Article

Large language models in translation quality assessment: The feasibility of human-AI collaboration

Chengxu Wang
Nankai University
Tianjin, China
asprion@hotmail.com
https://orcid.org/0000-0002-8456-9426

Abstract: This research explores the potential application of Large Language Models (LLMs) in translation quality assessment within the Chinese Academic Translation Project (CATP), from a human-AI collaboration perspective. The study integrates the LISA QA Model and the Chinese standard GB/T 19682-2005 to develop a multidimensional translation quality assessment system, including typologies and weights of errors specific to Chinese academic works. Using this system, three LLMs (GPT-4, Claude-3.7, and Deepseek-R1) were employed to evaluate the Portuguese version of the work *Introduction to Qing Dynasty Academic Thought*, analyzing their performance and comparing it with the results of an assessment conducted by human experts, with the aim of exploring the feasibility of a collaborative model between humans and AI. Based on the experimental results, the research proposes a hierarchical assessment process of "AI screening-refined human judgment" and an inter-linguistic assessment mechanism of "Chinese prompt-multilingual verification", constructing a translation quality assessment framework based on human-AI collaboration for the CATP. This study infuses elements of technological innovation into traditional translation quality assessment, providing a new technical support pathway for the strategy of "internationalization" of Chinese academic knowledge.
Keywords: translation quality assessment; LLMs; human-AI collaboration; translation of Chinese academic works; Prompt engineering.

1. Introduction

In recent decades, the rapid growth of the Chinese economy has exposed a gap between the country's rising international status and the limited global circulation of its intellectual output. As Sun (2023, p. 97) observes, "external cultural research and dissemination, pillars of national *soft power*, have yet to keep pace with China's new geopolitical position"[1]. To enable a deeper global

[1] My translation: 作为国家"软实力"的对外文化研究与传播，却与我国的国际大国角色极不适应。

understanding of China, it is essential to bring the "masterpieces" of Chinese academic scholarship to international audiences, thereby fostering a more balanced dialogue with the global academic community. Over the past decade, the country's translation landscape has shifted from an almost exclusive emphasis on "translating the world" to a dual focus on both "translating the world" and "translating China" (Ren & Li, 2021, p. 5).

A milestone in this shift is the Chinese Academic Translation Project (CATP), established in 2010 with funding from the National Social Science Fund of China (NSSFC). The CATP selects representative works in the humanities and social sciences for translation into foreign languages and promotes their publication through prestigious international channels. As Gao and Qiu (2022, p. 132) note, the project, "[...] planned, led and supervised by the government, aims at promoting Chinese culture on the international stage and is characterized by its large scale and institutionalization"[2] (Gao & Qiu, 2022, p. 132).

The works supported by the CATP are highly specialized and represent the most authoritative scholarship in Chinese humanities and social sciences. They are marked by theoretical and terminological density: alongside abundant technical terminology, they contain numerous neologisms and culturally specific concepts that pose significant challenges for translation. As Professor Yang Ping, editor-in-chief of the *Chinese Translators Journal*, emphasizes, "[...] the export of Chinese academic production requires careful standardization of terminology with Chinese characteristics; terminological data must undergo systematic auditing and management to ensure the output of discursive resources of Chinese knowledge"[3] (Gao & Qiu, 2022, p. 132).

To address these challenges, the CATP requires a hybrid team of experts, including specialists from China and abroad, native translators of the target language, and international reviewers. Nevertheless, multi-author collaboration can still generate terminological inconsistencies and compromise textual cohesion. To mitigate this, the project introduced quality monitoring mechanisms from the production stage, shifting translation criticism from the "post-text" phase to earlier stages of the process. Yet, as the project has expanded into multiple languages[4], the demand for a more efficient, standardized assessment system—one that is sensitive to the specificities of each language—has become increasingly urgent.

From a theoretical standpoint, Translation Quality Assessment (TQA) has a long history. In the early 21st century, it emerged as a central theme in Chinese translation studies (Xu & Mu, 2009). Wang (2017) classifies research on TQA into five major paradigms:

> (1) Empirical, based on experience and intuition; (2) Equivalence-oriented, which seeks to reproduce the source text; (3) Reader-response, centered on audience reactions; (4) Functional, which emphasizes the function and purpose of the translation; and (5) Linguistic, which applies pragmatics and discourse analysis to quality assessment (Wang, 2017, p. 108).

[2] My translation: 是由国家策划、组织、主导和监管的翻译活动，是国家借此实现长远战略目标、实施对外塑造国家形象的重要行为，具有规模化、机构化、制度化等特征。

[3] My translation: 中华学术外译需要特别注重中国特色术语翻译标准化，要对术语数据进行标准化审核与管理，确保中国知识话语资源输出。

[4] According to the 2023-2024 project list, released on August 2, 2024, by the National Office for Philosophy and Social Sciences of China, 257 masterpieces received funding, covering 17 languages, including Portuguese (2 works) (National Office for Philosophy and Social Sciences, 2024).

While nearly all translation theories address the notion of quality, the most systematic models include House's functional-pragmatic model (1977, 1997, 2015), Reiss's text typology model (2000), Williams's argumentative model (2004), and Si Xianzhu's systemic-functional model (2007). Despite their contributions, these approaches remain inadequate. The limitations stem partly from structural issues within the models—such as over-reliance on subjective qualitative criteria and the absence of standardized parameters—but more critically, from their inability to adapt to the new technological and institutional demands of the CATP.

One promising approach is to integrate parameters widely adopted in the language services industry, such as the LISA QA Model, with the Chinese standard GB/T 19682-2005—*Target Text Quality Requirements for Translation Services*. The former provides a weighted taxonomy of errors that enables quantification, while the latter reflects local practices and expectations. Together, these models offer a robust foundation for a multidimensional assessment system suited to the needs of the CATP. Yet, human-only evaluation remains slow and inconsistent. In large-scale programs such as the CATP marked by linguistic diversity and high terminological complexity, exclusive reliance on human reviewers is impractical for maintaining quality control.

The advent of Large Language Models (LLMs) introduces new opportunities. Trained on massive datasets, LLMs have achieved impressive results in natural language processing tasks, including translation (Brown et al., 2020; Hendy et al., 2023; Hu & Li, 2023). This has fueled growing academic interest in their potential as tools for translation quality assessment (Kocmi & Federmann, 2023). Within this context, the paradigm of Human-Centered AI (HCAI) has gained traction. Rather than seeking to replicate human abilities, HCAI aims to develop tools that extend human capacity, improve performance, and ensure that users remain supervisors of the system at every stage (O'Brien, 2024; Shneiderman, 2022; Wang & Zhang, 2025).

Building on this paradigm, the present study proposes a collaborative human-AI evaluation framework, leveraging recent advances in LLMs to strengthen the consistency, efficiency, and scalability of translation quality control. Acknowledging both the capabilities and limitations of LLMs, including their tendency toward hallucination, cultural bias, and terminological gaps, this study proposes a human-AI collaborative evaluation framework for the CATP. As a case study, it examines the Portuguese translation of *Introduction to Qing Dynasty Academic Thought*[5] by Liang Qichao (1873–1929), funded by the NSSFC in 2024. Building on this, a multilevel system of error categories and weights was developed by integrating the LISA QA Model with the Chinese standard GB/T 19682-2005.

