnan, 2025). According to Rizki and Masykuroh (2024), ChatGPT's translation of Harry Potter uses a variety of translation strategies and shows high accuracy, especially in the use of related words for paraphrasing (62%). This study demonstrates the potential of ChatGPT in handling complex literary translation tasks. A comparative study by Jiang and Zhang (2024) showed that ChatGPT received a positive rating from human evaluators in Chinese diplomatic text translation when presented with contextual cues.

However, significant challenges emerge when examining ChatGPT's limitations across different contexts and domains. Jiao et al. (2023) assert that while GPT models can generate natural language, their performance in MT remains underexplored. Park and Eunsil (2023) indicate that ChatGPT possesses a greater comprehension and interpretation of the source text's meaning and purpose than current MT, despite achieving a different level of completeness than human translations. In conversational domains, ChatGPT excels by producing spoken language that is more natural and diverse (Hendy et al., 2023; Jiao et al., 2023), yet its performance in translating politically nuanced texts has not been sufficiently investigated.

The limitations of ChatGPT become particularly evident when examined systematically. While ChatGPT can correct errors, it often displays overconfidence in inaccurate outputs, excessive verbosity, and sensitivity to input phrasing (Azaria, 2022; OpenAI, 2022). Although effective for high-resource languages, performance declines for distant language pairs (Peng et al., 2023) and in specialized domains such as biomedical and ICT terminology (Zhao et al., 2023). Araújo and Aguiar (2023) insist that while ChatGPT demonstrates potential as a resource for conducting individual translation assessments, its appropriateness for extensive and rigorous evaluations may be limited. They agree that prudence is advisable when considering ChatGPT as the sole basis for translation appraisals within formal contexts. Research conducted on Vietnamese translation, ChatGPT shows no clear advantage and struggles with specialized terminology and grammatical understanding (Yang et al., 2023). Puppel (2025) evaluated ChatGPT's translation of creative texts from English to German and found that, despite high overall quality, notable errors in style and fluency remained, necessitating human post-editing. Karabayeva and Kalizhanova (2024) identified that although ChatGPT demonstrated acceptable accuracy when processing relatively straightforward textual segments, significant shortcomings were observed in its handling of translating metaphors, similes, rhymes, and segmented sentences. GPT models demonstrate a high level of translation quality in languages with abundant linguistic resources, whereas their proficiency substantially diminishes when confronted with languages characterized by a dearth of linguistic resources (Hendy et al., 2023).

Despite extensive research on ChatGPT's general translation capabilities, several critical gaps remain. First, a systematic investigation of its performance with politically nuanced texts is insufficient, as such texts demand not only linguistic accuracy but also retention of tone, intent, and contextual meaning. Second, how conversational AI models perform in formal, specialized domain translation, particularly high-stakes government documents, remains unexplored. Third, existing studies focus primarily on general-purpose translation, with limited attention to political discourse challenges, including ideological precision and cultural sensitivity.

Assessing ChatGPT's translation capabilities and limitations comprehensively is of paramount importance and necessity. This is not only an unparalleled challenge that the translation industry has to confront (Cheng et al., 2023), but also a reality that many translation students and language professionals need to be aware of. Feng (2024) points out that the translations generated by ChatGPT may reflect certain ideological biases, particularly a bias towards left-wing economic views, which challenge the objectivity of translations in politically sensitive contexts. This study investigates ChatGPT's translation performance on CGWRs through a dual evaluation approach combining BLEU metrics and human assessment. Rather than seeking correlation between these approaches, this study examines their respective capabilities and limitations in evaluating political text translation quality, providing insights into the adequacy of current evaluation methodologies for specialized translation domains.

## III. Methodology

### A. Corpus Selection and Data Collection

The CGWR is an official document of the state administrative organ of the government of the People's Republic of China. As an authoritative governmental normative document, the CGWR provides a comprehensive summary of the endeavors undertaken in the past year, spanning various facets, including politics, economy, and culture. It represents a non-literary and paradigmatic political document. The English translation of the CGWR is integral to China's external propaganda efforts, serving as a vital medium for the international community to understand the country's policy directives and economic development, highlighting the utmost importance of translation credibility and fidelity (Du, 2023). Selecting the 2023 CGWR as a case study to evaluate the translation performance of ChatGPT strengthens the analysis by enhancing its representational and authoritative qualities. The CGWR presents inherent peculiarities, such as specialized political language and specific governmental terminology. Whether undertaken by ChatGPT or through manual translation, the process entails specific difficulties and challenges.

