Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

How Well Do Generative LLMs Translate Chinese Political Texts into English? A Multi-Metric Evaluation with Human Judgment

Liwei Zou, Yifu Wang

School of Languages and Literature, University of South China, Hengyang, Hunan, 421001, China

Abstract: Taking the 2026 New Year Address by China's president as its research corpus, this study selects ten representative sentences to evaluate English translations produced by four generative large-language models: Gemini, Doubao, DeepSeek and ChatGPT. It adopts a mixed-method evaluation framework that integrates three widely-used automatic evaluation metrics, namely METEOR, BERTScore and COMET, together with expert human scoring. On this basis, the paper makes a comparative analysis of the consistency between scores yielded by different automatic evaluation metrics and human expert judgments. The experimental results indicate that these automated metrics differ substantially in their respective evaluation priorities, and neural-network-based evaluation metrics achieve relatively higher consistency with human ratings. Within the scope of the present corpus, the comprehensive translation performance of the four large-language models is ranked in descending order as follows: ChatGPT, Doubao, DeepSeek, and Gemini.

Keywords: Machine Translation Quality Evaluation; Generative Artificial Intelligence; Political Text Translation; Automated Evaluation Metrics.

1. Introduction

As China's international exchanges continue to expand, translation has become increasingly prominent in cross-linguistic and cross-cultural communication. The process of translation not only involves linguistic shift, but also functions as an intra-lingual construction of cultural meaning and discourse stance [1]. In the context of global communication, the translation of political text poses a higher demand on accuracy, standardization and expressiveness. In recent years, the generative artificial intelligence represented by large pre-trained models has made substantial progress in the natural language generation and therefore has been applied to translation practice gradually. However, it has been found that there is still a gap between AI-generated translations and the work of professional translators [2]. However, very few empirical studies have been conducted on the systematic assessment of the actual performance of generative models in the translation of political texts, especially with clearly specified corpora. Based on the problem above, this paper integrates automatic evaluation indices and human-expert ratings. It compares the degree of agreement in outputs from automated metrics with the scores of human judgments to obtain a more objective assessment of the limits to which large language models (LLMs) can perform in translating this type of text.

Traditionally, the evaluation of machine translation was based on word matching. In recent years, the trend of research has been to construct evaluation frameworks based on semantic representations and neural networks. Therefore, one of the current problems is how closely the results of automated evaluation align with those of human evaluation. Banerjee and Lavie[3] proposed the METEOR index to better align with the results of human evaluation by accounting for word forms, word stems, synonym matches and penalizing incorrect word order.[4] Zhang and others used a pre-trained model to calculate the semantic similarity between the AI-generated translations and the reference text, and then automatically determined whether they were good based on the meaning, not just on word overlap at the surface level.[5] Rei and others proposed the COMET model. Based on all the above materials, COMET can generate scores that are close to those given by people. Based on the above, the three chosen indicators in this paper are METEOR, BERTScore and COMET. Compare the results of the evaluation indicators with human scores to determine if they are suitable for evaluating political text translation. It will also be determined how far the above-mentioned automatic tools can be used as reliable substitutes for professional human evaluation in areas with rich semantics and strict discourse norms.

2. Research Design

This paper has three related aims. First, it compares the translation quality of output produced by different large language models (LLMs) on political texts. Second, it examines the performance of the model on texts that are highly normative and stylistic. Thirdly, we should determine whether such models can be applied in the actual work of translators effectively. The four main LLMs developed in China and abroad have been chosen for empirical analysis: Gemini, Doubao, DeepSeek and ChatGPT. Comparison was made among the actual translation results produced by the various models.

The source of this study is the 2026 New Year Address, a public speech delivered by the Chinese president. Topics of discussion include national development, progress in science and technology innovation, social stability, etc. With an orderly and formal style, it is relatively brief; frequently used are four-character phrases and extended rhetorical figures. Therefore, translators must pay close attention to language details and adopt suitable styles. Ten representative sentences were taken from the whole text. The sample excerpts have various types of language, such as politically-charged conceptual terms, fixed four-character idiomatic phrases and

different rhetorical structures. The Chinese sample corpus is 512 characters long, and the official reference translation is about 410 English words. The three problems that this sampling method can help us investigate are how well a model handles political terms and expressions, and how well it can reproduce the stylistic qualities of rhetoric.