Assuming that a human-AI collaborative framework offers greater speed, consistency, and replicability for the CATP, assessments generated by ChatGPT-4, Claude 3.7, and DeepSeek R1 were compared with those of expert human reviewers to identify performance patterns, strengths, and weaknesses. On this basis, the article seeks to address the following research questions:

1. How can a multidimensional TQA model be designed to meet the specific requirements of Chinese academic works within the CATP?

[5] 清代学术概论 (original title in Chinese. Published by Zhonghua Book Company, 2022).

2. What is the comparative performance of ChatGPT-4, Claude 3.7, and DeepSeek R1 in evaluating translations from Chinese into Portuguese?
3. How can a human-AI collaborative framework, anchored in LLMs, be structured to enhance the efficiency, consistency, and scalability of quality control for future high-level Chinese-Portuguese translations?

2. Literature review

Scholarship on the Chinese Academic Translation Project (CATP) has expanded steadily since 2013. A search on the CNKI[6] platform for the term "中华学术外译" (the full Chinese designation of the CATP) yields 36 articles up to April 2025, of which 24 (66.7%) fall within the field of Translation Studies (CNKI, 2025). While adopting diverse perspectives, these studies converge on a central concern with translation quality.

Researchers have examined both internal and external factors shaping translation outcomes. Under the framework of Skopostheorie, He and Hu (2018) emphasize the importance of translator competence, text selection, and translation strategies. In a similar vein, Wang and Shi (2024, p. 97) propose developmental pathways for translational competence at three levels—individual translators, universities, and society—aimed at "[...] consolidating a talent base that supports the internationalization of Chinese scholarship and strengthening the capacity and effectiveness of global dissemination of academic discourse and national culture"[7]. Yao and Friedman (2019) further demonstrate empirically that intensive communication between translators and authors enhances fidelity and academic rigor in target texts.

Another recurring theme is the standardization of culture-specific Chinese terminology, regarded as essential for improving translation quality. Long and Zhou (2024), in their study of the English version of *The Transformation of China*, show that combining specialized corpora, computer-assisted translation (CAT) tools, and LLMs (e.g., ChatGPT) significantly reduces terminological inconsistency. Similarly, in research on the English translation of *The Academic History of the Han Dynasty*, Li (2023) advocates the creation of macro-glossaries to stabilize terminological choices across projects.

The most recent scholarship emphasizes the integration of emerging technologies. From the perspective of eco-translation, Zhang and Zhu (2024) propose a "linguistic-cultural-communicative" balance model enhanced by digital tools to increase both the acceptability and impact of translated texts. Jiang et al. (2024) and Yao (2024) explore the use of LLMs in translation assessment. Jiang et al. (2024) analyze Chinese-Portuguese translations in genres such as political discourse and poetry, though on the basis of a limited corpus of 20 sentences. Yao (2024), by contrast, focuses on technical translation and post-editing with LLMs but does not advance a systematic evaluation model.

With respect to TQA, no unified standard has yet been consolidated for the CATP. From the perspective of descriptive translation studies, Tao (2020) argues that evaluation criteria should account for the interaction between the target text and the sociocultural expectations of its

[6] CNKI (China National Knowledge Infrastructure) is the largest academic portal in China, aggregating journals, dissertations, conference proceedings, and other primary sources. The portal is available at: http://www.cnki.net
[7] My translation: 为中国学术"走出去"夯实人才根基，提升我国学术话语和文化国际传播的能力与效果。

readership. Wang et al. (2020), working from a dialogic cooperation model, call for the inclusion of metrics addressing co-authorship, terminology, and cultural adaptation.

In general, the existing literature contributes valuable insights into the pre-translation selection of academic works, the normative management of terminology during the translation process, and post-translation quality review. However, several limitations persist: evaluative methods remain largely qualitative, with quantitative metrics still underdeveloped; the dimensions and specificity of assessment criteria lack standardization, as no widely recognized scales or indicators exist within the language services industry; and the application of modern technologies—such as LLMs and automated quality-control tools—remains at an incipient stage.

3. Theoretical framework

This section presents the conceptual foundations of the study, organized around two main dimensions: (i) the premises for developing a multidimensional system of translation quality assessment for the CATP, and (ii) the theoretical basis for a human-AI collaborative evaluation model. A subsection on prompt engineering is also included, given its decisive role in operationalizing LLMs within the evaluative context.

3.1 Foundations for a multidimensional assessment system

I propose integrating the LISA QA Model (1995)—widely adopted in the localization industry—with the Chinese standard GB/T 19682-2005, *Target Text Quality Requirements for Translation Services* (Chinese GB Standards, 2005). This integration aims to bridge gaps identified in the literature and to establish a robust protocol specifically suited to the CATP.

3.1.1 LISA QA Model

As globalization and the language services industry expanded, TQA evolved from predominantly subjective judgments to structured, standardized, and technology-driven approaches. Developed in 1995 by the Localization Industry Standards Association (LISA), the LISA Quality Assurance Model was designed for localization projects but remains a versatile tool applicable to product documentation, help systems, user interfaces, and even computer-assisted training (Martínez Mateo, 2014). Its widespread adoption reflects the balance it strikes between systematic rigor and practical operability, making it one of the most influential assessment frameworks in the translation and localization sector.

The model is built on two complementary dimensions: error typology—covering mistranslation, terminology, language, style, consistency, cultural appropriateness (country/locale), among others—and severity, categorized as minor, major, or critical. The overall score derives from the weighted sum (error × weight) by severity, allowing translation quality to be quantified objectively and compared across projects.

A key strength of the model is its flexibility: evaluators can adjust categories, weights, and tolerance thresholds to match project objectives, text genres, and target audiences. This adaptability

renders the model particularly suitable for complex academic translations, such as those undertaken within the CATP.

3.1.2 Chinese Standard GB/T 19682-2005

The rapid expansion of China's translation services sector exposed disparities in quality and the absence of unified standards, prompting the need for nationwide regulatory intervention. To address these gaps, the State Administration for Standardization issued GB/T 19682-2005, *Target Text Quality Requirements for Translation Services*, in March 2005. As the country's first normative framework with regulatory force, it marked the beginning of standardized quality control in Chinese translation practice.

The standard consists of nine articles, with evaluation criteria detailed in Articles 4 through 8. These criteria are structured on two interdependent levels: (i) a quality requirements layer, subdivided into basic, specific, and other requirements; and (ii) a judgment and verification layer, which defines the analytical procedure for calculating scores or error rates using the formula "error type + severity + sampling scope" (Chinese GB Standards, 2005).

The basic requirements emphasize three principles: "faithfulness to the source, terminological consistency, and textual fluency". These echo the classical Chinese notions of *xin, da, ya* (faithfulness, expressiveness, elegance), but are reformulated in operational terms (Chinese GB Standards, 2005). The specific requirements prescribe best practices for handling numbers, proper names, measurement units, abbreviations, and formatting, offering translators practical guidance (Chinese GB Standards, 2005). The other requirements extend the scope to special cases—including neologism creation, complex textual structures, poetry, advertising, third-language conversion, supplementary texts, and correction of source-text errors—thus ensuring flexibility across diverse genres (Chinese GB Standards, 2005).