This study selects 100 passages from the 2023 CGWR as translation materials for ChatGPT. These passages were retrieved from the official bilingual document repository of the People's Republic of China (http://www.china.org.cn/), which provides both Chinese and English versions. As shown in Table 1, we selected the first paragraph, which outlines the entire CGWR, as a representative sample for Chinese-English translation material. Due to the unique nature of the selected material, which constitutes a portion of the CGWR, only a singular official English reference is available.

Table 1
Source Text & Reference
Adapted from http://www.china.org.cn/chinese/2023-07/27/content_85189666.htm

| | |
|---|---|
| Chinese Source Text | 2022年是党和国家历史上极为重要的一年。党的二十大胜利召开，描绘了全面建设社会主义现代化国家的宏伟蓝图。面对风高浪急的国际环境和艰巨繁重的国内改革发展稳定任务，以习近平同志为核心的党中央团结带领全国各族人民迎难而上，全面落实疫情要防住、经济要稳住、发展要安全的要求，加大宏观调控力度，实现了经济平稳运行、发展质量稳步提升、社会大局保持稳定，我国发展取得来之极为不易的新成就。 |
| English Reference | The year 2022 was a year of great importance in the history of the Communist Party of China (CPC) and our country. The Party successfully convened its 20th National Congress, during which it drew up an inspiring blueprint for building China into a modern socialist country in all respects. In the face of high winds and choppy waters in the international environment and challenging tasks in promoting reform, development, and stability at home, the Party Central Committee with Comrade Xi Jinping at its core brought together the Chinese people of all ethnic groups and led them in meeting difficulties head-on. We acted on the requirements of responding effectively to Covid-19, maintaining economic stability, and ensuring security in development, and intensified macro regulation. As a result, we stabilized the economy, steadily enhanced development quality, and maintained overall social stability, securing new and hard-won achievements in China's development. |

To ensure representativeness and methodological rigor, a stratified sampling strategy was applied in selecting 100 passages from the 2023 CGWR. The number was determined as an optimal balance: fewer passages would risk insufficient thematic coverage, while substantially more passages would impose impractical demands on human evaluation without proportionally enhancing insight. Thus, 100 passages provide sufficient statistical robustness for BLEU analysis while remaining manageable for detailed human assessment. These passages were chosen based on their thematic diversity, linguistic complexity, and political significance as illustrated in Table 2. Among them, the opening paragraph was selected as a representative example for in-depth analysis. This paragraph offers a succinct overview of the report's key themes, thereby serving as an optimal context for illustrating both the translation challenges and strengths of ChatGPT. Although the analysis focuses on a single paragraph, the systematic selection of 100 passages ensures that the findings reflect a broad spectrum of the report's content and translation challenges.

Table 2
Passages Selection Criteria

| Selection Criteria | Description |
|---|---|
| Thematic Diversity | Passages covering various aspects, such as political directives, economic policies, and social achievements, are included to ensure broad content representation. |
| Linguistic Complexity | Passages with sophisticated sentence structures, specialized terminology, and idiomatic expressions characteristic of official discourse. |
| Political Significance | Passages related to significant initiatives, strategic goals, or political events, highlighting critical information. |

