Global Language and Cultural Studies, 2026, Volume 1, Issue 1
https://ojs.bbwpublisher.com/index.php/GLCS

English Translation of Ancient Chinese Scientific Classics in the AIPE Era: Efficacy, Boundaries, and Pathways—An Empirical Study Based on the Translation Review of *Mengxi Bitan*

Zipeng Wang, Shiying Liu*

School of Foreign Languages, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing 400065, China

**Corresponding author: Shiying Liu, liusy@cqupt.edu.cn*

Abstract: Current research on Artificial Intelligence Post-Editing (AIPE) remains largely theoretical, lacking sufficient empirical, case-specific, and quantitative analysis. This study investigates the efficacy, boundaries, and developmental pathways of AIPE in translating ancient Chinese scientific classics, using the English translation of *Mengxi Bitan* (*Brush Talks from Dream Brook*) by Shen Kuo (1031-1095) as a case study and integrating quantitative and qualitative methods. The findings reveal that, despite multiple rounds of prompt engineering and Few-Shot Learning (FSL) training, AI continues to exhibit significant limitations: it is prone to factual and contextual hallucinations, lacks empathetic perception, and struggles to capture the emotional and cultural nuances of the source text or adapt to target readers' cognitive gaps. The study concludes that AI currently serves only as an auxiliary tool and is not yet capable of functioning independently as a translation reviewer in the short term. This research aims to fill the empirical gap in AIPE-assisted translation of scientific classics, underscores the necessity of building a human-AI collaborative model, and proposes strategies for enhancing the competence of human reviewers and optimizing AI performance to achieve synchronous improvements in both translation efficiency and quality.

Keywords: AIPE; English translation of scientific classics; *Mengxi Bitan*; hallucination; empathy

Online publication: February 10, 2026

1. Introduction

Machine Translation Post-Editing (MTPE) is the process of revising and refining machine-translated text to meet quality requirements[1]. Building upon MTPE, translation technology is evolving into a new phase known as AIPE. In the AIPE workflow, translators dynamically collaborate with Large Language Models (LLMs) through prompt engineering. This shifts their role from the passive "revisor" in MTPE to an active "guide" in the translation process. Meanwhile, LLMs significantly reduce repetitive tasks by leveraging their deep semantic comprehensio[2]. The fundamental distinction between the two lies in their collaborative logic: MTPE involves a one-way revision of machine output by human translators, following a linear process where humans and machines operate in separate stages. In contrast, AIPE employs multi-turn human-AI dialogue, utilizing a cyclical "prompt-response-quality control" interactive process to iteratively refine the translation, thereby achieving "deep collaboration" and producing high-quality output. Within this context, how to effectively utilize AIPE technology to efficiently and accurately disseminate Chinese scientific and cultural heritage to the world has become a crucial task aligned with the contemporary imperative of "telling China's stories well".

In the AIPE era, with their extensive knowledge bases and efficient processing capabilities, LLMs appear poised to assume a key role in translation quality control. In 2023, the research project "Study on the Collaborative Model between Large Language Models and Human Translation in the Intelligent Era" was initiated. Through a year-long practical project involving the English translation of *Mengxi Bitan* (hereinafter referred to as *Bitan*), the research team discovered that even after multiple rounds of prompt engineering training and the development of customized prompt templates, the AI-assisted translations still contained a significant number of factual, contextual, and cultural errors that the AI itself could not effectively identify, necessitating human intervention for correction. This indicates that while AIPE demonstrates clear advantages in efficiency, it still has fundamental limitations in handling the unique challenges of translating scientific classics, such as linguistic polysemy, culture-loaded terms, and emotional expression. Consequently, the in-depth participation of human translation reviewers remains indispensable at the current stage.

2. Literature Review

Bitan is a masterpiece by Shen Kuo, an outstanding statesman and scientist of the Northern Song Dynasty in China. Its content is extensive, covering multiple disciplines such as physics, chemistry, biology, geography, medicine, literature, and history, making it a canonical work among Chinese scientific and technological classics. The renowned British historian of science Joseph Needham hailed it as "a landmark in the history of Chinese science"[3].

Regarding its English translations, since the mid-19th century, both Chinese and foreign translators have successively undertaken the translation projects. According to previous studies, a total of 13 English translations of *Bitan* have been published to date[4]. Among the abridged versions, apart from Needham's translation which is relatively comprehensive, most others focus only on prominent technological contributions such as the "compass" and "movable-type printing," or on folk tales. Furthermore, these translations exhibit a variety of titles without standardization. In 2008, the collaborative translation by Hong Wang and Zheng Zhao was published by Sichuan People's Publishing House[5], presenting the world's first complete Chinese-English translation.

Theoretical research on its translation began with Hong Wang's description and discussion of translation strategies for this work[6]. Over the past 15 years, research on the English translation of *Bitan* has yielded some results, producing over 20 scholarly articles. However, most focus on static analysis and evaluation of completed translations, with their frameworks still rooted in the traditional triad of "translator-translation-reader." The research perspective has not yet extended to the dynamic translation production process involving AI as a new "collaborator." Besides, most of these studies have primarily focused on the translation strategies for scientific and technical discourse within the text, while little scholarly attention has been paid to the translation of its embedded literary elements, such as poetic verses, rhetorical flourishes, and narrative nuances, which are integral to the work's unique style as a "non-literary text with literary flavor" . For instance, Zonghao Lin and Hong Wang[7], integrating translation ontology and communication methodology, used the English translation of "Yandang Mountain" from *Bitan* as a case study to explore popularization strategies for translating classics. They proposed methods such as multimedia presentation, contextualization, and localization to bridge the gap with general readers, aiming to broaden the audience for Chinese scientific classics and enhance their acceptance and influence in international dissemination. Yanmeng Wang[8], from a translation historiography perspective, examined the first English translations of scientific entries from *Bitan* by 19th-century sinologists Alexander Wylie and Samuel Wells Williams. He pointed out that they hold significant pioneering importance despite the shortcomings in terminological accuracy and cultural understanding due to limitations by the era and their backgrounds. As early practices in Sino-Western scientific exchange, their translation work laid a preliminary foundation for subsequent sinological studies and mutual learning between scientific civilizations. Although previous research has largely concentrated on traditional translation theory, cultural studies, and communication, it is noteworthy that some scholars have shifted their focus to the intersection of translating classics and digital literacy. For example, Peiying Lü[9], in her study on the Russian translation of *Bitan*, employed bibliometric tools and parallel corpus methods to conduct quantitative analysis across dimensions such as terminology translation, syntactic features, and paratexts. This revealed potential pathways for deepening empirical and visual research on translating classics within the digital humanities perspective. However, while this study has taken initial steps towards empirical research in related fields, its core remains confined to using computer assistance for post-translation analysis.