As illustrated in Figure 1, the second layer provides a quantitative framework for assessment, logically aligned with the LISA QA Model and well-suited to the future integration of automated tools, including LLMs.

Figure 1: Composite Error Rate

$$\text{Composite Error Rate} = KC_A \frac{c_{\text{I}} D_{\text{I}} + c_{\text{II}} D_{\text{II}} + c_{\text{III}} D_{\text{III}} + c_{\text{IV}} D_{\text{IV}}}{W} \times 100\% \quad \cdots\cdots\cdots\cdots (\text{A.1})$$

Source: Chinese GB Standards (2005).

[Appendix A of GB/T 19682-2005 introduces formula (A.1) for calculating a composite error rate, which consolidates errors of varying severity, intended use, and textual difficulty into a single percentage indicator. The product of *K* and the coefficient *CA* adjusts the result to reflect both the complexity of the source text and the quality requirements of the translation's target use. The weighted sum of errors (numerator) is normalized by corpus size (*W*) and converted into a percentage, providing an objective benchmark for determining whether a translation meets or falls below the threshold defined by the project or the standard.]

In developing the multidimensional system for this study, GB/T 19682-2005 is treated as the local counterpart to the LISA QA Model. While the three key principles are retained, the categories are expanded to address the specificities of Chinese academic works, such as political sensitivity,

cultural appropriateness, and terminological rigor. This integration ensures both cultural relevance and international transferability, attributes essential to the CATP.

3.1.3 Proposed assessment system for the CATP

As Wei et al. (2020) note, contemporary approaches to TQA follow an analytical chain that begins with a multilevel typology of errors, proceeds through severity ratings and weighted penalties, and culminates in the calculation of a quantitative index (error rate or score). This logic of "error – severity - weight - sampling - score" aligns closely with the requirements of the CATP. Building on this framework, I propose a new assessment system designed specifically for Chinese academic works within the CATP. The system integrates the parameters of the LISA QA Model with the three core principles of the Chinese standard GB/T 19682-2005—faithfulness, terminological consistency, and fluency. The result is a multidimensional framework, the main components of which are summarized in Table 1.

Table 1: Translation Quality Assessment (TQA) System for the CATP

Error Level	Name	Category	Description	Weight
Level I	Critical Error	Country (a1)	Errors in translating politically sensitive terms, including those related to the Party, foreign relations, religion, sovereignty, Hong Kong, Macao, and Taiwan.	1.0
Level II	Major Error	Omission (b1)	Omission of sentences, words, or passages.	0.5
		Unwarranted Addition (b2)	Insertion of content absent from the source text.	0.2
		Mistranslation (b3)	Misinterpretation of cultural context, idiomatic expressions, or source-specific features leading to mistranslation.	0.5
		Terminology (b4)	Use of technical terms inconsistent with established standards or conventions.	0.3
		Grammar (b5)	Grammatical errors such as subject-verb disagreement, incorrect tense, or faulty sentence structure.	0.3
Level III	Minor Error	Inconsistency (c1)	Inconsistent use of terms, numbers, dates, acronyms, or capitalization.	0.2
		Spelling/Punctuation (c2)	Errors in spelling or punctuation.	0.2

Source: The Author (2025)

[Error Rate = (K×(a1×1.0+b1×0.5+b2×0.2+b3×0.5+b4×0.3+b5×0.3+c1×0.2+c2×0.2)) /Number of source characters ×1000, where K is the difficulty coefficient (set at 1.0 for academic works intended for publication). Repeated errors of the same type are counted only once.]

It is important to note that subcategory a1 is defined as a critical error with a weight of 1.0 and a veto effect. This designation reflects not only ethical principles of accuracy but also the high political sensitivity of the CATP, a state-led initiative in cultural diplomacy. Unlike commercial or purely academic projects, the CATP forms part of China's broader soft power strategy and its mission to "tell China's story well". Any impropriety concerning government institutions, sovereignty, ethnicity, religion, or issues related to Hong Kong, Macao, and Taiwan may distort China's position, provoke diplomatic misunderstandings, or fuel media controversies, with

potentially irreversible damage to the country's image. The "maximum weight + veto" logic therefore functions as a mechanism of discursive risk management, underscoring that, in this context, quality assessment extends beyond the linguistic dimension to encompass the alignment of international communication policy with high-level translation practice.

This design ensures traceability of errors, scalability through LISA-inspired modularity, and cultural localization via GB/T. Most importantly, it facilitates integration with LLMs, which can perform error detection while leaving judgments of severity and context to human evaluators.

3.2 Foundations for human-AI collaborative evaluation

The conceptual shift initiated by Shneiderman (2022) culminated in the "Human-Centered AI" (HCAI) framework, formulated in 2020. After decades of industry narratives centered on emulating or replacing human reasoning, Shneiderman (2022, p. 8) argues that technology should instead expand human capabilities, enhance performance, and empower users, while keeping them in continuous supervisory control of the system. Applied to Translation Studies, this principle is embodied in O'Brien's (2024, p. 402) notion of "Human-Centered Augmented Translation", which "can be viewed as a way of amplifying translators' abilities, empowering translators, while also allowing them to maintain control".

Recent scholarship reinforces this orientation. Wang and Zhang (2025, p. 55) argue that the rise of generative AI is moving the field toward a new paradigm of "human-AI collaboration—systemic integration—standards updating". LLMs, on the one hand, provide computational power for detecting terminological inconsistencies and rapidly flagging errors; on the other, the market is shifting from SaaS (Software as a Service) to MaaS (Model as a Service), necessitating governance frameworks in which translators oversee and refine algorithmic output.

Within this perspective, we embed the concept of HCAI into translation quality assessment for the CATP through a threefold circuit:

1. Capacity enhancement: LLMs perform structural error scans and generate preliminary statistics, allowing evaluators to concentrate on cultural context and political risk analysis (with particular attention to category a1 items subject to automatic veto).
2. Configured supervision: evaluators define weights, thresholds, and red lines in advance; the model functions strictly within these parameters and generates interpretable reports, ensuring alignment with the principle of "humans in the group; computers in the loop" (Shneiderman, 2022, p. 21).
3. Continuous learning: human judgments feed back into prompt engineering and model fine-tuning, activating a cycle of "algorithmic detection—human arbitration—model adjustment", through which terminology and cultural adequacy are progressively refined over the course of the project.

This design harmonizes algorithmic efficiency with human sensitivity, meeting both the operational demands of the CATP and the ethical and political imperatives of a cultural diplomacy initiative.

3.3 Prompt engineering

The effectiveness of large language models depends heavily on prompt engineering, understood as programming the model through textual instructions that guide, refine, or expand its capabilities (White et al., 2023). Increasingly, this practice is recognized as a critical lever for enhancing the "intelligence" of generative AI and adapting it to user needs (Wang, 2024).

In human-AI collaborative evaluation, prompts function as a control mechanism: they determine the relevance, traceability, and auditability of automated outputs, directly shaping the extent of human rework and, by extension, the overall cost of quality. Scholarship converges on a tripartite structure—Context (Background), Task/Objectives, and Specific Requirements—augmented by frameworks such as ICIO[8] (Elavis Saravia), CRISPE[9] (Matt Nigh), and BROKE (Chen C. M.) (Wang & Xie, 2024).

