ISSN: 2710-0170

A Comparison of Mainstream Large Language Models' Performance in Chinese-to-Japanese Political Text Translation: An Empirical Analysis Based on BLEU and TER

Qi Shi, Kui Zhu*

Beijing Forestry University, Beijing 100083, China

*Corresponding author email: kuizhu2000@126.com

Abstract

With the rapid development of generative artificial intelligence, Large Language Models (LLMs) are being increasingly applied in the field of machine translation. However, their performance in high-difficulty domains such as political text translation still requires systematic evaluation. Using the Report to the 20th National Congress of the Communist Party of China (CPC) as the research corpus, this study selects four mainstream LLMs—DeepSeek, Doubao, ChatGPT, and Gemini—as research subjects. Taking the official Japanese version translated by the Institute of Party History and Literature of the CPC Central Committee as the reference text, this study quantitatively evaluates the Chinese-to-Japanese translation results of each model using two automated evaluation metrics: BLEU (Bilingual Evaluation Understudy) and TER (Translation Edit Rate), supplemented by qualitative analysis through case comparisons. The results indicate that Gemini performed best across both BLEU and TER metrics, with its translations approaching human standards in terms of structural restoration, terminology handling, and stylistic conformity. ChatGPT and DeepSeek showed moderate overall performance, with differences that were not statistically significant. Doubao performed the worst in both metrics, with primary issues concentrated in the inappropriate use of honorifics (Keigo) and the mistranslation of specific technical terms. The conclusions of this paper provide empirical evidence for the application of generative AI in professional translation and offer references for the optimization of models for political text translation in the future.

Keywords

Generative AI; Chinese-to-Japanese Translation; Political Text; Machine Translation Evaluation; BLEU; TER.

1. Research Background

According to the definition by the Institute of Linguistics of the Chinese Academy of Social Sciences, translation is the activity of expressing one language or script in another (Dictionary Editorial Office, Institute of Linguistics, CASS, 2012). As a bridge for global communication, translation plays a vital role in modern society. As early as the ancient Greek era, methods for using machines to translate natural language were proposed. At that time, people attempted to design an idealized universal language to replace the diverse and varied forms of natural languages, aiming to establish a communicative bridge between different peoples speaking different languages (Feng Zhivei, 2018).

Until the early 20th century, translation activities relied entirely on manual operations. Since the 1990s, with the rapid development of computer technology, machine translation has achieved significant breakthroughs, and its accuracy has shown a qualitative leap. In recent years, thanks to breakthrough progress in natural language processing (NLP) technology, the reliability and application prospects of machine translation have gained wide recognition in

ISSN: 2710-0170

academia. With the release of the large language model ChatGPT by the American AI company OpenAI in 2022, generative AI models entered the public eye. As technology matures, many different generative AI products have emerged in the market. They not only bring more convenience and diversity to our lives but also offer immense possibilities for enhancing productivity (Chen Yongwei, 2025).

Based on the aforementioned technologies, this study adopts a combined quantitative and qualitative research method to systematically evaluate the translation quality of four mainstream LLMs—Gemini, Doubao, ChatGPT, and DeepSeek—in Chinese-to-Japanese translation tasks. The focus is on examining their performance characteristics in the specific field of political texts, aiming to provide empirical evidence and methodological references for the expansion of generative AI applications in professional translation.

2. Previous Studies

Political discourse is a challenging area in the field of translation. Countries communicate, engage in dialogue, and exchange ideas through political discourse. In diplomatic affairs, political discourse is closely linked to national stances, international relations, and expressions of sovereignty (Zhang Quanxin, 2017). The Report to the 20th National Congress of the CPC pointed out that under the new situation, it is necessary to "accelerate the construction of Chinese discourse and Chinese narrative systems, comprehensively improve the efficiency of international communication, and form international discourse power commensurate with China's comprehensive national strength and international status." Chinese political discourse refers to "expressions with specific meanings formed by the Party and the government during the process of national governance." The political discourse system formed by China over a long period is unique; therefore, political discourse translation emphasizes the accurate grasp of context and the connotations of terms. According to research by Xie Li et al., Chinese political discourse is characterized by high induction and generalization, distinct temporal characteristics, comprehensive coverage, and popular language style (Xie Li & Wang Yinquan, 2018).