Since the release of ChatGPT at the end of 2022, generative AI (GenAI) products based on LLMs have rapidly emerged worldwide. These models possess exceptional capabilities in language understanding and content generation. Huashu Wang first proposed the concept of AIPE, reflecting the shift in translation technology from "machine-aided human translation" to "human-AI collaboration" in the era of GenAI[10], marking 2024 as the "inaugural year of AIPE." The rise of LLMs has significantly enhanced machine translation performance, yet these models still face challenges in translation tasks[11]. However, existing research has paid insufficient attention to this technological and paradigmatic shift.

Systematically exploring the characteristics, mechanisms, and impacts of AIPE has become a critical issue in the field of translation studies. Furthermore, although previous studies have addressed inherent challenges in translating scientific classics, such as terminological accuracy and the handling of culture-loaded terms, these analyses are often based on human translators' cognitive biases and historical limitations. There is currently a lack of systematic, efficacy evaluation research based on real translation review workflows regarding the potential systemic limitations of LLMs in translating classics and their performance in complex tasks such as semantic understanding, cultural transference, and stylistic adaptation. In other words, what AIPE "can do" and "cannot do" in translating classics remains an area awaiting empirical exploration.

Using the *Bitan* English translation project as a representative case study, this research aims to be the first to systematically investigate the practical application of the AIPE model in translating scientific and technological classics. It will focus on evaluating its assistance efficacy, identifying its potential limitations in areas such as semantic fidelity, cultural transmission, and logical coherence. Based on this, the study will explore possible pathways for deep human-machine collaboration.

3. Research Design

To verify the actual efficacy and limitations of AI-assisted translation under the AIPE model, this study conducted an empirical analysis using the English translation project of the scientific classic *Bitan* as the research sample. The project encompassed approximately 123,000 characters of classical Chinese source text and 182,000 characters of modern Chinese source text. During the self-review process, the research team integrated elements from several prevalent prompt frameworks (such as ICIO, BROKE, CRISPE, etc.)[12]. Through structured prompt engineering, the team guided the AI in conducting preliminary draft reviews. It required multiple rounds of FSL to train the AI to master the project's terminology norms, logic for handling culture-loaded terms, and rules for syntactic transformation, ultimately forming a custom prompt template that met the project requirements (the Prompt design is detailed in the table below). Note that the study is not confined to any specific LLM product but aims to reveal general patterns across various products within the AIPE workflow.

Table 1. Example of translation revision prompt design

Structure	Prompt	Explanation
Role & Profile	"You are a professional translation reviewer specializing in proofreading Chinese-English translations of scientific classics, familiar with the background of ancient Chinese science and culture and the expression paradigms of English scientific texts."	Assigns a clear professional role to the AI, guiding it to execute tasks from the perspective and standards of a specific expert.
Input Data	"Please study the provided Chinese-English terminology database."	Provides the AI with a proprietary knowledge base for retrieval-augmented generation (RAG) to ensure consistency and professionalism in terminology translation.
Background & Context	"Please study this excerpt from [Chapter Title] in *Bitan*: [Classical Chinese + Modern Chinese excerpt + Annotations]"	Provides specific text fragments and background annotations to help the AI understand the specific context and culture-loaded information.
Tasks	"Please check this Chinese-English translation with the following requirements: (1) Check for grammatical errors; (2) Check for over-translation, under-translation, mistranslations and omissions ; (3) Check the accuracy of semantic connotation reproduction; (4) Check the formatting norms for culture-loaded terms and proper nouns; (5) Check the fluency of the translation and its suitability for the reading habits of the target audience."	Provides case examples of human translations and FSL corpus data. Clearly lists the review task checklist, breaking down quality assessment dimensions item by item to make the AI's output more structured and targeted.
Output Format	"If errors are found, please output in a three-column comparative format: 'Source Text Clause-Current Translation-Revision Suggestion'."	Standardizes the AI's response structure, making the feedback results clear and facilitating translators in quickly locating problems and synthesizing suggestions.
Restrictions	"Note: Please output only the sentence segments you consider problematic; revision suggestions must include a brief explanation."	Prevents the AI from outputting meaningless statements like "all correct," ensuring feedback is efficient, focused, and actionable.

Featured by its balanced design, this hybrid framework offers the clarity of structured role and task specification while preserving essential flexibility in addressing the inherent unpredictability of classical texts. Moreover, its iterative prompt-response structure supports progressive refinement throughout the translation review process.

Using the structured prompts above for AI preprocessing of the translations, the study systematically reviewed and recorded the specific performance of AIPE in handling various translation problems, taking the paragraph or semantic chunk of the translated text as the unit of analysis. Based on this, quantitative statistics and classification were conducted on all recorded problems to objectively delineate the efficacy boundaries and error distribution of AIPE. Typical cases were selected for qualitative analysis to reveal the specific mechanisms and underlying causes of the errors, providing a key basis for constructing a human-AI collaboration model.

4. Research Findings

Following AI preprocessing, subsequent human review still identified numerous problems that the AI had failed to detect. After excluding all format-related issues as required by the publisher, the undetected error types revealed two major trends: empathy-related errors (166 cases: 66 related to insufficient empathy for the author; 100 to lacking empathy for the readers) and hallucination-related errors (90 cases: 35 factual hallucinations; 55 contextual hallucination).