Aligned with these findings, this study developed an evaluative academic prompt (see Wang, 2025b, Appendix A) that integrates elements of CLEAR, BROKE, and CRISPE, adapted to the Portuguese version of *Introduction to Qing Dynasty Academic Thought*. The design unfolds in three steps:

1. Background: specification of the excerpt under evaluation, the evaluator's profile (Chinese-Portuguese bilingual with training in Chinese intellectual history), and the experimental purpose (validation of the proposed TQA system).
2. Role: the LLM is explicitly assigned the function of evaluation assistant, while the human translator retains decision-making authority. All responses must conform to the prescribed category system and glossary, ensuring adherence to the HCAI framework.
3. Objectives: instruction to generate a matrix of "number of errors – category – justification", with automatic veto applied to a1 errors (political sensitivity). This constraint leverages computational power without compromising red-line risks.

The resulting cycle—"algorithmic detection – human validation – prompt adjustment"—operationalizes the continuous learning loop outlined in Section 3.2, ensuring that terminology, stylistic coherence, and political compliance are progressively refined over the course of the project. In this way, prompt engineering serves as the integrative link between the analytical mechanism of TQA and the principles of human-centered AI.

4. Methodology

This section outlines the criteria for corpus selection, the rationale for assessing CATP translation quality with LLMs and human experts, and the experimental design, ensuring reproducibility and reliability. Three representative models were selected: GPT-4 (OpenAI, 2023), known for its multilingual robustness and global adoption; Claude-3.7 (Anthropic), noted for its emphasis on safety, accuracy, and refined reasoning in academic texts; and DeepSeek-R1 (DeepSeek-

[8] See DukeManh (2024).
[9] See DukeManh (2024).

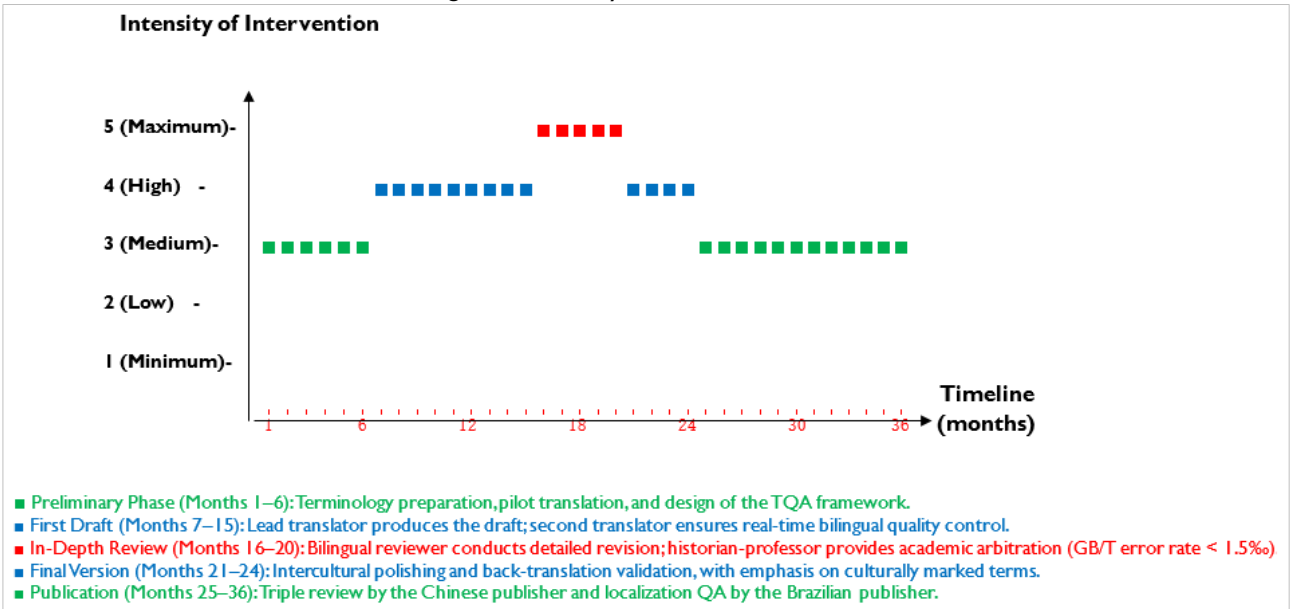
A1, 2025), a China-based model optimized for linguistic and cultural specificities of Chinese. Together, these models capture complementary trajectories of AI development—global, safety-focused, and locally optimized—providing a comprehensive basis for comparing machine translation quality in specialized and culturally sensitive contexts.

4.1 Object of study and corpus selection

This study examines the Portuguese translation of *Introduction to Qing Dynasty Academic Thought* by the renowned Chinese scholar Liang Qichao (1873–1929). Written from the perspective of "new historiography", the work systematizes Qing intellectual currents and analyzes the influence of Western ideas on late academic transformation, marking a turning point in Chinese historiography.

The 36-month translation project involves two translators (a Brazilian sinologist and a native Chinese translator) and a team of four specialists: two historians (Chinese and Brazilian), a Chinese scholar of classical texts, and a bilingual reviewer. Their roles span pre-translation terminology definition, clarification during translation, and post-translation global review. As shown in Figure 2, quality assessment is embedded throughout the entire process.

Figure 2: Intensity of TQA Intervention



Source: The Author (2025)