### B. Translation and Evaluation Protocol

In the assessment of MT quality, particularly in the context of ChatGPT, predominant research primarily relies on the utilization of automated evaluation metrics, such as BLEU (Jiao et al., 2023; Papineni et al., 2002), which is a standard quality evaluation index for MT (Wang & Mao, 2023). The validity of BLEU was confirmed through empirical evidence, demonstrating a correlation with human assessments of translation quality in natural language texts (Evtikhiev et al., 2023). Rafaeli (2021) also proves that BLEU was among the initial metrics to assert a strong correlation with human

assessments of quality, and it continues to be one of the most widely used automated and cost-effective metrics. The selection of BLEU as the automatic evaluation metric for this study is attributed to its ease of comprehension, efficiency, and comparability with other results. While acknowledging the merits of BLEU, it is imperative to integrate human evaluation into the translation quality assessment. Human evaluation can yield feedback of elevated quality by leveraging the evaluators' professional expertise, particularly in the context of divergent terminologies or cultural nuances. In contrast to human evaluation, which is costly and demands substantial human effort, automatic evaluation, while relatively lower in quality, proves cost-effective and convenient for comparing multiple systems (Chen & Cherry, 2014). Human evaluation and automatic evaluation mutually complement one another. We combine automatic evaluation and human evaluation in our study, as illustrated in Figure 1.

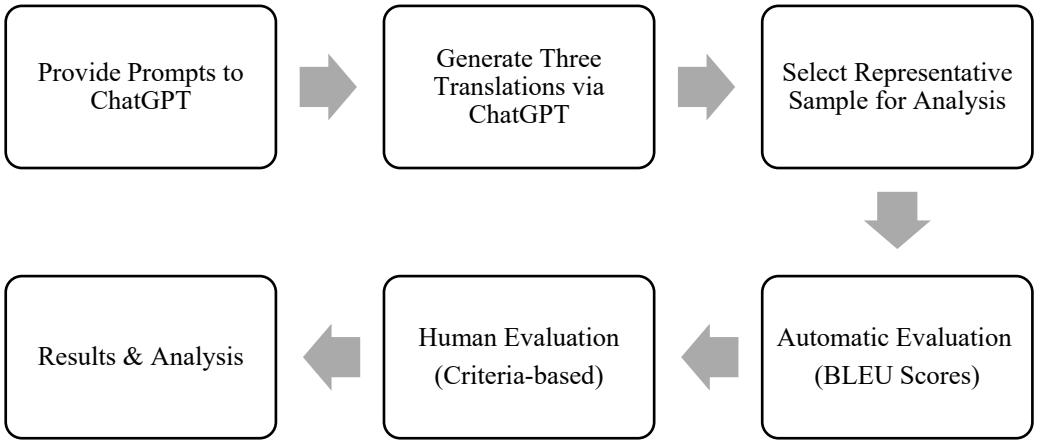


Figure 1. Translation Evaluation Methods Flow

The choice of prompts significantly shapes ChatGPT's responses. To ensure that the responses from ChatGPT remain unaffected by variations in prompts, we maintain consistency by employing identical prompts for translation tasks. We selected three translations generated by ChatGPT at different times to serve as case samples for our analysis. For human evaluation, building on the framework proposed by Araújo and Aguiar (2023), we developed a domain-specific evaluation rubric that addresses the unique requirements of political text translation, as summarized in Table 3. These criteria include syntax and grammar, cultural and ideological accuracy, fluency and coherence with scores ranging from 1 (poor) to 6 (excellent).

TABLE 3
HUMAN EVALUATION CRITERIA
Adapted from Araújo and Aguiar (2023)

| Evaluation Criteria | | Scores (1-6) |
|---|---|---|
| Syntax & Grammar | Whether there are grammatical errors in the translation, and whether the sentence structure of the translation is reasonable. | A score of 1 indicates a severe grammar error. A score of 6 means no grammatical errors. |
| Cultural & Ideological Accuracy | Whether the translation fully and accurately conveys the source information while preserving politically charged terminology, ideological nuance, and culturally embedded expressions. | A score of 1 indicates that political or cultural meaning is severely distorted, with key ideological terms mistranslated or omitted. A score of 6 means full accuracy: the translation is precise, politically appropriate, and culturally faithful to the source text. |
| Fluency & Coherence | Whether the translated sentence is fluent and coherent. | A score of 1 indicates very poor fluency and discontinuity. A score of 6 indicates excellent fluency and coherence. |

This study represents an exploratory investigation into ChatGPT's performance in specialized domain translation. While broader generalizability would benefit from expanded sampling across multiple domains and language pairs, the focused approach allows for in-depth analysis of political text translation challenges.

