Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(8), 453-469
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Integrated Evaluation of AI-Driven MT Systems' Performance for Chinese-English Translation of China's Political Documents

Daohua Hu

School of Languages and Cultures (School of International Communication and Exchange), Shanghai University of Political Science and Law, Shanghai, China
Email: hudaohua01@126.com

How to cite this paper: Hu, D. H. (2026). Integrated Evaluation of AI-Driven MT Systems' Performance for Chinese-English Translation of China's Political Documents. *Open Journal of Social Sciences, 14,* 453-469.
https://doi.org/10.4236/jss.2026.148026

Received: July 6, 2026
Accepted: August 17, 2026
Published: August 20, 2026

Abstract

With increasingly frequent international exchanges, the translation of China's political documents has become a critical factor in conveying China's voice and shaping its international image. The development of AI-driven MT systems has profoundly transformed the language service industry and its ecology. An integrated evaluation of these MT systems' performance for Chinese-English translation of China's political documents has been carried out. This study takes "Recommendations of the Central Committee of the Communist Party of China for Formulating the 15th Five-Year Plan for National Economic and Social Development" as the source text material, with the English version released by the State Council as the reference translation. First, the translation quality of English versions generated by DeepL, Baidu, Youdao, Doubao, ChatGPT, and Gemini 3 Flash was evaluated respectively by using three automatic evaluation metrics: BLEU, TER, and METEOR. Then, based on the MQM 2.0 core framework, the same English versions were evaluated by error analysis to assess the performance of each MT system. The results show that: 1) As for the automatic evaluation metrics, Doubao, DeepL, and Baidu ranked the top three for BLEU; the top three in ascending order were Doubao, Youdao, and DeepL for TER; Doubao, Gemini 3 Flash, and DeepL ranked the top three for METEOR. Based on the overall score of the above three metrics, the top three rankings are Doubao, DeepL, and Youdao. 2) As for MQM evaluation, all six MT systems had an error rate of 60% or above for the terminology dimension; only two MT systems had no errors in the language conventions dimension; two MT systems had no errors in the accuracy dimension; in the style dimension, one MT system had an error rate of only 4%, and others of at least above 85%. In terms of overall Chinese-English translation performance, Doubao, Youdao, and Gemini 3 Flash ranked the top three, with Doubao's perfor-

mance significantly surpassing other MT systems in the dimensions of accuracy, language conventions, and style. This study will shed light on the performance of MT systems for Chinese-English translation of China's political documents, and further research in translation studies and post-translation editing in this digital era.

Keywords

AI-Driven MT Systems, Political Documents, Translation Quality Assessment, Automatic Evaluation, MQM Framework

1. Introduction

Translation has long been an important medium for bridging linguistic and cultural gaps, facilitating the exchange of knowledge, and fostering mutual understanding between peoples. From the manual translation of ancient texts to modern technological advancements, the methods and tools used for translation have continuously evolved (Shahmerdanova, 2025).

Machine Translation (MT) has been profoundly reshaped by the rapid development of artificial intelligence and natural language processing technologies. Generative artificial intelligence (GenAI) technology, centered on large language models, has already permeated the entire process of language services, promoting the collaborative development of machine translation technology and human translation, forming a human-in-the-loop translation model (Cui, 2025). GenAI not only significantly enhances the efficiency and accuracy of translation (Dalayli, 2023; Liu, 2024), but also transforms the translation process, the participants involved, and the final product (Wang & Liu, 2026). However, no matter how people's understanding of translation expands, translation quality is always an issue that cannot be ignored and often constitutes one of the hot topics in translation studies (Sun, 2023; Lommel et al., 2013).

2. Literature Review

In the early 1970s, Holmes (2000) remarked that "the level of such criticism is today still frequently very low, … Doubtless the activities of translation interpretation and evaluation will always elude the grasp of objective analysis to some extent, and so continue to reflect the intuitive, impressionist attitudes and stances of the critic" (Holmes, 2000: p. 182). As Drugan (2013: p. 70) stated, research on translation quality in academia is confined to translated works while neglecting the processes and contexts of translation activity, and only validates various proposed theories and models within a relatively narrow scope.

In the 1990s, however, software localization began to emerge as a separate branch within translation, and its business-oriented requirements tended to focus on identifying and quantifying individual errors to produce a quality score that could be used for acceptance testing and to support other business decisions (Lommel

et al., 2013). The strong practical demand has prompted the development of translation quality evaluation models, typically the LISA QA Model and SAE J2450. The LISA QA Model was developed by the Localization Industry Standards Association (LISA) and is based on error counts to evaluate translation quality from both language and format aspects. Language evaluation parameters include mistranslation, accuracy, terminology, style, consistency, etc.; format evaluation parameters include layout, typography, images, illustration numbering and captions, indexes, etc. The severity of errors is further divided into minor, major, and critical, with each category assigned a certain weight. Translation quality is assessed by calculating the error scores (LISA, 2006).

SAE J2450 is a quality evaluation metric primarily used for technical translations in the automotive sector (SAE International, 2001). It is divided into seven error types: terminology errors, semantic errors, omission, structural errors, spelling errors, punctuation errors, and other errors. In the absence of other accepted models, both of these metrics were widely implemented, even for tasks beyond their original scope of application.