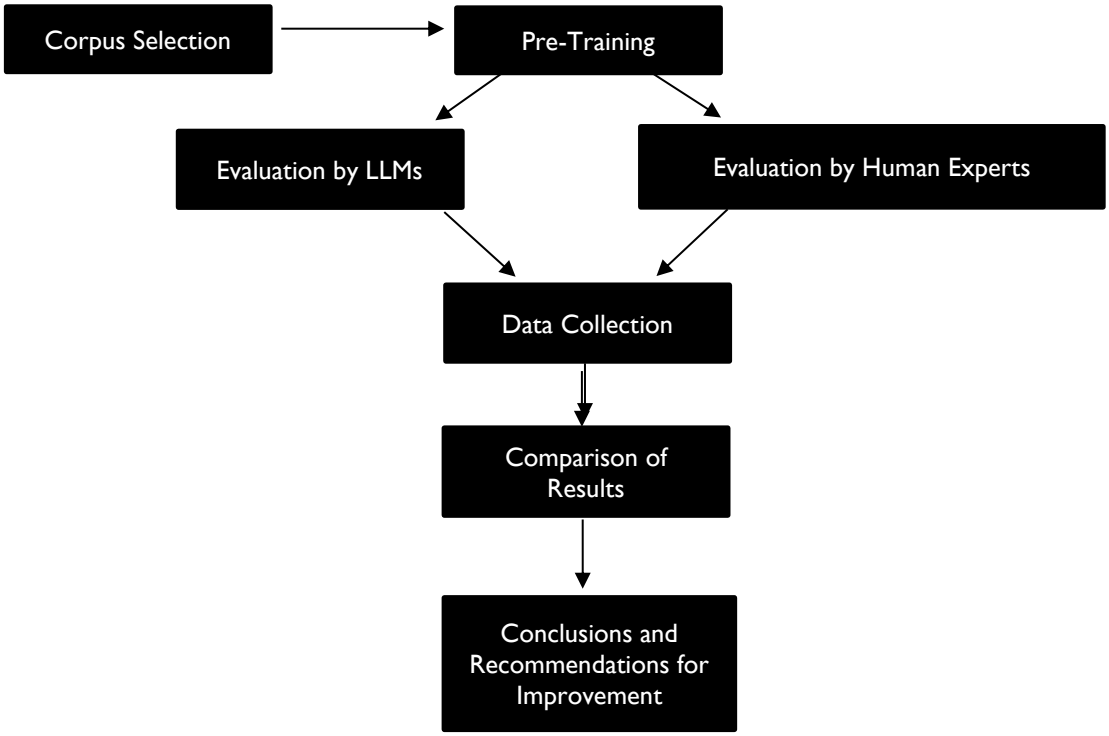
To build a representative evaluation corpus, a native Portuguese-speaking student was invited to translate Chapter 15, "地理学 天文算学" (Geography, Astronomy, and Mathematics). This chapter is rich in traditional Chinese terminology, combines classical and modern vernacular Chinese, alternates between critiques of Qing scholars and Sino-Western comparisons, and cites numerous works and authors, requiring consistency in proper names and references (see Wang, 2025b, Appendix C). Because the novice translator was not a native Chinese speaker and lacked training in classical Chinese, a modern Chinese explanatory version—prepared by the author—was provided to aid comprehension and minimize distortions (see Wang, 2025b, Appendix D). This support can

be viewed as a form of intralingual translation. The novice translator worked independently, with access to bibliographic and online resources but without machine translation or LLMs. No specialized revision was conducted in order to preserve the authentic limitations of a novice translator. The chapter's terminological density and intercultural complexity make it especially prone to errors, providing an appropriate test case for evaluating the effectiveness of LLMs in translation quality assessment.

4.2 Research design

As shown in Figure 3, the research design employs triangulation, integrating two verification chains: "LLMs + multiple prompts" and "AI × human experts".

Figure 3: Experimental Flowchart



Source: The Author (2025)

The experiment begins by aligning the knowledge base through the translation quality assessment system (Table 1, Subsection 3.1.3) and the Chinese-Portuguese bilingual glossary (see Wang, 2025b, Appendix B). Multiple LLMs then conduct independent evaluations using standardized prompts (see Wang, 2025b, Appendix A), followed by parallel assessments from bilingual experts without AI assistance. The two sets of results are compared quantitatively and qualitatively to assess model consistency, prompt sensitivity, and divergences between AI and human evaluators in error detection and scoring. Table 2 summarizes the experimental procedures in detail.

Table 2: Experimental Procedures

Dimension	Description
LLMs	ChatGPT-4, Claude-3.7, DeepSeek-R1
Prompt language	Chinese (3 repetitions) / Portuguese (3 repetitions)
Human evaluators	1 native Chinese speaker and 1 native Portuguese speaker
Total tests	2 languages × 3 rounds × 3 models = 18 AI evaluations + 2 human evaluations

Source: The Author (2025)

For the AI evaluation, three LLMs were employed—ChatGPT-4, Claude-3.7, and DeepSeek-R1. Prompts were prepared in Chinese and Portuguese, with three independent runs in each language, yielding 18 AI-generated evaluation sets[10]. All tests used the same bilingual bitext (see Wang, 2025b, Appendix C), ensuring consistent input conditions across models and languages.

For the human evaluation, two bilingual experts—one a native Chinese speaker and the other a native Brazilian Portuguese speaker—conducted independent assessments following standardized guidelines. Each was provided with the *Translation Quality Assessment System* (Table 1), the Chinese source text, the Portuguese target text, and the *Chinese-Portuguese Bilingual Glossary of Specialized Terms* (see Wang, 2025b, Appendix B).

The evaluation followed five steps: (1) preliminary study of the assessment system; (2) close reading of the source text and reference terms; (3) systematic evaluation of the translation according to established criteria; (4) error annotation in the translated text, coded by severity (Level I in red, Level II in yellow, Level III in green); and (5) preparation of an analytical report indicating the number, typology, and possible causes of errors.

To ensure comparability with LLM outputs, evaluators were allowed to consult dictionaries and reference materials but were strictly prohibited from using AI tools. This process yielded two independent sets of human evaluation data. To create a coherent and reliable "human gold reference", the two sets were cross-checked and merged through an equal-weighted average, producing a unified benchmark. This consolidated standard preserves the complementary strengths of both native bilingual perspectives and provides a single authoritative reference for future human-AI collaborative assessments.

4.3 Reliability and validity

To ensure reliability and generalizability, the study adopted three safeguards:

1. Internal consistency: a "multi-model × multilingual" cross-validation strategy, with three LLMs tested in both Chinese and Portuguese to verify stability of results.
2. External validity: use of representative chapters from *Introduction to Qing Dynasty Academic Thought*, marked by dense terminology, full textual structure, and strong intercultural content, reflecting typical challenges in academic translation.
3. Traceability: systematic archiving of experimental records, scoring details, and raw datasets to guarantee transparency and replicability.

[10] The complete raw data from the 18 test rounds are available in Wang (2025a).

These measures strengthen both reliability and validity while providing a transparent methodological foundation for future research.

5. Analysis and discussion of results

This section presents a systematic analysis of the test data, examining how three LLMs—ChatGPT-4, Claude-3.7, and DeepSeek-R1—performed in assessing translation quality within the CATP project. Their outputs are compared with human expert evaluations to provide empirical evidence for developing a translation quality assessment model based on human-AI collaboration.

5.1 Error-detection capacity of the three LLMs

This subsection examines the sensitivity of different models to various error types and the influence of prompt language on their performance. Such analysis helps clarify the strengths and limitations of LLMs in translation quality assessment.

As shown in Table 3, ChatGPT-4 detected 5, 10, 20, 8, 18, and 5 errors across the six tests, showing marked fluctuation. Most of these were concentrated in two categories—*mistranslation (b3)* and *terminology (b4)*—indicating that the model is particularly responsive to semantic deviations and terminological inconsistencies.

Table 3: Performance of ChatGPT-4

Evaluation dataset of ChatGPT-4						
Error type	No. of errors (Test 1: Chinese prompt)	No. of errors (Test 2: Portuguese prompt)	No. of errors (Test 3: Chinese prompt)	No. of errors (Test 4: Portuguese prompt)	No. of errors (Test 5: Chinese prompt)	No. of errors (Test 6: Portuguese prompt)
Country (a1)	0	0	0	0	0	0
Omission (b1)	0	1	1	1	1	0
Unwarranted addition (b2)	0	0	1	1	0	0
Mistranslation (b3)	2	2	6	2	6	4
Terminology (b4)	2	2	4	1	5	0
Grammar (b5)	0	1	3	1	2	0
Inconsistency (c1)	1	2	2	1	3	0
Spelling/punctuation (c2)	0	2	3	1	1	1
Total	5	10	20	8	18	5

Source: The Author (2025)

Notably, in Tests 3 and 5 with Chinese prompts, ChatGPT-4 identified 20 and 18 errors — substantially more than the 10, 8, and 5 errors found in Tests 2, 4, and 6 with Portuguese prompts. This suggests that prompt language significantly influences the model's performance, with Chinese instructions yielding greater error sensitivity. By category, ChatGPT-4 showed steady detection of *mistranslation (b3)* and some responsiveness to *terminology (b4)*, though with fluctuations. In contrast, performance was weak for *grammar (b5)* and *spelling/punctuation (c2)*—both tied to Portuguese norms—indicating limited sensitivity to language-specific conventions. For *Country-related errors (a1)*, no cases were identified across all tests, consistent with human expert evaluations. This outcome suggests both the actual absence of politically sensitive errors in the sample and the model's capacity to judge such errors accurately.