## IV. DISCUSSION AND RESULTS

### *A. BLEU Evaluation*

To commence, the initial step in conducting this study involves providing prompts to ChatGPT, allowing it to acquire a preliminary general understanding of its translation tasks. The researcher input prompts, providing ChatGPT with explicit cues and stipulated directives, delineating its prescribed tasks. This approach serves a dual purpose: first, to observe ChatGPT’s learning capabilities by examining its ability to accurately comprehend input prompts and standards and apply them to subsequent translations; second, to establish a controlled environment that facilitates a more favorable context for its task performance. In order to make ChatGPT get better translation results and have a general understanding

of its translation task, the context of the report is provided in detail in Table 4.

TABLE 4
PROMPTS FOR CHATGPT

| | |
|---|---|
| Prompts | The following Chinese part is the Chinese government's report on the background to the Party's 20th Congress in 2022. The 20th National Congress of the Communist Party of China (CPC) is a national congress held once every ten years to summarize the work of the previous stage and plan for future development. At the Party's 20th National Congress in 2022, the CPC Central Committee outlined a grand blueprint for comprehensively building a modern socialist country. The report sets out the main expected goals of social and economic development in 2023 and the work priorities for the current year; the report also mentions the challenges of the international environment and the importance of domestic reform, development, and stability tasks. Please translate the following Chinese sentences into English to better understand the content and context of the report. |

For the automatic evaluation of the translation of ChatGPT, we adopt the BLEU metric to compute sentence-level BLEU scores. BLEU scores are reported in normalized forms ranging from 0 to 1, with a higher proximity to 1 indicating a superior quality of the hypothesis. The BLEU scores were obtained by the use of the *sentence_bleu* function (Papineni et al., 2002) with smoothing method 7 (Chen & Cherry, 2014) from the NLTK (Natural Language Toolkit, https://www.nltk.org/) library in Python. Tokenization was performed using *nltk.word_tokenize* to ensure proper segmentation of words and punctuation. The default 4-gram setting was used without explicit modification. Specifically, we imported *sentence_bleu* from *nltk.translate.bleu_score* and applied it to hypothetical translations (ChatGPT's translations) against reference sentences to obtain sentence-level BLEU scores. The BLEU scores for the three translations of ChatGPT are presented in Table 5.

TABLE 5
BLEU SCORES

| | |
|---|---|
| T1 | 0.313 |
| T2 | 0.343 |
| T3 | 0.372 |

Analyzing the BLEU scores, it is evident that T3 achieves the highest score at 0.372, followed by T2 with a score of 0.343, while T1 attains the lowest score at 0.313. The BLEU scores, ranging from 0 to 1, show that even T3's highest score of 0.372 falls significantly below the upper limit, representing relatively low values by conventional MT evaluation standards. However, these scores primarily reflect the limitations of BLEU evaluation when applied to political discourse translation rather than indicating poor translation quality. It should be noted that BLEU in this study relies on a single reference translation, which may underestimate translation adequacy, especially in political discourse. Human assessment is integrated to establish a more comprehensive evaluation framework. Therefore, the primary evaluation focuses shifts to human assessment, with BLEU serving as a supplementary reference point while acknowledging its limitations in this specialized domain.

### *B. Human Evaluation*

In this section, three evaluators (experienced translators with expertise in political discourse) systematically assessed each of the three translations generated by ChatGPT based on the three predetermined criteria. Each passage was independently scored by all three evaluators on a 6-point scale, with final scores calculated as the average across evaluators. Inter-rater reliability was maintained through consistent application of the evaluation rubric. However, formal statistical analysis of inter-rater agreement was not conducted, representing a limitation of this study. Concurrently, a meticulous record is maintained for each translation, documenting scores obtained under each evaluation criterion, and the respective average scores for each criterion are computed. The scores are displayed below in Table 6.