The core of empathy-related errors lies in LLMs' lack of human-like perceptual capacity for textual emotion and cognitive differences. They cannot accurately capture the author's expressive intent nor adapt to the reader's comprehension needs, ultimately resulting in inaccurate or ineffective translations. These errors are divided into insufficient author empathy and insufficient reader empathy. The former refers to LLMs' inability to accurately capture the author's intended expression, including emotional tone, aesthetic qualities, and contextual purpose, leading translations to deviate from the author's true intent. The latter refers to LLMs' failure to precisely perceive the cognitive blind spots of target readers, resulting in translations that are difficult for them to understand. By neglecting readers' cognitive differences, the translations fail to effectively convey information, thus losing the communicative value of translation.

The root causes of hallucination-related errors may stem from AI's sycophancy[13], fictionality, and plausibleness. To superficially fulfill the requirement of " completing the translation task," AI may deviate from the factual content or contextual logic of the source text, generating content that appears reasonable but is actually biased. These errors can be further categorized into factual hallucination and contextual hallucination. The former can be understood as a deviation from objective facts or established standards, manifested as the AI' s excessive supplementation or forced inference of information from the source text' s, where such errors conflict withexternal facts or fixed criteria[14]. The latter focuses on the disruption of the source texts contextual logic and semantic coherence, manifested as AI generating content detached from the textual context, leading to a disconnect between the translation and the original context.

Quantitative analysis of the aforementioned errors (see Table 2 for details) yields the following data: "insufficient reader empathy" (100 cases, 39.06%) and "insufficient author empathy" (66 cases, 25.78%) together account for nearly 65% of total errors, constituting the most prominent systemic shortcoming of current AI in translating classics. Within these categories, "cultural default" and "low sensitivity to English-Chinese differences" are high-frequency problems, directly exposing AI's inherent deficits in understanding deep cultural backgrounds and adapting to the cognitive habits of target-language readers. In comparison, "factual hallucination" (35 cases, 13.67%) and "contextual hallucination" (55 cases, 21.48%), while comprising a lower proportion, are equally noteworthy. Particularly within Factual Hallucination, the "category mismatch" alone accounts for 20 cases, revealing the mechanical limitation of AI's reliance on statistical probability rather than cultural-semantic discernment [15]. Furthermore, the high frequency of "task focus shift" and "logical errors" within contextual hallucination further illustrates AI's propensity to deviate when handling the overall logic of complex texts and task priorities.

Table 2. Statistics of AIPE error types

Category	Sub-category	Count	Total	Percentage
Factual Hallucination	Category Mismatch	20	35	13.67%
	Over-translation	8		
	Lack of Idiomaticity	4		
	Lack of Terminological Coherence	3		
Contextual Hallucination	Task Focus Shift	20	55	21.48%
	Logical Errors	17		
	Spelling/Pinyin Errors	9		
	Omission	7		
	Grammatical Errors	2		
Insufficient Author Empathy	Inadequate Grasp of Emotion	30	66	25.78%
	Contextual Comprehension Bias	21		
	Ambiguity	11		
	Aesthetic Deficiency	4		
Insufficient Reader Empathy	Cultural Default	27	100	39.06%
	Low Sensitivity to E-C Differences	26		
	Overuse of Transliteration	18		
	Redundancy	10		
	Insufficient Back-Translationality	8		
	Punctuation Misuse	8		
	Inappropriate Rhetorical Translation	3		
Total	/	256		100%

5. Case Analysis

The experiment reveals that the AIPE exhibits major errors across multiple key dimensions. This section presents four representative cases, each illustrating common deviations and corresponding mitigation strategies in critical areas: semantic category mismatch, task focus shift, inadequate grasp of emotion and contextual comprehension bias.

5.1. Case One

A typical limitation of AIPE in translating cultural terminology is semantic category mismatch.

Classical Chinese: “序乃堂上东西壁，在室之外者。序之外谓之荣，荣，屋翼也。今之两**徘徊**，又谓之两厦……” (as cited from “Critical Thinking” of *Bitan*)

Modern Chinese: “序是指堂屋的东墙、西墙，或者是堂室之外的厢房。序之外称为荣，荣就是墙上和屋檐翼角相结合的部分。现在又把两侧的**回廊**称为两厦……”

Initial Translation: ... *xu* (序) (denotes the eastern and western walls of a house and is situated outside the main chambers. Outside the *xu* lies the *rong* (荣), or roof wings. Nowadays, the two **cloisters** are called *liang-sha* (两厦), or the two wing-rooms of a building...

Human Revision: ... *xu* (序) denotes the eastern and western walls of a house and is situated outside the main chambers. Outside the xu lies the *rong* (荣), or roof wings. Nowadays, the two **side corridors** are called *liang-sha* (两厦), or the two wing-rooms of a building...

In this case, the initial translation renders the " 回 廊 " of Chinese folk architecture as "cloister." While the words are literally related, this incorrectly maps the term into the specific cultural category of Western religious architecture, resulting in an inappropriate contextual transplantation. The AI fails to accurately identify such errors. The human reviewer must follow the strategy proposed by Changshuan Li[16]: when "there is no ready-made translation," one should use the superordinate term "the two side corridors" for a neutral description to avoid cultural misinterpretation. Similar situations frequently occur in the AIPE workflow for *Bitan*, such as the habitual use of "dragon," a term laden with negative connotations, to translate the " 龙 " in Chinese culture, instead of prioritizing the semantically more objective "loong" to minimize ambiguity[17].

5.2. Case Two

AIPE task focus shift mainly manifests in the literary elements within non-literary English translation.

Source Text: "堪恨隋家几帝王 (*wang*)，舞裀揉尽绣鸳鸯 (y*ang*)。" (as cited from "Musical Temperam -ent" of *Bitan*)

Revision Goal: To achieve semantic equivalence, formal equivalence, and phonological equivalence (specifically end rhyme) between the translation and the source text; simplify diction, consider linguistic rhythm, and opt for words with fewer syllables.