Table 1. Ten Chinese sentences

1. 我国经济总量连续跨越新关口，今年预计达到 140 万亿元，经济实力、科技实力、国防实力、综合国力跃上新台阶，绿水青山成为亮丽底色，人民群众获得感幸福感安全感不断增强。
2. 这一年，令人难忘的是，我们隆重纪念中国人民抗日战争暨世界反法西斯战争胜利 80 周年。
3. 科技与产业深度融合，创新成果竞相涌现，人工智能大模型你追我赶，芯片自主研发有了新突破，我国成为创新力上升最快的经济体之一。天问二号开启“追星”之旅，雅下水电工程开工建设。
4. 文旅市场人气火爆，“城超”“村超”热闹非凡，冰雪运动点燃冬日激情。
5. 从雪域高原到天山南北，各族群众心手相连，像石榴籽一样紧紧抱在一起，大家用洁白的哈达、热情的歌舞，表达对祖国的热爱、对幸福的礼赞。
6. 柴米油盐、三餐四季，每个“小家”热气腾腾，中国这个“大家”就蒸蒸日上。
7. 中国始终站在历史正确一边，愿同各国携手促进世界和平发展，推动构建人类命运共同体。
8. 民生无小事，枝叶总关情。
9. 锐始者必图其终，成功者先计于始。
10. 山海寻梦，不觉其远；前路迢迢，阔步而行。让我们拿出跃马扬鞭的勇气，激发万马奔腾的活力，保持马不停蹄的干劲，一起为梦想奋斗、为幸福打拼，把宏伟愿景变成美好现实。

3. Research Methods

To generate translation outputs, the four models received exactly the same set of translation instructions. We took the original Chinese passages as source inputs and the official authoritative translation as our reference benchmark. All model produced translations became the materials under evaluation. In the automated evaluation phase, we applied COMET, METEOR and BERTScore to assign numerical scores for each translation sample. These individual sentence-level scores were further pooled to calculate each model's aggregated performance over the ten test sentences.

A blind rating scheme was implemented for human assessment. Translations were randomized before being distributed to five raters who possessed advanced translation competence. When completing holistic quality assessments, raters remained unaware of the model behind each translation. Comparing metric generated scores against human ratings allows us to gain a full understanding of how LLMs cope with political text translation. Drawing on this two pronged evaluation design, we can interpret automated results in conjunction with expert opinions and build a more nuanced view of model performance in this highly norm governed translation domain.

Table 2. Official English translation

1. Our economic output has crossed thresholds one after another, and it is expected to reach RMB 140 trillion yuan this year. Our economic strength, scientific and technological abilities, defense capabilities, and composite national strength all reached new heights. Clear waters and lush mountains have become a salient feature of our landscape. Our people enjoy a growing sense of gain, happiness and security.
2. This year is full of indelible memories. We solemnly commemorated the 80th anniversary of the victory of the Chinese People's War of Resistance Against Japanese Aggression and the World Anti-Fascist War.
3. We integrated science and technology deeply with industries, and made a stream of new innovations. Many large AI models have been competing in a race to the top, and breakthroughs have been achieved in the research and development of our own chips. All this has turned China into one of the economies with the fastest growing innovation capabilities. The Tianwen-2 probe began its star-chasing journey to explore asteroids and comets. Construction of the hydropower project at the lower reaches of the Yarlung Zangbo River started.
4. The cultural and tourism sectors thrived. The "super league" football games in our cities and villages attracted numerous fans. Ice and snow sports ignited people's passion for the winter season.
5. From the snow-covered plateau to both sides of the Tianshan Mountains, people of various ethnic groups are united as one, like seeds of a pomegranate sticking together. With white khatas and passionate singing and dancing, they expressed their love of the motherland and the happiness they enjoy.
6. When the happy hum of daily life fills every home, the big family of our nation will go from strength to strength.
7. China always stands on the right side of history, and is ready to work with all countries to advance world peace and development and build a community with a shared future for humanity.
8. No issue of the people is too small; we care for every leaf and tend every branch in the garden of people's well-being.
9. A successful venture should start with a good plan and with clear goals set.
10. The dream lofty, the journey long-bold strides will get us there. Let us charge ahead like horses with courage, vitality, and energy, fight for our dreams and our happiness, and turn our great vision into beautiful realities.