TABLE 6
HUMAN EVALUATION SCORES

| Criteria | Syntax & Grammar | Cultural & Ideological Accuracy | Fluency & Coherence | Average |
|---|---|---|---|---|
| T1 | 4.67 | 4.3 | 5.33 | 4.767 |
| T2 | 5.33 | 4.33 | 5.67 | 5.11 |
| T3 | 5.67 | 4.67 | 5.6 | 5.313 |
| Average | 5.223 | 4.433 | 5.533 | |

Based on the outcomes of human assessment, it is evident that the results align with the BLEU score. T3 achieved the highest score among the three translations, with T2 following at 5.11. T1 attained the lowest score at 4.767. Additionally, an examination of the performance of each translation under distinct evaluation criteria reveals that the fluency and coherence criterion attained the highest score at 5.533. In contrast, the cultural and ideological accuracy criterion scored the lowest at 4.433, indicating this as the most challenging aspect for ChatGPT in political text translation. For syntax and grammar, the score amounted to 5.223.

*(a). Syntax and Grammar*

In the domain of syntax and grammar, an examination of the translation "2022年是党和国家历史上极为重要的一年" reveals a discrepancy in the tense employed among the translations. Both translations, T2 and T3, employ the past tense, while T1 utilizes the present tense. Given that the Government Work Report serves as a retrospective review and summarization of the government in the previous year, the usage of the past tense in T2 and T3 aligns with grammatical appropriateness. The use of the present tense in T1 is, however, misleading and has the potential to confuse the reader, as it inaccurately conveys events from the preceding year as if they are ongoing or current. It is inconsistent with the retrospective nature of the report. This issue underscores the difficulty faced by ChatGPT in accurately identifying and applying contextual temporal markers. When translating "面对风高浪急的国际环境和艰巨繁重的国内改革发展稳定任务，以习近平同志为核心的党中央团结带领全国各族人民迎难而上", T3 translated as "Confronted with a challenging international environment and the arduous tasks of domestic reform, development, and stability in the face of turbulent waters, the Party Central Committee, with Comrade Xi Jinping at its core, united and led the people of all ethnic groups across the nation to overcome difficulties", from which "in the face of turbulent waters" lacks a logically coherent syntactic relationship with the sentence body. The phrase incorporates an idiomatic expression; however, it lacks a clear grammatical connection to the subsequent clause, resulting in structural ambiguity. This issue exemplifies ChatGPT's difficulty in achieving syntactic alignment when processing complex metaphorical expressions that require precise logical integration.

*(b). Cultural and Ideological Accuracy*

Certainly, based on human evaluation, it is evident that the scores for the criterion of cultural and ideological accuracy for each of the three translations are notably low, with the average score ultimately reaching the lowest at only 4.433. Upon analyzing the three translations, it was discerned that politically charged terminology and ideological nuances were not always adequately preserved. For instance, in the translation of “2022年是党和国家历史上极为重要的一年”, the term “党” should be translated as “the Communist Party of China (CPC)” rather than simply “the party”. Given that it is the opening statement and initial mention in the political report, the complete appellation of the party should be mentioned. In the context of Chinese political discourse, the term "党" is uniquely associated with the Communist Party of China, whose identity holds a central role in the narrative of the Government Work Report. A generic translation, such as "the party", diminishes this specificity, potentially resulting in misinterpretations among target audiences lacking familiarity with the political nuances embedded in the source text. This issue underscores ChatGPT's limitations in handling politically charged terminology in the absence of explicit contextual training. Similarly, although the translations "epidemic" or "pandemic" for the term “疫情” are not linguistically wrong, their lack of explicit reference to “Covid-19” undermines the historical and political accuracy of the report. In the specific context of 2022, “疫情” refers explicitly to “Covid-19”, which had profound political, social, and economic consequences worldwide. These examples highlight ChatGPT’s limitations in handling politically sensitive texts, where cultural and ideological accuracy is as crucial as grammatical correctness or semantic fidelity.

*(c). Fluency and Coherence*

Fluency and coherence earn the highest score among the three criteria. However, the challenge also remains. The representative one is the translation of "面对风高浪急的国际环境和艰巨繁重的国内改革发展稳定任务，以习近平同志为核心的党中央团结带领全国各族人民迎难而上" which rendered as "Confronted with a challenging international environment and the arduous tasks of domestic reform, development, and stability in the face of turbulent waters, the Party Central Committee, with Comrade Xi Jinping at its core, united and led the people of all ethnic groups across the nation to overcome difficulties". “In the face of turbulent waters” exhibits a potential impact on the fluency and coherence of the entire sentence, with a lack of clarity in logical expression. This underscores the challenges encountered by ChatGPT in achieving a balance between fluency and logical coherence, especially in translations that demand the seamless integration of figurative expressions with precise contextual meaning.