The current academic community generally adopts a research method that combines automated evaluation metrics such as BLEU and TER with human evaluation. Several recent studies have shown that in the field of political literature translation, GPT-4 demonstrates translation capabilities superior to other tools (Wen Xu & Tian Yaling, 2024), although it still falls short when handling ideology-related expressions and culture-specific items. At the linguistic structure level, this model performs exceptionally well in controlling lexical density and syntactic complexity (Yu Lei, 2024). Notably, Liu Shijie (2024) conducted a multi-dimensional evaluation by constructing professional test sets, showing significant differences between systems in terminology recognition and semantic parsing—in the translation of maritime-related texts, Wenxin Yiyan 4.0 and TranSmart performed best, while GPT-4's English-Chinese bidirectional translation capability showed clear asymmetry. Furthermore, a comparative study by Li Mei and Kong Delu (2024) found that although GPT-3.5 surpassed traditional machine translation systems in overall quality, its accuracy and consistency still require improvement. Zhang Wenyu and Zhao Bi (2024) pointed out that while the model has made progress in professional terminology and literary translation, it has not yet formed a significant advantage compared to traditional neural machine translation (NMT) technology. These studies collectively reveal the technical bottlenecks currently faced by LLMs in professional translation. Obeidat et al. (2024) compared the performance of Google Translate, ChatGPT, and Gemini in translating 155 English idioms into Arabic, finding that they primarily used literal translation, paraphrasing, and idiom-to-idiom methods. Among them, Google Translate had the highest proportion of literal translation, while Gemini performed best in

ISSN: 2710-0170

paraphrasing and idiom-to-idiom translation. Wang Yudi (2025) found through BLEU evaluation that GPT-4o performed excellently in informative text translation for Japanese-to-Chinese tasks but poorly for expressive and operative texts, with deficiencies in emotional transmission and cultural connotation handling.

3. Research Design

3.1. Research Questions

This study aims to explore the performance differences among mainstream LLMs (DeepSeek, Doubao, ChatGPT, Gemini) in the Chinese-to-Japanese translation of political texts and to test the effectiveness of BLEU and TER metrics in measuring translation accuracy, fluency, and structural restoration. Through quantitative and qualitative analysis of the translations of the 20th National Congress Report, the study further reveals the main problems of each model in terminology handling, honorific usage, and syntactic structure.

3.2. Selection of Research Subjects

The language of the 20th National Congress Report is characterized by formal structure, profound cultural connotations, deep political thought, and distinct temporal features. It widely employs political metaphors and metonyms to frame political issues and guide ideology. Given that the report is accompanied by an official Japanese translation by the Institute of Party History and Literature of the CPC Central Committee, it possesses high authority and suitability as a research corpus.

This study uses the original text of the report as the source text and the official Japanese version as the reference translation. We introduced translations generated by four mainstream LLMs (DeepSeek, Doubao, ChatGPT, Gemini), selected 20 typical sentences for comparative analysis, and evaluated the translation quality by calculating BLEU and TER values.

3.3. Core application scenarios in investment management

3.3.1. BLEU (Bilingual Evaluation Understudy)

Proposed by Papineni et al. in 2002, BLEU aims to evaluate translation accuracy by calculating the n-gram overlap between the machine translation output and the reference translation. This metric matches 1-grams to 4-grams and uses a modified precision to avoid over-penalizing short translations. Scores typically range from 0 to 1, with higher values indicating better quality. While BLEU has been criticized for lower sensitivity to semantic accuracy and fluency, its simplicity and efficiency have made it a standard metric. The formula is as follows:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{w_n} \cdot \log\ \mathrm{P_n}\right)$$

Where P_n is the n-gram precision, w_n is the weight for each n-gram (usually evenly distributed). BP (Brevity Penalty) is the penalty coefficient for short translations.