Initial Translation:

How lamentable are those emperors of the Sui Dynasty;
The dancer wore out the mandarin ducks embroidered on the mats.

1st LLMs Iteration:

"How I regret those royal reigns long **gone**;
Which trod to pieces love-birds on the **lawn**!"

2nd LLMs Iteration:

How I regret those royal reigns of **old**;
Which trod to shreds the love-birds wrought in **gold**.

3rd LLMs Iteration:

How I hate those Sui lords, proud and **high**,
Who trod the love-birds on the mats to **die**.

…

Human Revision:

I sigh for lords of Sui, in their dance's reckless **sway**;
Who wore the mandarin ducks on their mats **away**.

Note: Here the lines primarily refer to Yang Guang (A.D. 569-618) notorious for his extreme extravagance and sheer brutality, who brought endless calamities upon the people, and ultimately led to the collapse and demise of the Sui Dynasty.

The contextual background of this source text was pre-provided to the LLMs: "Emperor Yang of Sui was indulged in song and dance, and the mandarin duck patterns embroidered on the mats were worn out by the dance steps." However, as demonstrated in the example above, during task execution, the LLMs shift their focus from semantic accuracy to an aesthetic goal—namely, achieving "rhyme", thereby neglecting the fundamental requirement of fidelity. In the first iteration, to create the end rhyme between "gone" and "lawn," the LLMs not only diluted the reference to “ 隋炀帝 ” (Emperor Yang of Sui) by ignoring the context but also mistranslated “ 舞裀 ” (dance mats) as “lawn”, creating an illusion for the translator that the “rhyme requirement had been perfectly met.” Similar to the first iteration, after the problem was pointed out to the LLMs, the second iteration gratuitously added “wrought in gold,” and the third iteration involved over-translation, introducing a metaphorical mapping absent from the source text and causing a shift in semantic focus. As the results after multiple iterations still failed to meet the requirements, a human translator conducted a comprehensive review, supplementing the cultural background of “ 隋炀帝 ”, thereby striking a balance between phonological equivalence and semantic equivalence. Furthermore, this phenomenon also explains why the count for “aesthetic deficiency” in Table 1 (4 cases) is significantly lower than that for “task focus shift” (20 cases): the sycophancy of AI, or its “pleasing mechanism,” fosters a misguided tendency to prioritize producing the so-called “beautiful” translations. Sycophancy essentially means the model believes what the user says is correct[18]. Therefore, human translators must exercise meticulous discernment to distinguish quality and intervene promptly.

5.3. Case Three

LLMs in the AIPE workflow also frequently overlook issues of mismatched evaluative tone, failing to achieve empathy with the author.

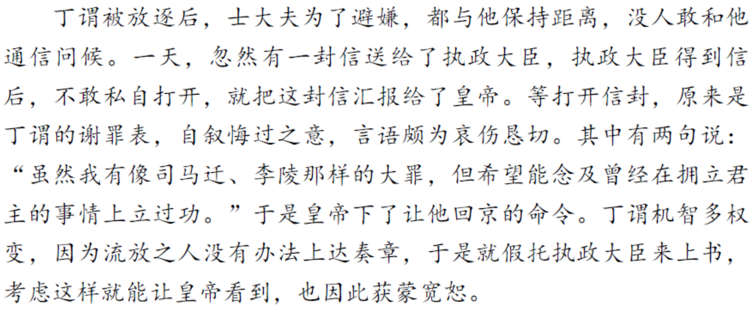
Source Text: “ 丁谓智变 ” (as cited from “Errors” of *Bitan*)

Initial Translation: Ding Wei’s Ingenious and Resourceful Strategies

Context: (Pre-provided to the LLMs, details in **Figure 2)**

Human Revision: Ding Wei’s Cunning/Maneuver (with a footnote added to explain Ding Wei’s life story)

Shen Kuo mentions “ 丁谓 ” in different chapters of *Bitan*, mostly to expose his cunningobsequious nature. However, the LLMs overlooked the positive connotations of “Ingenious” and “Resourceful, failing to reconstruct the authors original intent and mistakenly conveying a positive image of the subject as “resourceful and astute” to readers, which violates translation ethics.



丁谓被放逐后，士大夫为了避嫌，都与他保持距离，没人敢和他通信问候。一天，忽然有一封信送给了执政大臣，执政大臣得到信后，不敢私自打开，就把这封信汇报给了皇帝。等打开信封，原来是丁谓的谢罪表，自叙悔过之意，言语颇为哀伤恳切。其中有两句说：“虽然我有像司马迁、李陵那样的大罪，但希望能念及曾经在拥立君主的事情上立过功。”于是皇帝下了让他回京的命令。丁谓机智多权变，因为流放之人没有办法上达奏章，于是就假托执政大臣来上书，考虑这样就能让皇帝看到，也因此获蒙宽恕。

Figure 2. Vernacular Chinese Translation of “ 丁谓智变 ”

5.4. Case Four

LLMs in the AIPE workflow generally also fail to deeply understand cultural contexts, making it difficult for them to identify content in translations that readers cannot empathize with.

Classical Chinese: “ 冬月作小坐床，冰上拽之，谓之**凌床**。余尝按察河朔，见挽床者相属，问其所用，曰：此运使**凌床**，此提刑**凌床**也。闻者莫不掩口。”（as cited from “Wit and Satire” of *Bitan*）

Vernacular Chinese: “ 冬天的时候，就制作一种小型坐床，在冰上拉着走，称为**凌床**。我曾经察访

河朔一带，见到拉床的人接连不断，问是给谁使用的，说：这是转运使**凌床**，这是提刑**凌床**。听到的人都掩口而笑。”

Initial Translation: In winter, they make small benches and drag them across the ice, called "**ice beds**." When I was inspecting the regions north of the Yellow River, I saw a continuous line of people pulling these **ice beds**. Upon inquiring about their use, I was told, "This one is for the transport official's **ice bed**," and "This one is for the provincial judge's **ice bed**." Everyone who heard this couldn't help but cover their mouths in amusement.