In the analysis of the translation quality of ChatGPT, an observed phenomenon is the propensity for the translations generated by ChatGPT to show a degree of solidification in their outputs. The translations exhibit a marked degree of similarity. Consistencies are observed in sentence structures and vocabulary across all three translations, indicating minimal variance. Solidification refers to the phenomenon where ChatGPT exhibits a high level of repetition in its translations when tasked with repeatedly translating the exact source text. This phenomenon can be attributed to the fundamental architecture of GPT models, which are designed to prioritize the generation of fluent and statistically probable sequences grounded in patterns observed during pre-training. Although this consistency contributes to grammatical accuracy and terminological coherence, it simultaneously underscores the models’ limited capacity to accommodate contextual subtleties and adapt to discourse-specific nuances. As Hendy et al. (2023) explain, despite both NMT systems and GPT models being grounded in the transformer architecture, key distinctions exist between them. GPT models showcase exceptional proficiency in natural language generation. A decoder-only architecture characterizes GPT models, whereas NMT adopts the encoder-decoder framework. GPT models primarily undergo training on monolingual

data, while NMT models depend on extensive, meticulously curated parallel datasets. These architectural and training differences help explain why GPT models tend to produce solidified outputs, often lacking the flexibility needed to reflect pragmatic or contextual variations—an issue particularly critical in the translation of politically nuanced texts.

## V. Conclusion and Future Work

This study provides a comprehensive evaluation of ChatGPT's performance in Chinese-English political text translation, revealing both notable capabilities and significant limitations. While ChatGPT demonstrates reasonable fluency and grammatical accuracy, it struggles with political terminology precision and the preservation of cultural-ideological nuances essential for accurate political discourse translation. Human evaluation results indicate that ChatGPT performs reasonably well in syntax and grammar and in fluency and coherence, but falls short in accurately conveying political and cultural specificity. BLEU evaluation alone was found to be insufficient for assessing translation quality, as moderate scores primarily reflect the metric's limitations, including single-reference constraints, inability to assess ideological accuracy, and emphasis on surface-level lexical matching. The study also identifies ChatGPT's tendency toward output solidification, producing highly similar translations across attempts, which limits adaptability to contextual and pragmatic variations.

The findings highlight several avenues for future research, including the adoption of multi-reference BLEU evaluations and alternative metrics such as COMET and METEOR to better capture semantic and contextual accuracy. Enhancing human evaluation reliability through inter-rater consistency analysis and structured evaluator training is also recommended. Further studies should explore ChatGPT's performance across diverse political text genres and multiple language pairs, as well as domain-specific fine-tuning and prompt engineering strategies to improve political and cultural accuracy. Overall, this research underscores the value of a human-AI collaborative approach in political translation, where AI provides fluent initial translations that require human oversight, and calls for the development of specialized evaluation frameworks capable of addressing the unique demands of political discourse translation.

## Acknowledgements

The first author gratefully acknowledges her PhD supervisor, Mozhgan, for invaluable guidance throughout this research, and her family for their unwavering support.

## References

[1] AlAfnan, M. A. (2025). Large Language Models as Computational Linguistics Tools: A Comparative Analysis of ChatGPT and Google Machine Translations. *Journal of Artificial Intelligence and Technology*, *5*, 20–32. https://doi.org/10.37965/jait.2024.0549

[2] Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Isaac Abiodun, O. (2023). A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information*, *14*(8), Article 8. https://doi.org/10.3390/info14080462

[3] Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmology Science*, *3*(4), 100324. https://doi.org/10.1016/j.xops.2023.100324

[4] Araújo, S., & Aguiar, M. (2023). Comparing ChatGPT's and Human Evaluation of Scientific Texts' Translations from English to Portuguese Using Popular Automated Translators. *Notebook for the SimpleText Lab at CLEF 2023*. CEUR Workshop Proceedings.