Table 4 shows the performance of Claude-3.7 across six tests, with 26, 8, 19, 18, 18, and 12 errors detected. Although results fluctuated, the model achieved a higher average (16.8) than ChatGPT-4 (10.8). Notably, in the first test with a Chinese prompt, it identified 26 errors—significantly more than in other runs and closer to human evaluation results. Overall, Claude-3.7 demonstrates stronger alignment with human assessments, particularly when operating with Chinese prompts.

Table 4: Performance of Claude-3.7

Evaluation dataset of Claude-3.7						
Error type	No. of errors (Test 1: Chinese prompt)	No. of errors (Test 2: Portuguese prompt)	No. of errors (Test 3: Chinese prompt)	No. of errors (Test 4: Portuguese prompt)	No. of errors (Test 5: Chinese prompt)	No. of errors (Test 6: Portuguese prompt)
Country (a1)	0	0	0	0	0	0
Omission (b1)	2	1	3	2	3	2
Unwarranted addition (b2)	1	0	0	0	1	0
Mistranslation (b3)	9	2	5	4	5	2
Terminology (b4)	6	3	4	1	3	2
Grammar (b5)	1	0	2	3	1	2
Inconsistency (c1)	2	1	2	5	2	2
Spelling/punctuation (c2)	5	1	3	3	3	2
Total	26	8	19	18	18	12

Source: The Author (2025)

Claude-3.7, like ChatGPT-4, showed strong sensitivity to *mistranslation (b3)* and *terminology (b4)* errors. It was more consistent in detecting *omissions (b1)*, averaging 2.2 errors per test compared to ChatGPT-4's 0.5, though still below the human benchmark of 5. Its performance on *spelling/punctuation (c2)* was also stronger (2.8 vs. 1.3 for ChatGPT-4), but again fell short of the human result (5). For *State-related errors (a1)*, it consistently detected none, matching the human evaluation.

Table 5: Performance of Deepseek-R1

Evaluation dataset of Deepseek-R1						
Error type	No. of errors (Test 1: Chinese prompt)	No. of errors (Test 2: Portuguese prompt)	No. of errors (Test 3: Chinese prompt)	No. of errors (Test 4: Portuguese prompt)	No. of errors (Test 5: Chinese prompt)	No. of errors (Test 6: Portuguese prompt)
Country (a1)	0	0	0	0	0	0
Omission (b1)	0	1	1	1	2	1
Unwarranted addition (b2)	0	0	0	0	0	1
Mistranslation (b3)	5	1	1	4	4	2
Terminology (b4)	3	1	1	3	3	1
Grammar (b5)	1	1	1	0	6	1
Inconsistency (c1)	4	1	0	2	3	1
Spelling/punctuation (c2)	2	1	0	2	2	2
Total	15	6	4	12	20	9

Source: The Author (2025)

As shown in Table 5, DeepSeek-R1 detected 15, 6, 4, 12, 20, and 9 errors across six tests, displaying wide variation. Its average number of detected errors (11) was lower than Claude-3.7 (16.8) but slightly higher than ChatGPT-4 (10.8). Compared with the other two models, DeepSeek-R1's results showed a more pronounced effect of prompt language: Tests 1 and 5 with Chinese prompts yielded 15 and 20 errors, respectively, while Tests 2 and 3 with Portuguese prompts yielded only 6 and 4.

In terms of error distribution, DeepSeek-R1 showed instability in detecting *mistranslation (b3)*, with an average of 2.8 errors—lower than ChatGPT-4 (3.7), Claude-3.7 (4.5), and the human benchmark (4). For *terminology (b4)*, it identified an average of 2 errors per test, close to ChatGPT-4 (2.3) but higher than the human evaluation (1), suggesting heightened sensitivity to Chinese academic terminology, consistent with its development context. A striking feature was its fluctuation in *grammar (b5)* detection: 1, 1, 1, 0, 6, and 1 errors across the six tests. In Test 5 with a Chinese prompt, the model flagged six grammatical errors—exceeding the human result (5)—indicating that under certain prompt conditions, DeepSeek-R1 may display strong sensitivity to Portuguese grammar issues. Like the other two models, DeepSeek-R1 consistently reported no *Country-related errors (a1)*, aligning with the human evaluation and confirming the absence of politically sensitive issues in the sample.

Analysis of the three LLMs shows that prompt language significantly affects performance: all models identified more errors with Chinese prompts than with Portuguese ones. Two factors likely explain this pattern: limited Portuguese training data compared to Chinese or English, and closer alignment of Chinese prompts with the evaluative and cultural context of the CATP. This suggests that selecting the language in which a model is most proficient as the instruction medium can improve both efficiency and accuracy in LLM-based translation quality assessment.

5.2 Comparative Analysis of LLM-Based and Human Evaluation

A comparison between the results produced by the three LLMs and those of the human experts (Table 6) reveals significant differences.

Table 6: Human Evaluation Results

Human evaluation dataset	
Country (a1)	0
Omission (b1)	5
Unwarranted addition (b2)	2
Mistranslation (b3)	4
Terminology (b4)	1
Grammar (b5)	5
Inconsistency (c1)	1
Spelling/punctuation (c2)	5
Total	23

Source: The Author (2025)

Human experts identified 23 errors in total, while the LLMs ranged from 5 to 26 across tests, showing wide variability. Claude-3.7 reached 26 in one case, exceeding the human result, whereas DeepSeek-R1 dropped to only 4 in another. On average, ChatGPT-4, Claude-3.7, and DeepSeek-R1 detected 11, 16.8, and 11 errors respectively—well below the 23 recorded by experts. This suggests that while LLMs demonstrate some error-detection capacity, they cannot yet match human reliability in quantifying total errors.

In terms of error type, notable gaps remain: human evaluators consistently flagged more *omission (b1)*, *grammar (b5)*, and *spelling/punctuation (c2)* errors, indicating that LLMs underdetect structure-and form-related issues compared to humans.