Human Revision: In winter, they make small benches and drag them across the ice, called **"*ling-chuang* (凌床 , ice beds)**." When I was inspecting the regions north of the Yellow River, I saw a continuous line of people pulling these ice beds. Upon inquiring about their use, I was told, "This one is for the transport official's ***ling-chuang***," and "That one is for the provincial judge's ***ling-chuang***." Everyone who heard this couldn't help but cover their mouths in amusement. (**Footnote:** In the Chinese language, "凌床" and "灵床" share the same pronunciation "*ling-chuang*". While the latter means "bier" or "dead wagon". That's why they felt amused.)

In this case, due to linguistic and cultural differences between English and Chinese, "凌床" possesses a degree of untranslatability. Therefore, upon reading the initial translation, target readers are unlikely to appreciate the humor presented by the homophonic pun as deeply as native readers would. The LLMs failed to empathize with the readers, assuming that the initial translation had adequately conveyed the original meanings on a literal level, thus overlooking the serious error of a translation that, in effect, translates nothing meaningful.

6. Analysis of Error Causes

Based on the practical experience of the *Bitan* project, the aforementioned two major error trends are not caused by a single factor but are the result of the combined effects of objective textual characteristics and subjective operational biases. They can be analyzed from the following dimensions.

6.1. Objective Perspective

Objectively, the English translation of ancient Chinese S&T classics inherently presents significant difficulties and challenges:

(1) Ambiguity and polysemy of classical Chinese. Classical Chinese vocabulary is often concise and can carry multiple meanings simultaneously. For instance, terms like "南朝宋" (420-479 A.D.) frequently mentioned in *Bitan* can easily be confused with "南宋" (1127-1279). Without cross-referencing the original text or relying on context, such distinctions are difficult to discern. LLMs struggle to accurately capture this type of context-dependent semantic referent, making them prone to errors related to "insufficient author empathy," such as ambiguity/misinterpretation. Alternatively, forcing a semantic judgment can lead to factual hallucination. Furthermore, the conciseness of classical Chinese sentences, their implicit logical connectives, and complex components make it easy for AI to stray from the original logical flow when supplementing meaning, triggering logical errors or omissions within contextual hallucination.

(2) Complexity and uniqueness of cultural background. Ancient S&T classics carry the cultural connotations and scientific understandings of their specific era. Specialized technical concepts like "浑天仪" (armillary sphere) and "隙积术" (*xi-ji-shu*, the method of piled-up differences) are linked to ancient Chinese astronomy, calendrical systems, and mathematical thought. These culture-loaded contents lack universal translation standards and ready-made equivalents. Moreover, the relevant corpus in AI training data may

be biased or insufficient, making it difficult for AI to accurately convey the cultural essence, thereby leading to errors like cultural default and category mismatch. Simultaneously, the differences between ancient Chinese and Western scientific cognitive system make it challenging for AI to adapt to readers' cognitive blind spots, resulting in issues of insufficient reader empathy.

(3) Significant differences and specialization in linguistic style. *Bitan* combines the rigor of scientific discourse with the conciseness of classical Chinese expression, alongside literary flourishes. Consequently, translations must both accurately convey scientific concepts and replicate the stylistic register of the original. LLMs are prone to imbalance when processing this mixed "literary & non-literary" style. They may become overly literal, leading to aesthetic deficiency, or excessively embellished, causing over-interpretation. Additionally, the significant differences in expression habits between classical and modern Chinese make it difficult for AI to naturally transform sentence structures, resulting in errors like lack of idiomaticity and inappropriate rhetorical translation.

6.2. Subjective Perspective

Subjectively, during the translation practice, unavoidable operational biases arise in the interaction between human translators and AI:

(1) Ambiguity and imperfection in human annotations. In the prompt engineering phase, human annotations are the core basis for AI learning. If annotations fail to clearly define the applicable context of a term (e.g., not specifying that " 方家 " refers specifically to "alchemist" or "Taoist priest"), do not detail the author's expressive intent (e.g., not marking whether the original tone is serious or humorous), or omit key contextual information, it can lead to AI misunderstandings of the task objectives, subsequently causing errors like task focus shift and contextual comprehension bias. Furthermore, inconsistent annotation standards can cause cognitive confusion for AI, increasing the incidence of factual hallucination and empathy-related errors.

(2) Risk of information overload and disorganization in input corpus. The content of ancient scientific classics spans multiple fields (astronomy, pharmacology, physics, etc.). If massive, blended corpus is input to the AI at once without classified by theme, field, or paragraph, it may exceed the contextual processing limits of LLMs. The AI may become confused by the multi-domain information, leading to factual hallucination errors like category mismatch and factual inaccuracies. Simultaneously, disorganized corpus makes it difficult for AI to establish a stable translation style, prone to cause issues like inconsistent linguistic style and lack of idiomaticity.

(3) Insufficient refinement of AI instruction demands. When interacting with AI, if translators do not refine demands according to the characteristics of ancient scientific classics, such as failing to explicitly require the AI to prioritize checking the accuracy of technical terms, not defining the stylistic register of the translation (e.g., "colloquial exposition"), or not prompting scenarios requiring supplemental cultural background annotations, it can cause the AI's translation direction to conflict with project requirements. For example, without specifying the cultural background of the readers, the AI may default to a generic translation approach, triggering risks of insufficient reader empathy like overuse of transliteration and cultural default. Not emphasizing contextual logical consistency can lead the AI to overlook the text's internal connections, producing contextual hallucination.