[5] Azaria, A. (2022). *ChatGPT Usage and Limitations*. https://hal.science/hal-03913837

[6] Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, & C. Voss (Eds), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Association for Computational Linguistics. https://aclanthology.org/W05-0909

[7] Chen, B., & Cherry, C. (2014). A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, & L. Specia (Eds), *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp. 362–367). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3346

[8] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., & Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2107.03374

[9] Cheng, Y., Wang, R., Chen, J., Chao, Y., Maimaitili, A., & Zhang, H. (2023). Context-Based AI Translation From a Globalization Perspective: A Case Study of ChatGPT. *Sino-US English Teaching*, *20*(9). https://doi.org/10.17265/1539-8072/2023.09.005

[10] De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*. https://www.frontiersin.org/articles/10.3389/fpubh.2023.1166120

[11] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, 138–145.

[12] Du, H. (2023). A Corpus-Based Study on the Linguistic Features of the English Translation of the Report on the Work of the Government. *Modern Linguistics*, *11*, 2630. https://doi.org/10.12677/ML.2023.116356

[13] Du, L. (2025). *Chinese Political Discourse in Translation: A Corpus-based Critical Discourse Analysis*. Routledge. https://doi.org/10.4324/9781003544456

[14] Evans, D. (2023, January 27). *I asked ChatGPT why human intervention is so necessary and why we shouldn't be scared of progress.* LinkedIn. Retrieved July 6, 2024, from https://www.linkedin.com/pulse/i-asked-chat-gpt-why-human-intervention-so-necessary-we-darren-evans/
[15] Evtikhiev, M., Bogomolov, E., Sokolov, Y., & Bryksin, T. (2023). Out of the BLEU: How should we assess quality of the Code Generation models? *Journal of Systems and Software*, *203*, 111741. https://doi.org/10.1016/j.jss.2023.111741
[16] Feng, J. (2024). An Analysis of the Translation Output and Value Dissemination of ChatGPT. *Lecture Notes in Education Psychology and Public Media*, *35*(1), 212–218. https://doi.org/10.54254/2753-7048/35/20232108
[17] Floridi, L. (2023). AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, *36*(1), 15. https://doi.org/10.1007/s13347-023-00621-y
[18] Han, C., & Lu, X. (2025). Beyond BLEU: Repurposing neural-based metrics to assess interlingual interpreting in tertiary-level language learning settings. *Research Methods in Applied Linguistics*, *4*(1), 100184. https://doi.org/10.1016/j.rmal.2025.100184
[19] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation* [Preprint]. arXiv. http://arxiv.org/abs/2302.09210
[20] Jiang, Z., Lv, Q., Zhang, Z., & Lei, L. (2024). *Convergences and Divergences between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation* [Preprint]. arXiv.Org. https://arxiv.org/abs/2401.05176v3
[21] Jiang, Z., & Zhang, Z. (2024). *Can ChatGPT Rival Neural Machine Translation? A Comparative Study* [Preprint]. arXiv. http://arxiv.org/abs/2401.05176
[22] Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). *Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2301.08745
[23] Kalla, D., & Smith, N. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology*, *8*(3). https://ssrn.com/abstract=4402499
[24] Karabayeva, I., & Kalizhanova, A. (2024). Evaluating machine translation of literature through rhetorical analysis. *Journal of Translation and Language Studies*, *5*(1), Article 1. https://doi.org/10.48185/jtls.v5i1.962
[25] Khoshafah, Faten. (2023, April 17). *ChatGPT for Arabic-English Translation: Evaluating the Accuracy* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-2814154/v2
[26] Larroyed, A. (2023). Redefining Patent Translation: The Influence of ChatGPT and the Urgency to Align Patent Language Regimes in Europe with Progress in Translation Technology. *GRUR International*, *72*(11), 1009–1017. https://doi.org/10.1093/grurint/ikad099
[27] Lavie, A. (2011, September 19). Evaluating the Output of Machine Translation Systems. *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts*. https://aclanthology.org/2011.mtsummit-tutorials.3