Table 7: Distribution of Error Type Identification (LLMs vs. Human)

Error type	ChatGPT-4	Claude-3.7	Deepseek-R1	Human
Country (a1)	0	0	0	0
Omission (b1)	0,50	2,20	1,00	5
Unwarranted addition (b2)	0,17	0,33	0,17	2
Mistranslation (b3)	2,83	4,50	2,83	4
Terminology (b4)	2,83	2,17	2,00	1
Grammar (b5)	1,17	1,50	1,67	5
Inconsistency (c1)	1,33	2,33	1,83	1
Spelling/punctuation (c2)	1,83	2,83	1,50	5

Source: The Author (2025)

Particularly in the category of *omission (b1)*, the average detection rates of ChatGPT-4 (0.5), DeepSeek-R1 (1), and Claude-3.7 (2.2) were all significantly lower than those of human experts. By contrast, for *mistranslation (b3)* and *terminology (b4)*, the LLMs' detection rates were comparable to or higher than the human evaluation: human experts identified 4 *mistranslations*, while Claude-3.7 averaged 4.5; for *terminology*, humans identified 1 error, whereas ChatGPT-4, Claude-3.7, and DeepSeek-R1 averaged 2.83, 2.17, and 2, respectively.

This distribution suggests that LLMs are comparatively strong in detecting errors related to semantic comprehension and terminological standardization, but weaker in identifying *omissions*—which require detailed comparison between source and target texts—as well as *grammar* and *spelling/punctuation* errors, which demand a deep command of target-language norms.

Further analysis of overlapping errors identified by both LLMs and human experts highlights the precision of LLMs in specific categories. For *mistranslation (b3)* and *terminology (b4)*, overlap rates—defined here as agreement between LLMs and human experts in identifying the same errors—were relatively high: 75% for ChatGPT-4, 82% for Claude-3.7, and 65% for DeepSeek-R1. This indicates a high degree of reliability in these categories. In contrast, overlap dropped below 50% for *omission (b1)*, *grammar (b5)*, and *spelling/punctuation (c2)*, underscoring the difficulty LLMs face in detecting these types of errors.

Regarding hallucinations, DeepSeek-R1 displayed multiple instances across the six tests. In Test 2, it prematurely anticipated subsequent files and evaluation steps, and after receiving the evaluation framework, it generated assessment results without access to the actual translation corpus. In Test 3, it fabricated requests for terms such as "Cheng-Zhu Neo-Confucianism" (程朱理学) and "Qian-Jia Textual Studies" (乾嘉考据), which were absent from the source text. In Test 4,

it produced fabricated examples of translated sentences without reference to the authentic corpus, and in Test 5, it even initiated evaluation without receiving any translated text. These cases highlight DeepSeek-R1's instability and tendency toward speculative output, which severely undermines the accuracy and reliability of automated translation quality assessment.

In addition, the LLMs sometimes exhibited *over-identification*—classifying correct translations as errors. For example, in Tests 3 and 5, ChatGPT-4 misclassified 4 and 3 correct translations, respectively, as *mistranslation* or *terminology* errors; in Test 1, Claude-3.7 marked 5 correct translations as errors. Such *false positives*—cases where a system incorrectly flags correct items as errors—diminish the reliability of LLM-based evaluations and underscore the necessity of human expert review and correction in practical applications.

6. Conclusion

This study, based on the Portuguese translation of *Introduction to Qing Dynasty Academic Thought*, examined the potential of Large Language Models (LLMs) in translation quality assessment (TQA) within the Chinese Academic Translation Project (CATP) from a human-AI collaboration perspective. By integrating the LISA QA Model with the Chinese standard GB/T 19682-2005, I developed a multidimensional evaluation system that incorporates both international practices and local specificities, including a veto mechanism for politically sensitive errors.

Experimental results with ChatGPT-4, Claude-3.7, and DeepSeek-R1, compared to human expert evaluations, revealed distinct performance profiles. Claude-3.7 most closely approximated human judgments but tended toward over-identification. ChatGPT-4 showed stability in semantic and terminological detection but underperformed in omissions. DeepSeek-R1 proved less stable, with frequent hallucinations, though it demonstrated particular sensitivity to grammatical issues. Overall, the findings indicate that LLMs can serve as effective screening tools, but expert oversight remains indispensable. Based on these results, I propose a hierarchical process of "AI screening – refined human judgment" and an interlinguistic mechanism of "Chinese prompts – multilingual verification". Together, these approaches harness the efficiency of LLMs while safeguarding accuracy and political sensitivity through human validation.

The study nevertheless faces limitations: the restricted corpus, the "black-box" nature of LLMs, and unresolved challenges such as hallucinations and false positives. Importantly, no politically sensitive errors (a1) were present in the corpus, limiting conclusions about this critical category. Future research should design targeted experiments with sensitive content to fully test LLMs in high-risk domains. In conclusion, this research demonstrates that human-AI collaborative frameworks hold promise for enhancing efficiency, consistency, and scalability in academic translation quality assurance, offering valuable support to the internationalization of Chinese scholarly knowledge.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,

Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/ARXIV.2005.14165

China National Knowledge Infrastructure (CNKI) (2025). Global Academic Insights from China and Beyond. *China National Knowledge Infrastructure*. http://www.cnki.net/index/

Chinese GB Standards. (2005, Mar 24). *Target text quality requirements for translation services (GB/T 19682-2005)*. https://iplogger.com/2hhR16

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Cornell University*. https://arxiv.org/abs/2501.12948

DukeManh. (2024, Mar 5). Prompting Introduction. *Github.com*. https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/guides/prompts-intro.md

Gao, Q., & Qiu, H. M. (2022). 中国文化外译与国家翻译实践 [Chinese Culture Translation and National Translation Practice]. *中国翻译* [*Chinese Translators Journal*], *43*(4), 129–132.

Giray, L. (2023). Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering*, *51*, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

He, H. Z., & Hu, W. H. (2018). 目的论视域下中华学术外译策略研究 [A Study on the Strategies of Translating Chinese Scholarship from the Perspective of Skopos Theory]. *新西部* [*New West*], *36*(18), 87–102.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *Cornell University*. https://doi.org/10.48550/arXiv.2302.09210

House, J. (1977). *A Model for Translation Quality Assessment*. G. Narr.

House, J. (1997). T*ranslation Quality Assessment: A Model Revisited*. G. Narr.

House, J. (2015). *Translation as Communication across Languages and Cultures* (1st ed.). Routledge.

Hu, K. B., & Li, X. Q. (2023). 大语言模型背景下翻译研究的发展: 问题与前景 [The Development of Translation Studies in the Context of Large Language Models: Issues and Prospects]. *中国翻译* [*Chinese Translators Journal*], *44*(6), 64–73.

Jiang, L., Jiang, Y., & Han, L. (2024). The potential of ChatGPT in translation evaluation: A case study of the Chinese-Portuguese machine translation. *Cadernos de Tradução*, *44*(1), 1–22. https://doi.org/10.5007/2175-7968.2024.e98613

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *Cornell University*. https://doi.org/10.48550/arXiv.2302.14520

Li, Y. B. (2023). 论中华学术外译项目术语库构建的缘由——以《汉代学术史论》(英文版)为例 [The Rationale for Building a Terminology Database for the Chinese Academic Translation Project: A Case Study of *The Academic History of the Han Dynasty* (English Edition)]. *华中学术* [*Central China Humanities*], *15*(1), 228–236.