But LISA QA and its predecessors followed a "one-size-fits-all" approach (Marheinecke, 2016: p. 72), and cannot be used for machine translation evaluation. SAE J2450 is limited to the automotive industry and its general applicability is somewhat insufficient, although machine translation evaluation is included. The concepts, methods, and standards used by existing evaluation models vary, lacking a basis for horizontal comparison, let alone compatibility and interoperability, which to some extent causes confusion. The language service industry urgently needs a unified evaluation framework to guide practice, and the Multidimensional Quality Metrics (MQM) model emerged accordingly.

The MQM model, developed by the German Research Center for Artificial Intelligence, is one of the research outcomes of the EU-funded QTLaunchPad project and is currently being updated by the QT21 (Quality Translation 21) project. This model integrates various existing resources to form a comprehensive, open, and customizable quality assessment framework. Within this framework, it provides a hierarchical error classification system and, based on this, constructs a family of related metrics. This system can be used not only for evaluating translations but also for assessing source texts, with applications beyond the traditional translation industry, including localization and transcreation domains (Lommel & Melby, 2015).

Functionalism and Multiple Evaluations Are the Underlying Concepts of the MQM Model

First, the MQM model adopts a functionalist evaluation approach, meaning that translation quality depends on the extent to which the text fulfills its communicative purpose. Quality is the fulfillment of consumer expectations; quality applies to specific products, services, people, processes, and environments; and quality is a continuously changing state (Goetsch & Davis, 2013: p. 2). Therefore, the accu-

racy and fluency that characterize translation quality are limited by the audience and purpose, must conform to the established standards of both parties served, and must take into account the needs of the users (Koby & Melby, 2013: p. 178).

Second, the MQM model also incorporates the concept of multiple evaluations. As Geoffrey Kingscott stated, "All aspects of translation quality are relative" (1996: 138). Different client groups have different business needs, and their uses, purposes, timeframes, and budgets for translation products vary, which naturally leads to differing requirements and assessments of translation quality. The MQM model does not intend to provide a "one-size-fits-all" standard; rather, it gives users the freedom to evaluate translation quality according to different standards, levels, and granularity.

MQM 2.0 categorizes issues in its top-level structure into eight primary dimensions, namely Terminology, Accuracy, Linguistic Conventions, Style, Locale-Conventions, Audience Appropriateness, Design and Markup, and Custom (MQM Counsel). The MQM model integrates the quality evaluation of machine translation and human translation, ending the estrangement between the two in terms of evaluation methods, which is a significant contribution of the MQM model (Tian, 2020).

When conducting fidelity/fluency evaluations, these definitions are very vague, and it is difficult for evaluators to be consistent in their application (Koehn, 2010: p. 219). Therefore, evaluation methods have gradually shifted toward automation. The basic idea is to compare the similarity between machine translations and reference translation and quantify it, thereby "objectively" assessing the quality of machine translations by simulating human scoring. The main automated evaluation metrics include BLEU, TER, METEOR, and others.

Automated evaluation methods can quickly yield scores and have relative stability, so they are widely used, but their reliability has always been a subject of debate. Taking the most commonly used BLEU as an example, its main drawback is that it heavily depends on reference translation and has low correlation with human evaluation (Turian et al., 2003; Callison-Burch et al., 2006). Therefore, in its recommendations to the European Union, the QT21 project stated that "mainstream MT quality assessment methods based on automatic metrics are incompatible with the methods used for professional human translation, and typically do not reflect the needs of actual users of translation" (Melby, 2015: p. 5). In order to overcome this drawback, the MQM model introduced methodological innovations. It no longer pursues rapid automatic evaluation, but instead follows the error classification approach commonly used in human evaluation, employing a unified descriptive language to characterize the quality of machine translation and human translation, serving as an important supplement to the automatic evaluation of machine translation quality (Cornelius, 2016: p. 15).

The MQM model places machine translations and human translations on an equal footing for evaluation, helping users choose a reasonable translation approach and highlighting that the relationship between the two is cooperative ra-

ther than competitive (Melby, 2015: p. 10), which adequately addresses the aforementioned issue.

As for the automatic metrics, they can only show the quality of the translations quantitatively, but they cannot demonstrate what kind of mistakes are in the translations. The MQM framework can point out precisely what kind of mistakes are made by the MT systems, which will make up for the shortcomings of the automatic analysis. Only with the integration of automatic analysis and MQM analysis can the quality of the MT translations be evaluated more efficiently, objectively, and reliably.

3. Methodology

3.1. Data

The Chinese political document used in this study comes from the "Recommendations of the Central Committee of the Communist Party of China for Formulating the 15th Five-Year Plan for National Economic and Social Development" (short as "the Recommendations"). The specific source text is Part Three, 1078 words, with the official English translation from the State Council website serving as the reference. English translations generated by AI-driven MT systems such as DeepL[1], Baidu AI[2], Youdao AI[3], Doubao AI[4], ChatGPT 4.0[5], and Gemini 3 Flash[6], and the translations were directly obtained by inputting the Chinese source text, without any more prompts or directions, on Feb 9 or 10, 2026. No major post-editing is applied to the MT translations.