(4) Potential arbitrariness and oversight in preliminary human review. During the first round of review following

AI preprocessing, if translators rely overly on the AI's output, fail to verify key information (e.g., terminology and historical facts), or if errors are missed due to work fatigue or process simplification, problems uncorrected by the AI can flow into subsequent stages. For instance, if a translator fails to notice the AI's spelling error for " rarely-used Chinese characters like "鮰" and "鮠"" or overlooks contradictions between the translation and the original text's logic, such contextual or factual hallucinations will remain in the text, increasing the workload for subsequent review.

7. Research Implications

Based on the analysis of AIPE error types and their causes in the *Bitan* translation project, AI and human translation reviewers exhibit distinct complementary characteristics in the practice of translating ancient S&T classics. Their respective strengths and weaknesses are concentrated in core dimensions such as efficiency, accuracy, empathy, and adaptability (see Table 3 for details). Concurrently, the project offers insights for enhancing reviewer competence in the AIPE era and for product development by translation technologists.

7.1. Comparative Analysis of LLMs and Human Reviewers

The core advantage of LLMs lies in their efficiency in processing massive amounts of information. They can rapidly conduct preliminary reviews of large-scale projects, such as the English translation of *Bitan*'s 123,000-character classical Chinese text and its 182,000-character modern counterpart. They excel at handling repetitive and standardized content, significantly shortening project cycles. Capable of integrating knowledge across multiple domains and operating around the clock without fatigue, they provide human reviewers with a foundational draft, reducing workload intensity and meeting the timeliness demands of large-scale projects.

The limitations of LLMs are equally significant. They lack empathetic capacity, unable to capture the author's emotional tone and contextual intent or perceive readers' cognitive blind spots, which easily leads to empathy-related errors and makes it difficult to restore the source text's cultural essence. They carry risks of both factual and contextual hallucinations, with errors being covert and difficult to self-correct. They are poorly adapted to the polysemy of classical Chinese and cultural complexity, prone to falling into standardized expressions. Highly dependent on explicit instructions, they are vulnerable to misunderstandings when annotations are vague or demands are insufficiently detailed, exhibiting strong passivity.

The core advantage of human reviewers is their outstanding capacity for empathy and cross-cultural perception. They can accurately capture the author's intent and emotion and adapt to the reader's cultural background, thereby avoiding empathy-related errors. Possessing critical thinking, they can verify facts and identify AI hallucinations and logical deviations. They can flexibly address the challenges posed by classical Chinese and cultural elements, proactively optimize AI interaction to enhance collaborative outcomes, and bear final responsibility for translation quality, mitigating ethical risks.

Human reviewers also have objective shortcomings. Their workload per unit of time is far lower than that of LLMs, making them susceptible to time pressure during large-scale projects, leading to fatigue and an increased likelihood of oversight. Their short-term knowledge reserves are limited, requiring extra verification for obscure terminology, and individuals have knowledge blind spots. Their judgment is influenced by subjective experience or translator habitus, potentially leading to inconsistencies in translation style and review standards; fluctuations in their condition over long-term projects can affect consistency.

Table 3. Comparison of LLMs and human reviewers

Dimension	LLMs (AI)	Human Reviewers
Efficiency	**Strength:** High volume processed per unit time, rapid preliminary review of massive/repetitive/standardized content and 24/7 operation without fatigue.	**Weakness:** Low output per unit time, prone to time pressure in large projects and fatigue from prolonged work leading to oversight errors.
Knowledge Base	**Strength:** Vast corpus data across fields, quickly calling common terminology and basic concept translations, and foundational drafts provided.	**Weakness:** Limited short-term knowledge reserves, extra verification for obscure professional terms required, individual knowledge gaps.
Stability	**Strength:** Standardized and uniform output, free from subjective experience interference, no condition fluctuation in long projects and high consistency.	**Weakness:** Judgment influenced by subjective experience, cognitive bias; inconsistent translation style and review standards and consistency affected by condition fluctuations.
Initiative	**Weakness:** Relying on instruction input, cannot proactively raising questions or supplement information, and lacking independent optimization awareness.	**Strength:** Prompts optimized, supplementary annotations, human-AI collaboration effects enhanced, and bearing final responsibility for quality.
Adaptability	**Weakness:** Poor adaptation to classical Chinese polysemy and cultural complexity, prone to standardized expression, and lacking flexible responsive capability.	**Strength:** Flexibly handling complex contexts, solving challenges in rhetorical translation and sentence structure conversion, and optimizing AI interaction logic.
Empathy	**Weakness:** Lacking empathetic perception, cannot capture author's emotional tone, contextual intent; difficult to perceive reader cognitive blind spots, prone to empathy errors and cannot restoring source culture essence.	**Strength:** Strong empathy and cross-cultural perception, accurately grasping author's expressive intent and emotion, adapting to reader's cultural background, and avoiding empathy errors.
Accuracy	**Weakness:** Prone to factual/contextual hallucinations, covert errors that cannot be self-corrected; relying on explicit instructions and vague annotations or insufficient demands leading to misunderstandings.	**Strength:** Critical thinking possessed, verifying facts through professional knowledge and external databases, and identifying and correcting AI hallucinations.

This comparative analysis underscores that the relationship between LLMs and human reviewers is fundamentally complementary rather than competitive. The efficiency and consistency of AI significantly enhance the throughput of large-scale translation projects, while human expertise remains irreplaceable in areas requiring empathy, critical judgment, and cultural discernment. Therefore, the optimal pathway forward lies not in pursuing the replacement of human expertise by AI, but in strategically integrating their respective strengths within a structured collaborative workflow. Such a human-AI synergy is key to achieving both scalability and high quality in the translation of ancient scientific classics.

7.2. Pathways for Enhancing AIPE Performance

In the AIPE era, the core competence of human reviewers is directly proportional to the quality of content produced through human-AI collaboration. Given a fixed corpus and category of LLM products, the higher the reviewer's competence, the higher the output quality, and vice versa. The core competencies that need enhancement can be summarized as follows:

(1) Consolidating the foundation of traditional translation theory is the core prerequisite. In the AIPE era, reviewers must still systematically study classical translation theory systems, accumulating the core logic and practical methods of translation theories from both past and present, east and west. Only by strengthening this foundational theoretical reserve can they use theory to guide practice, accurately judging quality and identifying the nature of errors the moment AI outputs a translation. Simultaneously, by combining GenAI principles and prompt engineering logic, they can build a compound knowledge system of "traditional translation theory & modern technical understanding," clarifying the technical

boundaries of LLMs to provide effective support for error identification.