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *Journal of Academic Librarianship, 49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

Long, Y. Q., & Zhou, X. L. (2024). Exploration of Technology-enabled Terminology Translation and Management: A Case Study on Chinese Academic Translation Project of "C-E Translation of The Transformation of Rural China". *China Terminology, 26*(2), 49–58.

Martínez Mateo, R. (2014). A Deeper Look into Metrics for Translation Quality Assessment (TQA): A Case Study. *Miscelánea: A Journal of English and American Studies, 49*, 73–93. https://doi.org/10.26754/ojs_misc/mj.20148792

O'Brien, S. (2024). Human-Centered augmented translation: against antagonistic dualisms. *Perspectives, 32*(3), 391–406. https://doi.org/10.1080/0907676X.2023.2247423

OpenAI. (2023). GPT-4 technical report. *Cornell University*. https://arxiv.org/abs/2303.08774

Reiss, K. (2000). *Translation Criticism: The Potential and Limitations*. Routledge.

Ren, W., & Li, J. J. (2021). 国家翻译能力研究：概念、要素、意义 [A Study on National Translation Competence: Concepts, Components, and Significance]. *中国翻译* [*Chinese Translators Journal*], *42*(4), 5–14.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Si Xianzhu, Z. (2007). *功能语言学与翻译研究: 翻译质量评估模式建构* [*Translation Studies from the Perspective of Systemic-functional Linguistics*]. Peking University Press.

Sun, P. H. (2023). 做好中华学术外译, 助力对外法治传播 [Enhancing the Translation of Chinese Scholarship to Promote the International Dissemination of the Rule of Law]. *语言与法律研究* [*International Journal of Language, Culture & Law*], *5*(2), 97–122.

Tao, Y. (2020). 外译质量评估的描写译学范式——内涵与路径 [The Descriptive Paradigm of Translation Studies in the Evaluation of Translation Quality into Foreign Languages: Connotations and Approaches]. *跨语言文化研究* [*Cross-Linguistic & Cross-Cultural Studies*], (1), 171–185.

Wang, C. (2025a). *Large Language Models in Translation Quality Assessment: The Feasibility of Human-AI Collaboration* [Data set]. Harvard Dataverse, V1. https://doi.org/10.7910/DVN/YBAMLW

Wang, C. (2025b). *Replication Data for: Large Language Models in Translation Quality Assessment: The Feasibility of Human-AI Collaboration* [Data set]. Harvard Dataverse. https://doi.org/p6x8

Wang, H. S., & Xie, F. (2024). 大语言模型技术驱动下翻译教育实践模式创新研究 [A Study on the Innovation of Translation Education Practice Models Driven by Large Language Model Technology]. *中国翻译* [*Chinese Translators Journal*], *45*(2), 70–78.

Wang, H. S., & Zhang, C. Z. (2025). Translation practice model in the GenAI era: Technological iteration, industrial transformation, and trend outlook. *Foreign Language Education, 46*(1), 53–58. https://doi.org/10.16362/j.cnki.cn61-1023/h.2025.01.011

Wang, J. H., & Shi, J. (2024). 学术译者素养:概念、内涵与提升路径 [Academic Translator Competence: Concept, Connotations, and Paths for Improvement]. *西安外国语大学学报* [*Journal of Xi'an International Studies University*], *32*(4), 91–97.

Wang, S. (2024). The Ethical Risks and Regulations of Generative Artificial Intelligence Prompt Engineering. *Studies in Science of Science.* https://doi.org/10.16192/j.cnki.1003-2053.20241126.003

Wang, S. S. (2017). 翻译质量研究的新视角——《职业化翻译中的质量:评估与改进》述评 [A New Perspective for Translation Quality Research: Review of Quality in Professional

Translation: Assessment and Improvement]. 外国语 [*Journal of Foreign Languages*], *40*(1),108–112.

Wang, S. S., Liu, Z. Q., & Li, D. (2020). 学术外译对话合作模式构建 [Constructing a Dialogic and Collaborative Model for Academic Translation into Foreign Languages]. *上海翻译* [*Shanghai Journal of Translators*], (5), 36–41.

Wei, Y. W., Li, N., & Zhao, L. W. (2022). Quality Standards, Quality Assessment and Development Trend of Machine Translation Based on High Frequency Error Type Analysis. *Computer Science and Application*, *12*(10), 2275–2281. https://doi.org/10.12677/CSA.2022.1210232

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *Cornell University*. https://doi.org/10.48550/arXiv.2302.11382

Williams, M. (2004). *Translation Quality Assessment: An Argumentation-Centered Approach (Perspectives on Translation)*. University of Ottawa Press.

Xu, J., & Mu, L. (2009). 中国翻译学研究 30 年(1978-2007) [Thirty Years of Translation Studies Research in China (1978-2007)]. *外国语* [*Journal of Foreign Languages*], *32*(1), 77–87.

Yao, B., & Friedman, U. D. (2019). 中文社科文献外译的挑战、对策与建议——以《20 世纪中国古代文化经典在域外的传播与影响研究》英译为例 [Challenges, Strategies, and Recommendations for Translating Chinese Social Science Literature into Foreign Languages: A Case Study of the English Translation of *Study on the Extra-Territorial Dissemination and Influence of Ancient Chinese Cultural Classics in the 20th Century*]. *中国翻译* [*Chinese Translators Journal*], *40*(2), 149–156.

Yao, Y. Z. (2024). 大语言模型在汉英技术文献翻译中的应用实证研究 [An Empirical Study on the Application of Large Language Models in Translating Chinese Technical Texts into English]. *翻译界* [*Translation Horizons*], (2), 1–17.

Zhang, J. P., & Zhu, Y. P. (2024). 中华学术外译项目《中国陶瓷史》陶瓷文物图片说明英译研究——以生态翻译学"三维"转换为视角 [A Study on the English Translation of Captions for Ceramic Relic Images in the CATP *History of Chinese Ceramics*: From the Perspective of the "Three-Dimensional" Transformation in Eco-Translatology]. *景德镇陶瓷* [*Jingdezhen Ceramics*], *52*(4), 16–22.

Editorial notes

Authorship contribution

Conceptualization: C. X. Wang
Data collection: C. X. Wang
Data analysis: C. X. Wang
Results and discussion: C. X. Wang
Writing - review and editing: C. X. Wang

Research dataset

The raw data from the 18 test rounds and the four appendices of this research are deposited in the Harvard Dataverse repository (Wang, 2025a, 2025b).

Funding
Chinese Academic Translation Project, NSSFC Grant No. NSSFC23WZSB034.

Image copyright
Not applicable.

Approval by ethics committee
This research is part of Project No. NSSFC23WZSB034 at Nankai University and is fully in compliance with established ethical standards for research.

Conflict of interests
Not applicable.

Data availability statement
The data from this study that are not included in the article are available from the author upon request.

License

Publisher
Cadernos de Tradução is a publication of the Graduate Program in Translation Studies at the Federal University of Santa Catarina. The journal *Cadernos de Tradução* is hosted by the Portal de Periódicos UFSC. The ideas expressed in this paper are the responsibility of its authors and do not necessarily represent the views of the editors or the university.

Guest editors

Section editors

Style editors

Article history
Received: 19-04-2025
Approved: 30-06-2025
Revised: 05-07-2025
Published: 09-2025