3.2. Automated Evaluation Metrics

The automatic evaluation metrics used in this study are BLEU, TER, and METEOR. BLEU (Bilingual Evaluation Understudy) evaluates the overall quality of a translation by calculating the n-gram precision between the candidate translation and one or more reference translations, combined with a length penalty factor. Its core idea is that "the closer the candidate translation is to the reference translation, the higher its quality" (Papineni et al., 2002). BLEU shows good correlation with human evaluation at both the discourse and sentence levels, and it is fast and low-cost to compute, ushering in a new era of automatic machine translation evaluation and becoming the *de facto* standard for nearly two decades. However, the BLEU metric heavily relies on the quality and quantity of reference translations and is not friendly to diverse expressions. Being based on surface-form matching, it cannot capture semantic similarity (such as synonyms or word order variations); its reliability is low for shorter translation units (such as sentences); and it

[1]https://www.deepl.com/zh/translator/l/en/en, translated on Feb. 9, 2026.
[2]https://fanyi.baidu.com/mtpe-individual/transText, translated on Feb. 10, 2026.
[3]https://fanyi.youdao.com/#/TextTranslate, translated on Feb. 10, 2026.
[4]https://www.doubao.com/chat/?channel=hw_db_csqixiang&source=hw_db_csqixiang, translated on Feb. 9, 2026.
[5]https://xsimplechat.com/chat?topic=tpc_wfJ6qWKut9oG, translated on Feb. 9, 2026.
[6]https://deepmind.google/models/gemini/flash/, translated on Feb. 10, 2026.

tends to favor "safe" translations with a high overlap of vocabulary with the reference translation, potentially penalizing reasonable creative translations.

To address the shortcomings of BLEU, the METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric introduces stem matching, synonym matching, and paraphrase matching, while balancing precision and recall, improving its correlation with human judgments. In multilingual and multidomain tests, METEOR shows a significantly higher Pearson correlation with human scores at the sentence level than BLEU, aligning more closely with how humans perceive translation quality (Banejee & Meteor, 2005). METEOR extends BLEU's concept of "co-occurrence" by proposing three modules for counting co-occurrences: First is the "exact" module, which counts co-occurrences of words that are exactly the same in the candidate and reference translations. Second is the "porter stem module," which uses the Porter stemming algorithm to count co-occurrences of word "variants" with the same stem in the candidate and reference translations, such as "happy" and "happiness." Third is the "WN synonymy module," which matches synonyms in the candidate and reference translations using the WordNet dictionary, counting co-occurrences, such as "sunlight" and "sunshine." However, METEOR relies on external dictionaries (like WordNet), making it less adaptable to low-resource languages. Its core is still word-level matching, so it has limited ability to evaluate complex syntax, semantics, and discourse coherence.

TER (Translation Edit Rate) is a metric based on edit distance. TER quantifies the minimum number of edit operations required to change a machine translation into exactly the same text as the reference translation. TER scores usually range between 0 and 1, with higher values indicating lower quality, and it can more accurately capture subtle differences in translations (Snover et al., 2006). By combining three different metrics to evaluate various aspects of machine translations, a more comprehensive and precise assessment of machine translation quality can be achieved.

3.3. Human-Evaluated MQM Metrics

MQM 2.0 is organized into eight top level dimensions: Terminology, Accuracy, Linguistic Conventions, Style, Locale Conventions, Audience Appropriateness, Design and Markup, and Custom. According to the mistake types of the MT systems, an indicator-based rather than a full-dimensional MQM assessment is carried out in this study. That is to say, we only chose four dimensions of terminology, accuracy, linguistic conventions, and style for the error analysis, with one or more representative indicators for each dimension; for details, see section 5.

4. Automated Quality Evaluation of Translations Generated by the MT Systems

This study uses the translation evaluation tool on the Shiyibao[7] website and takes the translation provided by the State Council as the reference translation. It measures the BLEU, TER, METEOR scores, and comprehensive scores of transla-

[7]https://www.shiyibao.com/tools/MTPEtest

tions generated by DeepL, Baidu, Youdao, Doubao, ChatGPT4.0, and Gemini 3 Flash, as shown in Table 1.

Table 1. BLEU, TER, METEOR scores, and overall score.

MI ID	BLEU	TER	METEOR	Overall Score
DeepL	0.7042	0.5699	0.3026	48
Baidu	0.6871	0.5981	0.2923	46
Youdao	0.6828	0.5739	0.3001	47
Doubao	0.7505	0.5347	0.3413	52
ChatGPT 4.0	0.3811	0.9529	0.2962	24
Gemini 3 Flash	0.6715	0.6026	0.3031	46
Mean	0.6462	0.6387	0.3059	43.83

As shown in Table 1, in terms of the BLEU metric, Doubao, DeepL, and Baidu ranked in the top three, while only ChatGPT 4.0 was significantly below the mean score of 0.6462. According to Zhou and Liu (2022), if the BLEU score is over 31.4%, the translation quality is fairly good. In Table 1, all the translations generated by the MT systems are quite good.

In terms of the TER metric, the top three in ascending order of scores were Doubao, Youdao, and DeepL, with only ChatGPT 4.0 significantly exceeding the mean score of 0.6387. Academia has not reached an agreement about the TER score at which translation quality can be defined as fairly well, but they are below those reported by Wen and Tian (2024) of above 0.70.