(2) Strengthening practical “human-AI collaboration” skills, proficiently applying prompt engineering to optimize AI interaction effects. Master multi-round FSL methods, guiding AI to establish a proprietary knowledge base through structured prompts; cultivate critical thinking, maintaining a skeptical attitude towards AI-generated answers to improve error identification efficiency. For factual hallucinations, focus on verifying the accuracy of proper nouns, culture-loaded terms, and objective facts against external knowledge bases. For contextual hallucinations, focus on contextual logic and semantic consistency, correcting deviations based on the source text context, and optimizing project process management.

(3) Honing empathetic perception and cross-cultural communication skills. On one hand, accurately capture the author’s emotional tone, expressive intent, and aesthetic pursuit to compensate for AI’s insufficiency in author empathy, striving to restore the original text’s emotional nuances and rhetorical value as much as possible. On the other hand, based on the target readers’ cognitive habits, cultural background, and understanding blind spots, balance the transmission of culture-loaded terms with readability during translation, avoiding reader empathy errors such as overuse of transliteration and cultural default, thereby ensuring the effectiveness of the translation’s communicative function.

(4) Establishing rigorous quality control awareness and an ethical responsibility mindset. Facing the “black-box nature” and potential risks of AI, enhance critical thinking, avoid blind reliance on AI output, and assume final review responsibility for translation quality. Proactively guard against potential biases and false information in AI-generated content, upholding the bottom line of translation’s authenticity and objectivity.

Taking the section “Newly Composed Song of Triumph” from “Musical Temperament” in *Bitan* as an example, the following demonstrates how to prepare effectively before AIPE (see the table below) to minimize errors caused by human inconsistency.

Table 4. Case of human annotation for “Newly Composed Song of Triumph”

Source Poem	Stratified Annotation
先取**山西十二州**，	**Geography:** The term “ 山西 ” here does not refer to the modern Shanxi Province. It denotes the area west of Mount Heng (located in present-day Shaanxi). Caution against literal interpretation. The “ 十二州 ” (12 prefectures) were territories lost by the Song Dynasty. Specifically: Suizhou, Yinchou, Xihe, Yihe, Huaicheng, Nanping, Guangyuan, Lanzhou, Jialu, Mizhi, Futu, and Anjiang. This information can be appended as a footnote.
别分**子将打衙头**，	**Military Rank:** “ 子将 ” refers to a deputy or lieutenant general under a senior commander in ancient military organization. In this context, it denotes a mid-to-low-ranking officer commanding approximately 1,000 troops. Caution against category mismatch. **Scenario:** “ 打衙头 ” is a unique expression in the military context of the Song Dynasty, specifically meaning to attack the enemy commander’s tent, i.e., the enemy’s command center. Caution against literal interpretation.
回看**秦塞**低如马，	**History:** “ 秦塞 ” refers to the sites of the ancient Great Wall built during the Qin Dynasty in the Shaanxi-Gansu region, not a specific section of the wall. Here, it metaphorically represents the northwestern frontier fortifications; “ 低如马 ” (low as horses) describes their appearance in the field of view, indicating the army had advanced deep beyond the national border, reclaiming lost territory. Note the emotional and logical relationship in context (the wording should convey the triumphant sentiment of recovering lost land).
渐见黄河**直北流**。	**Orientation:** This depicts the flow characteristic of the Yellow River in the section south of the Qingtong Gorge in present-day Ningxia. In this region, the river flows from south to north. It was a key geographical marker on the Song-Xia border (near the Western Xia capital) and a significant route indicator for the army’s westward advance (building upon the previous line’s emotion for further escalation).

In the AIPE era, translation technology developers also need to make further efforts:

(1) Optimize model performance to reduce the incidence of hallucinations. Focus on the core causes of factual and contextual hallucinations by strengthening the model's fact-checking and contextual comprehension capabilities. Expand high-quality, multi-domain corpora to optimize training data, enhancing accurate mapping of professional terminology, culture-loaded terms, and objective facts. Improve the model's context modeling mechanisms to strengthen control over textual logical relationships and contextual consistency, reducing contextual hallucinations like task focus shift and logical errors. Implement dynamic verification of factual content during the translation process to lower errors such as fabrication and category mismatch.

(2) Enhance the adaptability of model interaction. Develop interactive features that better align with the reviewer's workflow, supporting custom error type annotation and feedback mechanisms. This allows reviewers to directly feed error cases back to the model, achieving a complete "error correction-model optimization" feedback loop. Increase the explainability of model decisions by providing key rationale for translation outputs (e.g., source of terminology translation and logic of semantic understanding) to help reviewers quickly determine error types. Optimize the prompt engineering interface to lower the technical barrier to entry, supporting reviewers in customizing translation rules and quality standards based on different text types (e.g., scientific classics and literary works).

(3) Strengthen empathy and cultural adaptability. Integrate emotional semantic recognition and cross-cultural communication modules into model training to improve perception of textual emotional nuances and rhetorical styles, aiding the model in better capturing authorial intent. Collect translation corpora from multiple languages and cultural backgrounds to optimize the processing logic for culture-loaded terms, providing cultural adaptability suggestions (e.g., supplementary annotations and optimized paraphrasing) to reduce issues like cultural default and inappropriate rhetorical translation. Develop specialized models for specific text types (e.g., ancient classics and lyrical texts) to adapt to their unique expressive paradigms and emotional transmission needs.