4. Results

Table 3. COMET-Based Evaluation of Translation Quality Across Generative Large-Language Models

Model	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Mean Value
Gemini	0.8414	0.8369	0.7899	0.7107	0.7303	0.5585	0.8653	0.7528	0.6697	0.6950	0.74505
Doubao	0.8286	0.8304	0.8164	0.7044	0.7608	0.6564	0.9063	0.7625	0.7436	0.7194	0.77288
Deepseek	0.8329	0.8478	0.8074	0.7120	0.7589	0.6335	0.8908	0.7539	0.6635	0.7531	0.76538
ChatGPT	0.8404	0.8496	0.8144	0.7155	0.7658	0.6485	0.8941	0.7843	0.7335	0.7580	0.78041
Mean Value	0.835825	0.841175	0.807025	0.71065	0.75395	0.624225	0.889125	0.763375	0.702575	0.731375	

Table 4. METEOR-Based Evaluation of Translation Quality Across Generative Large-Language Models

Model	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Mean Value
Gemini	0.5389	0.6693	0.2680	0.2269	0.3410	0.1854	0.7391	0.3482	0.1027	0.3941	0.38136
Doubao	0.6718	0.7331	0.5250	0.1910	0.4285	0.3897	0.9506	0.2907	0.1592	0.4060	0.47456
Deepseek	0.4786	0.7655	0.4300	0.1582	0.4094	0.1991	0.8102	0.2852	0.1379	0.3589	0.4033
ChatGPT	0.6416	0.7562	0.4030	0.2631	0.4468	0.1562	0.8396	0.2516	0.1678	0.3226	0.42485
Mean Value	0.582725	0.731025	0.4065	0.2098	0.406425	0.2326	0.834875	0.293925	0.1419	0.3704	

Table 5. BERTScore-Based Evaluation of Translation Quality Across Generative Large-Language Models

Model	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Mean Value
Gemini	0.6154	0.7261	0.4193	0.3781	0.4864	0.0587	0.7823	0.4493	0.2894	0.2209	0.44259
Doubao	0.7209	0.6800	0.5366	0.4340	0.5521	0.4108	0.9248	0.4698	0.3704	0.3689	0.54683
Deepseek	0.6619	0.8266	0.5276	0.4717	0.5484	0.4134	0.9074	0.4955	0.3700	0.3568	0.55793
ChatGPT	0.6900	0.8193	0.5825	0.4278	0.5948	0.3525	0.8877	0.4609	0.3634	0.3812	0.55601
Mean Value	0.67205	0.763	0.5165	0.4279	0.545425	0.30885	0.87555	0.468875	0.3483	0.33195	

Using the corpus and research methods specified in previous sections, we performed automated evaluation over translation outputs from the four large-language models. All computed results are summarized in Tables 3–5. In the analysis that follows, we inspect score distributions across metrics and unpack model-level performance patterns, so as to unpack variations in translation quality and the degree of consistency among evaluation metrics.

4.1. Comparative Analysis of Automatic Evaluation Results for All Models and Metrics

Based on the scores at the sentence level for the ten samples, ChatGPT performed relatively consistently and had the smallest difference between high and low scores. Gemini reached a BERTScore of 0.0587 for Sentence 6 and was far below the mean. Therefore, it can be seen that the model is not performing well on rich rhetorical expressions. Sentence 6 of the original Chinese text is "柴米油盐、三餐四季，每个‘小家’热气腾腾，中国这个‘大家’就蒸蒸日上" (C#1 - Daily necessities and food, all seasons; every "little family" is warm; China as a whole is flourishing). Gemini failed to convey the metaphorical meaning of "热气腾腾" (hot and lively), and thus the translated expression was lost. Therefore, culturally-loaded linguistic items are still one of the main problems of this model.

At the level of model comparison, although the three indicators had different scores for the same source material, the general order of performance was the same. Most of the sentences that scored well according to COMET also performed well in METEOR and BERTScore. Conversely, the source materials that are rich in metaphors, condensed rhetorical expressions or dense semantics generally have low scores in all three indicators. Sentence 7 shows that this is correct; due to its fixed syntactic structure and clear political declarative purpose, LLMs had fewer difficulties with it and consistently scored high in all indicators. Sentence 6 is a contrasting case; it has concise language and culturally specific meaning that are difficult for models to learn and generate, thus scoring relatively low. Although the three indicators are based on different theoretical foundations, our results show that they can be combined reasonably well to gauge the overall quality of translation and identify the difficulties in the source text.