In terms of the METEOR metric, the top three scores were Doubao, Gemini 3 Flash, and DeepL, with only Baidu and ChatGPT 4.0 falling below the mean score of 0.3059.

The overall score was roughly calculated by the formula: Overall score = BLEU × 63 + METEOR × 17 − TER × 3, and my overall score was directly obtained from the shiyibao website. Based on the overall score of the above three metrics, the top three rankings are Doubao, DeepL, and Youdao.

5. Multidimensional Analysis of Translation Quality

The MQM model assigns a certain weight to each issue. In terms of severity, it is divided into none (no error, scored 0), minor (minor error, does not affect the use or understanding of the content, scored 1), major (major error, affects the use or understanding of the content but still within an acceptable range, scored 10), and critical (severe error, making the content unusable, scored 100). The MQM model is quite flexible. Users can either directly use the model's preset metrics for quality assessment or customize them. The steps are as follows: 1) Defining the evaluation criteria based on parameters; 2) Selecting the evaluation dimensions; 3) Determining the evaluation method; 4) Choosing several questions for each dimension;

5) Setting weights according to the relative importance of each question; 6) Determining the threshold for translation acceptability; and 7) Implementing the evaluation deployment (Lommel & Melby, 2015).

Based on the specific performance of translation outputs by the MT Systems, this study evaluates the translations from four dimensions: terminology, language conventions, accuracy, and style, with 25 points for terminology, 25 points for language conventions, 20 points for accuracy, and 30 points for style. Errors in each category are classified as minor or major, with a deduction of 1 point for a minor error, 5 points for a major error, and 1.5 points for a major error in style. Minor errors are those that do not impact meaning or usability; major issues are those that impact meaning or usability but do not render the text unusable.

We invited two English teachers, each holding a PhD in Linguistics and Translation Studies respectively, with over 20 years of experience in higher education and extensive experience in teaching translation courses, to evaluate independently the quality of translations generated by six MT systems based on the MQM 2.0 core standards.

Cohen's Kappa (Wieckowska et al., 2022) was employed for inter-rater agreement, which is 0.85. Using an error analysis approach, any disagreements regarding error types were resolved through consultation between the two teachers for final confirmation.

5.1. Terminology Translation Error of MT Systems

In the latest national standard "Terminology Work and Terminology Science Vocabulary" (GB/T 15237-2025/ISO 1087:2019), a term is defined as "a designation used in language to represent a general concept." Feng (2011: p. 29) believes that a term is "a conventional symbol used to express or define a specialized concept through speech or writing".

It points out that "Using Chinese theory to interpret Chinese practice, elevating Chinese theory through Chinese practice, creating new concepts, categories, and expressions that integrate China and the world, and more fully and vividly showcasing China's story and the ideological and spiritual power behind it."[8] This excerpted text contains 1087 Chinese characters and is themed "Building a Modernized Industrial System and Reinforcing the Foundations of the Real Economy," with many new concepts, expressions, and categories included. Five Chinese terms with high frequency were chosen, namely zhongguoshi xiandaihua (Chinese modernization, Term 1), weilai chanye (industries of the future, Term 2), jushen zhineng (embodied artificial intelligence, Term 3), dujiaoshou (unicorn companies, Term 4), and fuwuye (the service sector, Term 5) as indicators to measure the translation capabilities of these six MT systems.

Based on the official translation by the State Council, DeepL and Doubao performed better in the translation of these five terms, with an error rate of 60%, followed by Baidu (error rate 80%), while Youdao, ChatGPT 4.0, and Gemini 3

[8]https://www.gov.cn/xinwen/2021-06/01/content_5614684.htm

Flash had an error rate of 100% (Table 2).

Table 2. Terminology translation errors by MT systems (Total score: 25).

MT ID	Term1	Term2	Term3	Term4	Term5	Deducted scores	Error rate
DeepL	0	-1	-1	-1	0	−3	60%
Baidu	−1	−1	−1	−1	0	−4	80%
Youdao	−1	−1	−1	−1	−1	−5	100%
Doubao	0	−1	−1	−1	0	−3	60%
ChatGPT 4.0	−1	−1	−1	−1	−1	−5	100%
Gemini 3 Flash	−1	−1	−1	−1	−1	−5	100%

Taking 中国式现代化 (Zhongguoshi Xiandaihua) as an example, the official translation by the State Council is "Chinese modernization." It appears 21 times in "the Recommendations," and the translation in the 2026 Government Work Report also adopts this translation (appearing 3 times). However, for such a fairly familiar term, only DeepL and Doubao provide the correct translation, whereas three MT systems translate it as "Chinese-style modernization," and Baidu AI translates it as "the Chinese path to modernization."