(4) Refine the technical ethics and normative framework. Establish AI translation ethical guidelines and error annotation standards to clarify boundaries for models regarding data usage and bias avoidance. Develop error warning functions that provide real-time alerts for high-risk error types (e.g., factual errors and cultural bias), assisting reviewers in rapid identification. Strengthen collaboration with the translation industry and educational institutions to collect error cases and demand feedback from real translation scenarios, promoting the coordinated development of technological iteration and industry standards to achieve both the technological empowerment and humanistic values.

8. Conclusion

Based on the empirical foundation of the *Bitan* English translation project, this study reveals the core limitations of AI-assisted review and proofreading for classics in the AIPE era: even after multiple rounds of prompt engineering and FSL training, AI still struggles to identify hallucination-related and empathy-related errors. As Walter Benjamin stated in "The Task of the Translator": "to release in his own language that pure language which is under the spell of another, to liberate the language imprisoned in a work in his re-creation of that work"[19]. In essence, AI is merely a knowledgeable, polite, and emotionally stable "foreign language learner," lacking the empathetic perception, cross-cultural insight, and critical thinking possessed by human reviewers, rendering it incapable of fulfilling the full mission of translation. This signifies that AI can only serve as an efficient auxiliary tool, and there remains a long journey before it can assume the role of a professional reviewer and truly achieve independent proofreading for English translation projects of Chinese classics.

Research can be further pursued in the following directions: first, expanding the empirical scope by incorporating a wider variety of scientific and technological classics into the AIPE empirical research paradigm for comparative analysis; second, exploring the construction of a comprehensive “human-AI co-translation” quality evaluation system integrating linguistics, cultural studies, and metrics; third, addressing the issues of hallucination and empathy by introducing AI agents, enriching external knowledge bases to handle complex tasks[20], and developing specialized prompt engineering frameworks and model fine-tuning paradigms better suited for ancient text translation.

In the future, the enabling potential of AIPE technology is undeniable, but its application in translating classics is by no means a simple tool replacement. Only by upholding the agency of human translators, continuously enhancing their theoretical grounding, technical collaboration, and ethical responsibility, while persistently advancing AI’s technical iteration in cultural adaptation and logical credibility, can a deep integration of “human-AI collaboration” achieve a dual leap in both efficiency and quality for the translation of Chinese classics into foreign languages.

Funding

This research was supported by Special Research Project on Foreign Language Education and Teaching Reform in Universities (Project No.: No. 03, 2024, Chongqing), Research on the Talent Training Mode for Language Service Engineers in the “Foreign Language + AI + Industry” Context.

Disclosure statement

The author declares no conflict of interest.

References

[1] Wang H, Zhu H, 2025, *Artificial Intelligence Translation and Post-editing*. Foreign Language Teaching and Research Press, 51.

[2] Wang H, 2025, From MTPE to AIPE: The Evolution of Translation Models in the GenAI Era and Its Implications for Translation Education. *Shandong Foreign Language Teaching*, 46(3): 111-121.

[3] Needham J, 1954, *Science and Civilisation in China, Vol. 1*. Cambridge University Press, 135.

[4] Wang YM, 2022, *A Multidimensional Comparative Study on the English Translation of Tiangong Kaiwu*. China Social Sciences Press, 246.

[5] Wang H, Zhao Z, 2008, *Library of Chinese Classics: Dream Pool Essays (Chinese-English)*. Sichuan People’s Publishing House.

[6] Wang H, 2010, A Study on Translation Strategies for the English Version of Mengxi Bitan. *Shanghai Journal of Translators*, (01): 18-22.

[7] Lin ZH, Wang H, 2019, A Study on Popularization Strategies for the English Translation of Shen Kuo’s Mengxi

Bitan: Yandang Mountain. *Translation Forum*, (01): 19-21.

[8] Wang YM, 2023, A Study on the English Translation of Scientific Entries from Mengxi Bitan by British and American Missionaries. *Journal of Dialectics of Nature*, 45(05): 89-95.

[9] Lü PY, 2023, The Dissemination and Reception of the Translation of Mengxi Bitan in Russia from the Perspective of Digital Humanities (Master's thesis, Dalian Maritime University).

[10] Liu SJ, Zhang Y, 2025, From MTPE to HACT: A Study on Translation Process Innovation Driven by Large Language Models. *Foreign Language Education Research*, 13(1): 18-26.

[11] Pang J, et al., 2025, Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. *Transactions of the Association for Computational Linguistics*, 13: 73-95.

[12] Wang ZP, Liu SY, 2024, When ChatGPT 4.0 Encounters Traditional Chinese Medical Texts—Taking the English Translation of Huangdi Neijing: Suwen as an Example. In *The 28th Global Chinese Conference on Computers in Education*, 698.

[13] Petrov I, Dekoninck J, Vechev M, 2024, *BrokenMath: A Benchmark for Sycophancy in Theorem Proving with LLMs*. arXiv preprint arXiv:2510.04721.

[14] Li Q, Cui LZ, Liu L, 2025, *Understanding and Application of Large Language Models*. China Machine Press, 91.

[15] Miracchi Titus L, 2024, Does ChatGPT Have Semantic Understanding? A Problem with the Statistics-of-Occurrence Strategy. *Cognitive Systems Research*, 83: 101174.

[16] Li CS, 2022, *Theory and Practice of Non-Literary Translation: Understanding, Expression, Adaptation*. China Translation & Publishing House, 167.

[17] Zhang Z, 2024, How to Say " 龙 年 " in English? Dragon or Loong?. *China Daily Bilingual News*. Available at: https://mp.weixin.qq.com/s/4I1Lgqi6eJhCHR23j_ay6w.

[18] Naddaf M, 2025, AI Chatbots Are Sycophants—Researchers Say It's Harming Science. *Nature*, 647: 13-14.

[19] Xie TZ, 2009, *A Brief History of Translation in China and the West*. Foreign Language Teaching and Research Press, 30.

[20] Zhao JF, Li X, 2024, Research on the Construction and Application of Translation Agents Driven by Large Language Models. *Technology Enhanced Foreign Language Education*, (5): 22-28.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.