Based on metrics, the same LLM would obtain different scores in various evaluation tools because the tools themselves are different. COMET had a relatively high score and a small spread; most of the BERTScore results were in the middle, and METEOR had a wide distribution of scores and applied a more severe penalty for poor-quality translations. The difference is not to be taken as a failure in assessment. The Aims for Developing these Indicators are also varied. COMET realises the whole-semantic correspondence between the candidate translation and the reference text. BERTScore computes the contextual semantic similarity of words. METEOR is very sensitive to superficial lexical matches and syntactic parallelism. It is thus methodologically justifiable for multiple metrics to return numerically different scores for the same translated text. Nevertheless, automatic evaluation tools cannot fully capture subtle pragmatic meanings, cultural imagery or value-oriented judgments. To compensate for these shortcomings, human expert assessment is introduced in the present study to establish a comparative benchmark.

4.2. Inter-Rater Reliability Analysis of Human Ratings

To measure inter-rater agreement, we invited five senior academics specializing in translation studies to perform blind scoring on translations generated by the four models. Raters assessed each output along three dimensions: semantic fidelity, syntactic fluency, and stylistic appropriateness. Table 6 presents the collected human evaluation results.

Table 6. Human evaluation scores for the translation tools.

Tool	Gemini	Doubao	DeepSeek	ChatGPT
Human Translator 1	90.1	91.7	91.3	97.2
Human Translator 2	80.7	82.2	81.0	85
Human Translator 3	94.8	97.2	98.4	98
Human Translator 4	74.5	77.5	73.5	71
Human Translator 5	71.5	76	75	76
Mean Value	82.32	84.92	83.84	85.44

To test the reliability of the human ratings, inter-rater agreement was analyzed using the intraclass correlation coefficient (ICC) and Kendall's coefficient of concordance (Kendall's *W*), with the results reported in Table 7.

The reliability analysis showed that the ICC for the ratings of the five raters was 0.837 (95% CI: [0.618, 0.965]). According to the criteria proposed by Koo and Li (2016) [6], an ICC between 0.75 and 0.9 indicates good reliability, suggesting a high level of inter-rater consistency. In addition, Kendall's coefficient of concordance (Kendall's *W*) was 0.783, and the significance test yielded $p = 0.008$ ($p < 0.01$), indicating that the raters' consistency in ranking translation quality reached a statistically significant level. These results demonstrate that the human rating data obtained in this study have a robust reliability foundation and can serve as a reference standard for subsequent validity analysis.

Table 7. Results of the inter-rater reliability analysis.

Indicator	Value	95% Confidence Interval	Interpretation
ICC(2,k)	0.837	[0.618, 0.965]	Good
Kendall's *W*	0.783	χ^2=11.74, *p*=0.008	High Consistency

4.3. Criterion-Related Validity of Automatic Evaluation Metrics

To assess the criterion-related validity of the automatic evaluation metrics, Pearson correlations between the three automatic metrics and human ratings were computed. The results are reported in Table 8.

Table 8. Pearson correlations between automatic evaluation metrics and human ratings.

Automatic Metric	Pearson's r	Coefficient of Determination R^2	Significance (*p*-value)	Validity Level
COMET	0.947	0.897	0.053	Excellent *
BERTScore	0.655	0.429	0.345	Moderate
METEOR	0.389	0.151	0.611	Poor

Note: *p*=0.053 is marginally significant at the conventional α=0.05 threshold and significant at the α=0.1 level.

The results showed a very strong positive relationship between COMET-generated scores and human evaluator ratings (*r*=0.947), with a coefficient of determination (R^2) of 0.897. In practical terms, 89.7 % of the variance within human ratings can be explained by the COMET metric. Although its significance level (*p*=0.053) was only marginally significant at the conventional α=0.05 threshold, this size is already considerable, given the small sample size (*n*=4). The result points to a close agreement between COMET measurements and human expert evaluation.

BERTScore had a relatively small positive correlation with human judgments (r=0.655, R^2=0.429). The R^2 value shows that the index can account for about 42.9 per cent of the variation in human ratings, and it is relatively weak. METEOR, on the other hand, had a weak correlation with human ratings (r=0.389, R^2=0.151), explained only 15.1% of the variance, and did not reach statistical significance (*p*=0.611). The above shows that evaluation methods based mainly on the matching of words cannot reproduce the comprehensive judgments people make about the quality of translated works.