Fang and Zhang (2026) collected 623 reports from 16 mainstream media sources in the UK and the US from December 1978 to December 2025, which contain the English translation of the Chinese term Zhongguoshi Xiandaihua, namely "Chinese modernization," "Chinese-style modernization," "Chinese path to modernization," and "modernization with Chinese characteristics." Among these, UK media accounted for 428 reports and US media for 195 reports, with 994,694 words and 904,022 words respectively. They found that in mainstream US media, the most frequently used translations are "Chinese modernization" and "Chinese-style modernization", whereas in mainstream UK media, "Chinese-style modernization" and "Chinese modernization" are the most commonly employed. "Chinese-style modernization" is a literal translation of Zhongguoshi Xiandaihua. The term "style" often collocates with words indicating "different, special, or unique," and the widespread dissemination and acceptance of this translation benefit from China's consistent emphasis in external discourse that Chinese-style modernization is unique—it shares the common characteristics of modernization in Western countries, as well as Chinese characteristics based on its national conditions. "Chinese modernization" can be directly translated as "China's modernization." Its frequent use in mainstream US and UK media indicates that the concise "modifier + core noun" structure is more readable and acceptable to audiences. From the perspective of international discourse dissemination, the translation as a conceptual name is easy to spread, although its literal meaning does not directly convey the semantic connotation of Zhongguoshi Xiandaihua, so the understanding of it relies on the discourse context. Of course, with deeper dissemination, this translation and its semantic connotations gradually integrate as a whole.

Li and Zhang (2026), based on multiple official documents and 187 news reports from six foreign publicity media in China, conducted a discourse analysis of the different translations and co-occurrence patterns of Zhongguoshi Xiandaihua. The study found that, when reporting on Zhongguoshi Xiandaihua, foreign publicity media in China generally chose different translations depending on the context. When Zhongguoshi Xiandaihua is used as a proper noun, it is usually translated as "Chinese modernization." When emphasizing the path or method of Zhongguoshi Xiandaihua, a translation containing "path," such as "Chinese path to modernization," is preferred. Other translations, such as "Chinese-style modernization" or "China's path to modernization," appear occasionally. Although the contexts in which they appear have slight differences, they can all be substituted with the two main translations: "Chinese modernization" and "Chinese path to modernization."

Since the concept of Zhongguoshi Xiandaihua was proposed, its English translations have included Chinese Modernization, Chinese path to modernization, Chinese-style modernization, and modernization with Chinese characteristics, etc. In order to more accurately convey the original meaning and to be easily understood by English readers, it was ultimately decided at the 20th National Congress of the CPC held in 2023 to adopt "Chinese modernization" as the official translation of Zhongguoshi Xiandaihua.

CGTN (2023) interviewed foreign expert Sean Slattery, working at the Party History and Literature Research Institute, who has participated multiple times in the translation of documents for the Chinese "Two Sessions." How was the English translation of one of this year's key terms, Zhongguoshi Xiandaihua, determined? The 20th National Congress of the CPC outlined a grand blueprint for comprehensively advancing the great rejuvenation of the Chinese nation through Chinese modernization. "Chinese-style modernization" has also become a high-frequency term during this year's Two Sessions. In previous translations of Zhongguoshi Xiandaihua, Sean and his colleagues chose the expression "Chinese modernization." This concept may seem simple, but in reality, there are many ways to translate it, such as "Chinese-style modernization" or "modernization with Chinese characteristics." We ultimately selected "Chinese modernization" as the translation, as it is closest to the original meaning while also being easily understandable to English readers. Our Chinese colleagues will first translate the content from Chinese into English, and then we will revise the English version. We will discuss various translation issues together, especially the parts that are not satisfactorily translated. For important concepts like these, we need to accurately understand their meaning and go through multiple rounds of discussion to determine the most appropriate translation.

Through the discussion above, we can conclude that the term translation needs the translator to understand the connotative meaning of the original term and its context in the source text. The translation of a term should be normalized and updated as time goes on.

5.2. Accuracy Errors of MT Systems

Accuracy has long been a critical concept in translation, for example, Tytler's first principle for translation that "Translation should give a complete transcript of the ideas of the original work" (Tytler, 1978: p. 16), or Yan's first principle for translation "Xìn (fidelity/faithfulness/trueness)" (Munday et al., 2020: p. 38). As for the translation of China's political documents, accuracy is a central and paramount requirement. Many studies have focused on the accurate translation of content or notional words in Chinese political documents, for example, new Chinese concepts and expressions (Xiong, 2018), but little attention has been paid to the translation of function or grammatical words.

In this research, we chose a Chinese function word 等 (děng, etc./ et al.), which belongs to the subcategory of auxiliary words. First, it is used at the end of a list of things or persons, but the list is not exhaustive; second, it is used at the end of a list of things or persons to show complete enumeration. The Chinese character děng in this political document is for the first usage, which means that the list is not exhaustive.

Table 3. Translation errors for děng in the MT systems (Total score: 20).

MT ID	děng1	děng2	děng3	děng4	Deducted Score	Error Rate
DeepL	−5	−5	−5	−5	−20	100%
Baidu	0	−5	0	−5	−10	50%
Youdao	0	0	0	0	0	0%
Doubao	0	0	0	0	0	0%
ChatGPT 4.0	−5	0	0	−5	−10	50%
Gemini 3 Flash	−5	0	0	−5	−10	50%

As shown in Table 3, only Youdao and Doubao have accurately understood and translated all four instances of děng in the text (error rate 0%), while DeepL completely omitted the four occurrences of děng, resulting in inaccurate semantic translation (error rate 100%); Baidu AI, ChatGPT 4.0, and Gemini 3 Flash correctly translated only two instances of děng (error rate 50%).