3.3.2. TER (Translation Edit Rate)

TER is a common quality metric proposed by Snover et al. in 2006. It measures the minimum number of edits required to change a machine translation into a human reference translation, normalized by the length of the reference. A lower TER value indicates better quality. The formula is:

$$TER = \frac{I + D + S + Sh}{N_{ref}}$$

Where I is insertions, D is deletions, S is substitutions, Sh is shifts, and N_{ref} is the word count of the reference.

ISSN: 2710-0170

4. Research Results

In the empirical evaluation, we conducted a detailed quantitative analysis of the performance of the four LLMs. The results clearly reveal significant differences in translation quality. All models were prompted with the same input to obtain results, which were then compared with the official Japanese translation.

With the development of economy, the global environmental problems become more and more serious, people pay more and more attention to wetlands, and wetland research has become a hot spot. As shown in Table 1.

Table 1. Comparison of BLEU and TER Scores for Model Translations

Model	NLTK BLEU	TER
ChatGPT	0.40	0.86
Deepseek	0.38	0.87
Gemini	0.50	0.71
Doubao	0.37	0.97

Note: Results are rounded to two decimal places.

The data shows that the Gemini model performed with high translation capability, leading in all indicators. Gemini achieved the highest BLEU score of 0.50, significantly higher than competitors. Simultaneously, its TER score was as low as 0.71, the lowest among all models. This indicates that Gemini's output is highly faithful to the original in both semantics and syntax, approaching human standards with minimal post-editing required.

In contrast, the Doubao model showed relatively weak performance, with the lowest BLEU (0.37) and the highest TER (0.97). This indicates that Doubao's output has the lowest match with the reference, containing more errors and deviations.

ChatGPT and DeepSeek performed at a medium level and were very close to each other. ChatGPT (BLEU 0.40, TER 0.86) slightly outperformed DeepSeek (BLEU 0.38, TER 0.87), but the differences were not statistically significant.

5. Translation Case Analysis

Due to the volume of text, we selected two representative examples to analyze the differences between model translations and human translation.

Table 2. Comparison of Model Translations (Example 1)

Type	Content
Source Text	坚持党管干部原则，坚持德才兼备、以德为先、五湖四海、任人唯贤，把新时代好干部标准落到实处。
Reference	党が幹部を管理する原則を堅持し、才徳兼備、徳の優先、全国各地から優れた人材のみを任用することを堅持し、新時代の優秀幹部基準を徹底させる。
DeepSeek	党の幹部管理原則を堅持し、徳才兼備・徳を優先・五湖四海・人材を賢く任用することを堅持し、新時代の良い幹部基準を着実に実行する。
Doubao	党が幹部を管理する原則を堅持し、徳と才能を兼ね備え、徳を優先とし、五湖四海から人材を集め、賢者を任じることを堅持し、新時代の優れた幹部の基準を着実に実行します。
ChatGPT	党が幹部を管理する原則を堅持し、徳と才を兼ね備え、徳を先と

ISSN: 2710-0170

	し、五湖四海から人材を集め、賢者を任用するという方針を、新時代の良い幹部の基準として着実に実行する。
Gemini	党が幹部を管理する原則を堅持し、徳才兼備・徳を優先し、出身を問わず賢者を登用することを堅持し、新時代の優れた幹部の基準を確実に実行する。

In Example 1, all models retained the main structure of the source text. However, there were significant differences in handling the four-character idioms. DeepSeek leaned towards literal translation (e.g., directly using 「徳才兼備」), preserving the political rhetorical style. Doubao and ChatGPT used a mix of literal translation and paraphrasing. Gemini preferred paraphrasing, translating "五湖四海" (from all corners/the five lakes and four seas) as 「出身を問わず」 (regardless of origin), which is closer to the official reference's approach of emphasizing semantic transmission over formal retention.

Notably, Doubao used the polite Keigo form (ending in 「～ます」), which does not fit the stylistic requirements of formal political texts. ChatGPT had a logical error, treating the entire first half of the sentence as the definition of the standard, rather than a parallel structure of "upholding [principles] and implementing [standards]," which could lead to ambiguity. Gemini's translation was the most refined and aligned with Japanese idiomatic expression.