[28] Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, *11*(4), Article 4. https://doi.org/10.3390/math11041006
[29] Liu, S., & Zhu, W. (2023). An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems. *Applied Artificial Intelligence*, *37*(1), 2214460. https://doi.org/10.1080/08839514.2023.2214460
[30] Marie, B., Fujita, A., & Rubino, R. (2021). *Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2106.15195
[31] Mattas, P. (2023). ChatGPT: A Study of AI Language Processing and its Implications. *International Journal of Research Publication and Reviews*, *4*, 435–440. https://doi.org/10.55248/gengpi.2023.4218
[32] OpenAI. (2022, November 30). *Introducing ChatGPT*. Retrieved June 5, 2025, from https://openai.com/blog/chatgpt.
[33] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135
[34] Park, S., & Eunsil, C. (2023). A Study of Translatability of Irony in ChatGPT. *The Journal of Translation Studies*, *24*(2), 131–160. https://doi.org/10.15749/jts.2023.24.2.005
[35] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., & Tao, D. (2023). *Towards Making the Most of ChatGPT for Machine Translation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.13780
[36] Puppel, M., & Borg, C. (2025). Evaluating ChatGPT's Performance in Creative Text Translation for Communication: A Case Study from English into German. *Media and Intercultural Communication: A Multidisciplinary Journal*, *3*(1), 1-27. https://doi.org/10.22034/mic.2024.480506.1023
[37] Rafaeli, O., Abend, O., Choshen, L., & Nikolaev, D. (2021). *Part of Speech and Universal Dependency effects on English Arabic Machine Translation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2106.00745
[38] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003
[39] Rizki, K. A. M., & Masykuroh, Q. (2024). Evaluating ChatGPT's Translation of Harry Potter: A Qualitative Study of Translation Techniques, Accuracy, and Acceptability. *JELITA*, *6*(1), Article 1. https://doi.org/10.56185/jelita.v6i1.902
[40] Wang, Z., & Mao, C. (2023). ChatGPT yiwen zhiliang de pinggu yu tisheng — yi taoci lei wenben hanying fanyi wei li [Evaluation and Enhancement of ChatGPT Translation Quality: A Case Study of Chinese-English Translation in Ceramic Texts]. *Shandong Ceramics*, *46*(4), 20–27. https://doi.org/10.3969/j.issn.1005-0639.2023.04.003
[41] Wu, J. (2023). A Comparative Analysis of Chinese-English Translation Quality Based on ChatGPT: A Case Study of Chinese Characteristic Words. *Journal of Social Science Humanities and Literature*, *6*(5), Article 5. https://doi.org/10.53469/jsshl.2023.06(05).08
[42] Yang, Y., Liu, R., Qian, X., & Ni, J. (2023). Performance and perception: Machine translation post-editing in Chinese-English news translation by novice translators. *Humanities and Social Sciences Communications*, *10*(1), Article 1. https://doi.org/10.1057/s41599-023-02285-7
[43] Zhao, Y., Zhang, J., & Zong, C. (2023). Transformer: A General Framework from Machine Translation to Others. *Machine*

*Intelligence Research*, *20*(4), Article 4. https://doi.org/10.1007/s11633-022-1393-5

[44] Zhu, G., & Wang, X. (2023). ChatGPT de yunxing moshi, guanjian jishu ji weilai tujing [ChatGPT: Operation Mode, Key Technology and Future Prospects]. *Xinjiang Normal University Journal (Philosophy and Social Sciences Edition)*, *44*(4), 113–122. https://doi.org/10.14100/j.cnki.65-1039/g4.20230217.001

[45] Zhu, P. (2023). Translation of Personal Pronouns in Government Work Report from the Perspective of Explicitation. *International Journal of Education and Humanities*, *9*(2), Article 2. https://doi.org/10.54097/ijeh.v9i2.9911

**Linli He** is a PhD candidate at the School of Languages, Literacies, and Translation, Universiti Sains Malaysia. Her research interests include machine translation and machine translation evaluation.

**Mozhgan Ghassemiazghandi**, PhD, is a Senior Lecturer at the School of Languages, Literacies, and Translation, Universiti Sains Malaysia. Her research interests include translation technology, machine translation, and audiovisual translation.