ST：巩固提升矿业、冶金、化工、轻工、纺织、机械、船舶、建筑等产业在全球产业分工中的地位和竞争力。

We should upgrade key industries and consolidate and enhance the position and competitiveness of China's industries such as mining, metallurgy, chemical industry, light industries, textiles, machinery, vessels, and construction in the global industrial division of labor.

DeepL TT: consolidating and enhancing the position and competitiveness of the mining, metallurgy, chemical, light industry, textiles, machinery, shipbuilding, and construction industries.

Baidu TT: consolidate and enhance the global industrial division of labor and competitiveness in mining, metallurgy, chemicals, light industry, textiles, machinery, shipbuilding, construction, and other sectors.

Youdao TT: Consolidate and enhance the positions and competitiveness of industries such as mining, metallurgy, chemicals, light industry, textiles, machinery, ships, and construction in the global industrial division.

Doubao TT: consolidate and elevate the status and competitiveness of mining, metallurgy, chemicals, light industry, textiles, machinery, shipbuilding, construction, and other industries in the global industrial division of labor.

ChatGPT 4.0 TT: consolidating and elevating the status and competitiveness of mining, metallurgy, chemicals, light industry, textiles, machinery, shipbuilding, and construction in global industrial divisions.

Gemini 3 flash TT: consolidating and elevating the status and competitiveness of mining, metallurgy, chemicals, light industry, textiles, machinery, shipbuilding, and construction in global industrial divisions.

The official translation uses "such as" to list eight industries, but the list is not exhaustive; in other words, these eight industries are literally emphasized, but many other industries should also be included. Baidu's translation, in addition to listing the eight industries, adds "and other sectors," while Doubao adds "and other industries" after the eight industries, both fully conveying the meaning of děng in the original text. Youdao uses a listing structure with "such as" to express the meaning of děng. DeepL, ChatGPT 4.0, and Gemini 3 Flash, however, only translate these eight industries, omitting the meaning of děng, therefore narrowing the range of industries that should be upgraded and enhanced.

5.3. Language Conventions Errors by MT Systems

In the MQM framework, Language Conventions focus on the linguistic well-formedness of the text, including problems with grammaticality, idiomaticity, and mechanical correctness. One of the sub-categories in the language conventions dimension is textual conventions, which means that when a text string (word, phrase, sentence, phrase, other) violates the text-building (discourse) norms of the target language, an error occurs.

English is a language with a clear hierarchy and distinction between subject and object. But Chinese is different; even long sentences are composed of several short clauses of comparable status, connected in a call-and-response manner, without a clear distinction between subject and object (Yu, 2010: p. 2).

In the practice of Chinese-English text translation, the behavior of switching thinking patterns involves that, in the comprehension stage, the translator operates based on the Chinese thinking pattern; in the expression stage, the translator operates based on the English thinking pattern. In other words, in the former stage, the translator should adhere to a spiral thinking pattern, while in the latter stage, adhere to a linear thinking pattern (Lv, 2011).

The title of a text conveys the core content of the article to the reader in the most concise words, and its important function is to attract the readers' attention. The title has a vocative function, so the characteristics of communicative translation (domestication) are relatively evident (Zhao, 2018: p. 126).

As for the five titles in this Chinese political document, the Chinese titles are all expressed with verb phrases, and the English translation should be in the form of nouns or gerunds according to English textual norms.

Table 4. Title translation errors of the MT system (Total score: 25).

MT ID	Title 1	Title 2	Title 3	Title 4	Title 5	Deducted Score	Error Rate
DeepL	0	−1	−1	−1	−1	−4	80%
Baidu	−1	−1	−1	−1	−1	−5	100%
Youdao	0	−1	−1	−1	−1	−4	80%
Doubao	0	0	0	0	0	−0	0%
ChatGPT 4.0	−1	−1	−1	−1	−1	−5	100%
Gemini 3 Flash	0	0	0	0	0	−0	0%

The English translation of titles is generally considered acceptable when rendered in the form of nouns or gerunds, whereas using the verb form is deemed incorrect. This study uses this criterion to determine whether English title translation conventions are met. In this text, there are five titles in total. As can be seen in Table 4, only Doubao and Gemini 3 Flash have recognized the differences between Chinese and English texts and translated the five titles with gerunds (error rate 0%); Baidu uses all five in the verb form (error rate 100%), while DeepL, Youdao, and ChatGPT 4.0 each use four in the verb form (error rate 80%).

5.4. Translation Style Errors of MT Systems

In the MQM framework, if it deviates from organizational style guides or exhibits inappropriate language style, even if it is grammatically acceptable, it is inappropriate in style. One subcategory of style error is called "awkward style," which means excessive wordiness or overly embedded clauses, often due to inappropriate retention of source text style in the target text.

English is an analytic language, where logical relationships between sentences are often made explicit through grammar, vocabulary, and other linguistic forms. In contrast, Chinese is a synthetic language, characterized by discreteness in its structure, and the logical connections between sentences are mostly achieved through semantic coherence. As a result, logical connectives are often unnecessary, making flowing sentences and subjectless sentences relatively common, particularly in Chinese political documents. Subjectless sentences are a common sentence structure in Chinese, referring to cases where the subject is omitted or missing. They are usually made up of a predicate or a predicate phrase, specifically for sentences where the exact subject cannot or does not need to be stated. In translating Chinese subjectless sentences, common approaches include adding a subject, translating them into passive sentences, or turning them into existential sentences, imperative sentences, or other special sentence patterns (Li & Shen, 2017: p. 156).