4.4. Supplementary Test of Rank Consistency

Given that the sample size is small and the relationship in the data may not be linear, Spearman's rank correlation test was used instead. The corresponding indicator for automatic evaluation is the distribution of human ranking scores for translation quality. It should be noted that with only four LLMs in the dataset (n=4), Spearman's coefficient can only be used to describe the concordance of rank order between metrics and human scores. Statistical inference based on significance testing was therefore not performed here. Full results from this supplementary test are shown in Table 9.

Table 9. Spearman correlations between automatic evaluation metrics and human ratings.

Automatic Metric	Spearman's ρ	Interpretation
COMET	1.000	Perfect Consistency
BERTScore	0.400	Low Consistency
METEOR	0.800	High Consistency

Note: Given the small sample size (*n*=4), Spearman correlations are reported solely to describe rank consistency and are not used for significance inference

The results showed that the Spearman correlation between COMET and human ratings reached ρ=1.000, indicating that its quality ranking among the four translation engines was completely consistent with the human ratings, namely ChatGPT > Doubao > DeepSeek > Gemini. The rank correlation of BERTScore with human rankings is 0.400. Although it shows a positive correlation, the consistency is moderate to low: this index correctly places ChatGPT at the top and Gemini at the bottom, but it incorrectly puts DeepSeek (human rank 3) in first and Doubao (human rank 2) in third. METEOR has a Spearman correlation of 0.800 and is in good agreement with the ranking produced by people. The main reason was that the two models had been ranked differently by the human evaluators; Doubao was ranked second, and ChatGPT took the top spot. Although there are still some inconsistencies, METEOR can still be used to compare different translation models. The above results indicate that COMET is relatively good at automated translation quality evaluation. BERTScore and METEOR are not the same in terms of the problems of rank-order consistency. Therefore, the choice of the index should be in line with the specific aims of the evaluation problem.

5. Discussion

The agreement between all the analyses of automatic metrics and human expert ratings varies greatly in this corpus. According to the data we have collected, COMET is closer to the human evaluation results than BERTScore and METEOR overall. Therefore, it can be concluded that evaluation tools based on neural network representation learning are more suitable for dealing with the quality differences in complex political-text translation. COMET's training method is to add human-evaluation signals so that its judgment logic aligns better with that of people. BERTScore is another way that considers context and meaning for similarity measurement. It can find some semantic consistency, but it is not very sensitive to small performance differences among the various large-language-model systems. METEOR, on the other hand, is based on a different principle and thus lacks the explanatory power of analysing texts with complex discourse structures and rich rhetorical devices.

The above deficiencies are due to the various reasons mentioned for each index. A single index cannot show all the difficulties in translating complex political language. To improve the reliability of the assessment results, many automatic indicators will be combined and then checked by an expert.

6. Conclusion

Given that the human-produced evaluation scores are reliable, we will now carry out criterion-related validity tests for the three main machine-translation evaluation indices. The degree of agreement between each index and human perception was different. COMET is the best overall, and BERTScore is next. Although METEOR has a small numerical correlation with human judgements, it can still be used to rank different translation systems relatively. Based on the combined evaluation results of automation and people, the performance ranking for the four generative large language models in our political-text corpus is ChatGPT > Doubao > DeepSeek > Gemini. This finding indicates that these LLMs generate translation outputs of noticeably uneven quality when handling Chinese political texts.

Several caveats should be acknowledged. The sample is limited, and our corpus only covers Chinese-English political translation. It remains unclear whether these findings hold for expanded corpora or other language combinations. Future research could examine metric behaviors and their inner working mechanisms using richer and more varied text resources. Insights from such work will help optimize metric selection for political-text translation and boost the reliability of automatic translation quality evaluation in this field.

Acknowledgments

This work was financially supported by the fund of the key scientific research project of the Hunan Provincial Education Department “Construction and Application of Human-Computer Interactive Collaborative Translation Model in the Era of ChatGPT” (No. 23A0321).

References

[1] Yuan, X. Y. (2024). Translation studies from the perspective of international communication. Journal of Foreign Languages, 47(2), 74–81.

[2] Wang, H. S., & Liu, S. J. (2022). Exploring post-editing competence in machine translation in the era of artificial intelligence. Chinese Science & Technology Translators Journal, 35(4), 21–24.

[3] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Michigan.

[4] Zhang, T., et al. (2020). BERTScore: Evaluating text generation with BERT. ICLR.

[5] Rei, R., et al. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[6] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163.