Table 3. Comparison of Model Translations (Example 2)

Type	Content
Source Text	健全反制裁、反干涉、反“长臂管辖”机制。完善国家安全力量布局，构建全域联动、立体高效的国家安全防护体系。
Reference	反外国制裁・反内政干涉・反「管轄権の域外適用」の仕組みを整える、国家安全保障にかかわる配置を最適化し、各方面が連携する、立体的かつ高効率な国家安全保障維持体系を構築する。
DeepSeek	反制裁・反干涉・反「長腕管轄」メカニズムを健全にする，国家安全力の配置を完善し、全域連動・立体高效的国家安全防護体系を構築する。
Doubao	反制裁、反干涉、反「ハンドルメークス」機制を整備します，国家安全保障の力の配置を完善し、全域連動型で立体的かつ高効率な国家安全保障防護システムを構築します。
ChatGPT	反制裁、反干涉、反「長腕管轄」メカニズムを健全化する，国家安全保障力量の配置を整え、全域連動、立体的かつ効率的な国家安全防護体系を構築する。
Gemini	対制裁、対干涉、対「ロングアーム管轄」の仕組みを健全化する，国家安全保障戦力の配置を整備し、全領域が連動する、立体的で高効率の国家安全防護体系を構築する。

The core difficulty in Example 2 is "long-arm jurisdiction." DeepSeek and ChatGPT translated it literally as 「長腕」(long arm), which is uncommon in Japanese legal contexts. Doubao produced a significant mistranslation, rendering it as 「ハンドルメークス」(handle makers), which is completely irrelevant. Gemini used the transliteration 「ロングアーム管轄」, which

ISSN: 2710-0170

is common in international news, though less formal than the reference's 「域外適用」 (extraterritorial application).

The case analysis confirms the quantitative findings. Gemini's refined processing and linguistic naturalness resulted in the best scores, while Doubao's issues with honorifics and hallucinations resulted in the worst.

6. Conclusion

This study conducted quantitative and qualitative analyses of four LLMs in Chinese-to-Japanese political translation. The conclusions are as follows:

First, Gemini stood out with superior accuracy, fluency, and idiomaticity, requiring the least post-editing. Doubao was the weakest due to terminology errors and stylistic inconsistencies. ChatGPT and DeepSeek occupied the middle ground.Second, the study shows that while some models have reached a high level, human intervention and review remain essential for texts with high political sensitivity and cultural depth. LLMs currently act more as assistants than total replacements for professional human translators.Finally, while this study provides empirical evidence, it has limitations, such as sample size and a lack of multi-dimensional human evaluation. Future research should expand the corpus, introduce more diverse evaluation metrics, and explore the impact of prompt engineering on translation quality to provide more refined guidance for the professional application of LLMs.

References

[1] Chen, Y. (2023). Beyond ChatGPT: Opportunities, Risks, and Challenges of Generative AI. Journal of Shandong University (Philosophy and Social Sciences Edition), (03), 127-143.

[2] Feng, Z. (2018). Parallel Development of Machine Translation and Artificial Intelligence. Foreign Languages (Journal of Shanghai International Studies University), 41(06), 35-48.

[3] Feng, Z., Zhang, D., & Rao, G. (2023). From Turing Test to ChatGPT: Milestones and Enlightenments of Human-Computer Dialogue. Language Strategy Research, 8(02), 20-24.

[4] Wang, Y. (2025). Quality Assessment of Generative AI in Japanese-to-Chinese Translation: A Case Study of GPT-4o. Japanese Learning and Research, (02), 12-24.

[5] Xie, L., & Wang, Y. (2018). A Study of Political Discourse Translation from the Perspective of China's International Image Construction. Foreign Language Education, 39(05), 7-11.

[6] Zhang, Q. (2017). International Communication and Translation of Chinese Political Discourse. In Foreign Language Research Papers. Tianjin: School of Foreign Languages, Tiangong University, 42-47.

[7] Obeidat, M. M., et al. (2024). Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic. FWU Journal of Social Sciences, 18(4), 1-18.

[8] Papineni, K., et al. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th ACL, 311-318.

[9] Snover, M., et al. (2006). A study of translation edit rate with targeted human annotation. Proceedings of AMTA, 223-231.