In the Chinese text, there are 22 sentences except for the five titles, among which

21 are subjectless sentences, and only one is a subject-predicate sentence. In the official English translation, all the subjectless sentences are translated into subject-predicate sentences, adding "We should" at the beginning of each sentence, which emphasizes the agent and its obligations.

Table 5. Awkward style errors of the MT systems (Total score: 30).

	No. of Sentences	No. of Subject-Predicate Sentences	No. of imperative sentences	Deducted Score	Error Rate
DeepL	26	2	24	−24	92%
Baidu	26	2	24	−24	92%
Youdao	24	2	21	−21	88%
Doubao	24	23	1	−1	4%
Chatgpt 4.0	26	4	22	−22	85%
Gemini 3 Flash	28	4	24	−24	86%

As shown in Table 5, Doubao transformed the Chinese subjectless sentences into English subject-predicate sentences except for one sentence. All other MT systems retained most of the Chinese source text style in the English target text, translating them into imperative sentences, and failed to translate most of the Chinese subjectless sentences into English subject-predicate sentences, and the error rate is at least above 85%.

Table 6. MQM errors of the MT systems.

MT ID	Terminology (Terms, 25)	Accuracy (děng, 25)	Linguistic conventions (Titles, 20)	Style (Sentence Patterns, 30)	Deducted Score
DeepL	−3	−20	−4	−24	−51
Baidu	−4	−10	−5	−24	−43
Youdao	−5	0	−4	−21	−30
Doubao	−3	0	−0	−1	−4
Chatgpt 4.0	−5	−10	−5	−22	−42
Gemini 3 Flash	−5	−10	−0	−24	−39

The errors of MT systems were incorporated into Table 6, as it can be seen that Doubao excels significantly in four dimensions: terminology, accuracy, linguistic conventions and style, and overall performance. Youdao and Gemini 3 Flash followed Doubao in overall performance.

6. Conclusion

This study has carried out an integrated approach to evaluate the performance of six AI-driven MT systems for Chinese-English translation of China's political documents. Part three of the Recommendations has been taken as the source text,

and the English version released by the State Council as the reference translation. This study has found that: First, the English versions generated by DeepL, Baidu, Youdao, Doubao, ChatGPT, and Gemini 3 Flash were evaluated respectively by using three automatic evaluation metrics: BLEU, TER, and METEOR. Second, based on the MQM 2.0 core framework, the same English versions were evaluated by error analysis to assess the performance of each MT system. Doubao, Youdao, and Gemini 3 rank in the top three in overall performance in terms of the four dimensions of terminology, accuracy, linguistic conventions, and style. Third, Doubao excels significantly in overall performance and each of the four dimensions. Youdao performs well in the accuracy dimension, and Gemini 3 Flash performs well in the linguistic conventions dimension.

Just as Zhang (2020) points out, machine translation technology is essentially the result of human intellectual development, and its progress relies on the accumulation and summarization of existing human knowledge, experience, and technological achievements, so the creativity of translators will be a technical challenge that machine translation will find difficult to overcome.

This study will shed light on the performance of MT Systems for Chinese-English translation of China's political documents, and further research in translation studies and post-translation editing in this digital era. Collaboration between human-MT workflow will be needed to facilitate translation efficiency and translation quality as well.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Banejee, S., & Lavie, A. (2005). *Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* IEEvaluation@ACL.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-Evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 06) (pp. 249-256). Association for Computational Linguistics (ACL).

CGTN (2023). *20th CPC National Congress: Irish Foreign Expert Understands Party's Value in Copy Editing English Version of Xi's Work Report.* https://news.cgtn.com/news/2022-10-17/VHJhbnNjcmlwdDY4ODIx/index.html

Cornelius, E. (2016) Potential Impact of QT21. In *Proceedings of the 38th Conference Translating and the Computer* (pp. 10-18).

Cui, Q. L. (2025). Development Trends of Language Services Industry in the Era of Artificial Intelligence. *Journal of Beijing International Studies University, 47,* 62-72. (In Chinese)

Dalayli, F. (2023) Use of NLP Techniques in Translation by ChatGPT: Case Study. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC).* INCOMA Ltd.

Drugan, J. (2013). *Quality in Professional Translation: Assessment and Improvement.*

Bloomsbury.

Fang, H., & Zhang, X. (2026). The Translation, Dissemination, and Reception of "Chinese Modernization" in Mainstream British and American Media. *Observation and Ponderation, No. 3,* 65-76. (In Chinese)

Feng, Z. W. (2011). *Introduction to Modern Terminology (Revised Edition).* The Commercial Press. (In Chinese)

Goetsch, D. L., & Davis, S. (2013) *Quality Management for Organizational Excellence: Introduction to Total Quality* (7th ed.). Pearson Education Limited.

Holmes, J. (2000). The Name and Nature of Translation Studies. In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 172-185). Routledge.

Koby, G. S., & Melby, A. K. (2013). Certification and Job Task Analysis (JTA): Establishing Validity of Translator Certification Examinations. *The International Journal of Translation and Interpreting Research, 5,* 174-210. https://doi.org/10.12807/ti.105201.2013.a10

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Li, C. M., & Shen, D. M. (2017). *Detaied Study of Translation between English and Chinese.* China Ocean University Press. (In Chinese)

Li, Y. M., & Zhang, W. R. (2026). A Study of Different Translations and Co-Occurrences of "Zhongguoshi Xiandaihua" in Chinese Foreign Propaganda Media. *Foreign Language Research in Northeast Asia, No. 1,* 110-129. (In Chinese)

LISA (2006). *LISA QA Model 3.1—Assisting the Localization Development, Production and Quality Control Processes for Global Product Distribution.* http://dssresources.com/news/1558.php

Liu, S. J. (2024). Evaluating the Application Efficacy of Machine Translation in Maritime Contexts: A Rigorous Evaluation via BLEU, chrF++, and BERTScore Metrics. *Journal of Ocean University of China (Social Sciences), No. 2,* 21-31. (In Chinese)

Lommel, A. R., Burchardt, A., & Uszkoreit, H. (2013). *Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality.* https://aclanthology.org/2013.tc-1.6.pdf

Lommel, A., & Melby, A. K. (2015). *Assessing Translation Quality with Multidimensional Quality Metrics MQM.* https://www.fbcinc.com/e/LEARN/e/Translation/presentations/Wednesday/JIAM-CATT-2015-v7b.pdf

Lv, S. S. (2011). Non-Correspondence of Chinese and English Text Structure and Transformation of Thinking Patterns. *Chinese Translators Journal, No. 4,* 60-63. (In Chinese)

Marheinecke, K. (2016). Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective. In *Proceedings of the LREC 2016 Workshop Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem* (pp. 71-75). http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MT%20Evaluation_Proceedings.pdf

Melby, A. K. (2015). QT21: A New Era for Translators and the Computer. In *Proceedings of the 37th Conference Translating and the Computer* (pp. 1-11). AsLing.

Munday, J., Pinto, S. R., & Blakesley, J. (2020). *Introducing Translation Studies: Theories and Applications* (5th ed.). Routledge.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *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

SAE International (2001). *SAE J2450: Translation Quality Metric.* https://www.sae.org/standards/j2450_200112-translation-quality-metric

Shahmerdanova, R. (2025). Artificial Intelligence in Translation: Challenges and Opportunities. *Acta Globalis Humanitatis et Linguarum, 2,* 62-70. https://doi.org/10.69760/aghel.02500108

Snover, M., Dorr, B., Schwartz, R. et al. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (ACL)* (pp. 223-231). Association for Machine Translation in the Americas.

Sun, L. (2023). Reflections on Translation Quality Assessment. *Shanghai Journal of Translators, No. 5,* 37-41. (In Chinese)

Tian, P. (2020). Translation of the Multidimensional Quality Metrics (MQM) Model: Evaluation and Insights. *East Journal of Translation, No. 3,* 23-30. (In Chinese)

Turian, J. P., Shen, L., & Melamed, I. D. (2003). Evaluation of Machine Translation and Its Evaluation. In *Proceedings of MT Summit IX* (pp. 386-393). https://aclanthology.org/volumes/2003.mtsummit-papers/

Tytler, A. F. (1978). *Essay on the Principles of Translation.* John Benjamins. https://doi.org/10.1075/acil.13

Wang, H. S., & Liu, S J. (2026) he "Deconstruction" and "Reconstruction" of Translation Ethics in the Era of Generative Artificial Intelligence. Foreign Language Research, No. 1, 15-24. (In Chinese)

Wen, X., & Tian, Y. L. (2024). The Effectiveness of ChatGPT in Translating China-Specific Discourse Text. *Shanghai Journal of Translators, No. 2,* 27-34. (In Chinese)

Wieckowska, B., Kubiak, K. B., Jóźwiak, P., Moryson, W., & Stawińska-Witoszyńska, B. (2022). Cohen's Kappa Coefficient as a Measure to Assess Classification Improvement Following the Addition of a New Marker to a Regression Model. *International Journal of Environmental Research and Public Health, 19,* Article 10213. https://doi.org/10.3390/ijerph191610213

Xiong, D. H. (2018). Answering-Questions Work in the Translation of the 19th CPC National Congress Documents: Research on the Political Language and Working Mechanism. *Journal of Tianjin Foreign Studies University, 25.* (In Chinese)

Yu, G. Z. (2010). *Preface to the Chinese Translation of "The Old Man and the Sea".* Phoenix Publishing & Media Group/Yilin Press. (In Chinese)

Zhang, F. L. (2020). On Machine Translation Technology in Legal Translation. *Technology Enhanced Foreign Language Education.* (In Chinese)

Zhao, X. Y. (2018) *Norm Evolution of the English Translation of the State Leaders' Selected Works.* East China Normal University. (In Chinese)

Zhou, C. B., & Liu, Z. B. (2022). Machine Translation of Ancient Chinese Text Based on Transformer of Semantic Information Sharing. *Technology Intelligence Engineering, No. 6,* 114-127.