Xiong Shao* and Yue'e Zhu

The assistance role of LLMs and NMT in student translators' Chinese–English post-editing: differences in workload, translation quality and user perception

https://doi.org/10.1515/jccall-2025-0014
Received April 23, 2025; accepted August 11, 2025; published online September 29, 2025

Abstract: This study examines the workload, translation quality, and user perception associated with Chinese–English (C–E) post-editing (PE) performed by student translators using three tools: DeepL, ChatGPT-4o, and DeepSeek-V3. Thirty Master of Translation and Interpreting (MTI) students from a comprehensive university in China participated in the study. Source texts of varying types, but similar complexity level, served as the materials for raw machine translation outputs. By analyzing variables such as task duration, keyboard events and pause length, translation quality assessment data of 120 post-edited files, and questionnaire responses, the study yielded the following findings: (1) The use of neural machine translation (NMT) or large language models (LLMs) did not significantly impact overall PE workload, though students assisted by LLMs required fewer keyboard events when editing popular science texts; (2) LLM-assisted PE demonstrated higher quality than those edited with DeepL, as evidenced by fewer error counts and lower error scores; (3) Participants perceived ChatGPT-4o to be the most effective tool for error correcting ability and task efficiency. DeepSeek-V3 was rated highest in terms of user experience, while DeepL was regarded as the most reliable in terms of information accuracy. By comparing students' PE workload, translation quality and user perception across different scenarios, this study offers new insights into human–AI collaboration in PE workflows and its implications for translation pedagogy in the AI era.

Keywords: post-editing; large language models; student translator; comparative study

***Corresponding author: Xiong Shao**, College of Foreign Languages, Central South University of Forestry and Technology, Changsha, China, E-mail: 1300913542@qq.com. https://orcid.org/0009-0001-9446-6295
Yue'e Zhu, College of Foreign Languages, Central South University of Forestry and Technology, Changsha, China, E-mail: 445447701@qq.com

1 Introduction

The emergence of generative artificial intelligence (AI) has significantly transformed the landscape of machine translation (MT) and translation education. Large language models (LLMs), such as OpenAI's ChatGPT, represent a milestone in this evolution, achieving unprecedented performance in generating accurate, fluent, and contextually appropriate translations. These advances have not only revitalized MT research (Jiao et al. 2023) but also promoted the development of interactive post-editing (PE) practices, posing both opportunities and challenges for translation pedagogy and translation workflows in the AI era (Hu and Li 2024; Wang and Wang 2023). Translation service providers, including SDL Trados and Lionbridge, have swiftly adopted AI-powered tools, and consequently, translation workflows have evolved from human translation, computer-assisted translation, and human-machine collaborative translation to a new model of human-AI collaborative translation (Wang 2024; Wang and Zhang 2025).

While AI technologies continue to reshape the translation landscape and offer new pathways for human–AI collaborative translation (Geng and Hu 2023), existing research mostly focused on evaluating the output quality of LLMs – such as ChatGPT and Gemini – and analyzing LLM-assisted PE processes (Farghal and Haider 2024; Gao et al. 2024; Hendy et al. 2023; Zhong et al. 2024). These studies highlight the remarkable potential of LLMs in translation tasks. However, as Huang (2022) cautions, "Translation technologies should serve as tools to assist human translators, with full respect to their agency." Therefore, in the midst of AI-driven innovations, it is essential to consider not only the capabilities of AI systems but also human performance and perceptions of engaging with such tools.

Meanwhile, LLMs developed in China, such as DeepSeek, have shown promising results in translation tasks (Liao 2025) and deserve further empirical investigations. In parallel, although NMT systems may lack interactivity, they continue to offer stable and high-quality outputs through advanced translation algorithms (Adawiyah et al. 2023), making them a relevant comparator to LLMs in PE tasks.

Against this backdrop, the present study aims to examine the assistance role of LLMs and NMT tools in Chinese–English (C–E) PE. Focusing on the performance and perceptions of Master of Translation and Interpreting (MTI) students, the study compares three widely used tools – DeepL, ChatGPT-4o, and DeepSeek-V3 – in terms of their impact on PE workload, translation quality, and user perception. A mixed-methods approach is adopted, integrating keystroke logging, error annotation, screen recording, and questionnaire surveys to provide a comprehensive understanding of the assistance role of different tools used by student translators in PE practices.

2 Literature review

2.1 LLMs in translation practice

LLMs, a type of language models that utilizes neural networks containing billions of parameters, are generative mathematical models of the statistical distribution of tokens in the vast public corpus of human-generated text. Their abilities have been proved in various language-related tasks, including text synthesis, translation, summarization, question-answering, and sentiment analysis, by leveraging deep learning techniques and large datasets (Raiaan et al. 2024; Shanahan 2024).

In the context of translation, LLMs offer assistance through two primary affordances: access to expansive knowledge bases and real-time interactive feedback. Compared with traditional NMT systems, LLMs are pretrained on diverse and extensive textual corpora, which allows them to capture various forms of knowledge – including event, relational, factual, and commonsense knowledge (Da et al. 2021; Han and Chai 2024; Heinzerling and Inui 2021; Kauf et al. 2023; Safavi and Koutra 2021). This breadth of information enables LLMs to support translators in generating target texts that are more contextually relevant and factually accurate, thereby improving both translation quality and efficiency.

A further advantage of LLMs lies in their capacity for real-time, dialogic interaction. Through multi-turn conversation, these models can tailor feedback to users' individual needs, cognitive styles, and translation goals. Techniques such as chain-of-thought prompting simulate human cognitive processing, thereby enhancing the depth and personalization of the interaction (Dai et al. 2023; Lu and Chen 2024). Within translation contexts, feedback serves not merely a corrective function but also a pedagogical one – enabling translators to read their translation work from the perspectives of readers or users, and fostering the capacity for self-directedness and self-assessment (Washbourne 2014).

ChatGPT, as a prominent example of LLMs, can provide real-time feedback and offer targeted suggestions for revisions (Dai et al. 2023). It is characterized by short response time, high efficiency, numerous outputs, diverse options, and personalized, context-aware supports. Previous empirical studies demonstrate that both post-graduate and graduate students of translation major significantly improve their translation quality through revision assisted by ChatGPT's interactive feedback (Zhu and Shao 2024, 2025).

Although ChatGPT and other models have demonstrated strong translation competence and are increasingly adopted in both translation practice and academic research, relatively few studies have focused on LLMs developed in China. Their

potential to support student translators during PE – especially in comparison with established NMT systems – remains under-investigated.

2.2 AI-assisted PE

PE refers to the systematic evaluation, refinement, and modification of MT outputs to meet specific quality standards. This process involves correcting errors, improving accuracy, and enhancing overall readability (Bowker and Ciro 2019; Feng and Cui 2016; ISO 2014). PE has become an integral part of modern translation workflows, especially in the context of increasingly advanced AI systems.

Operationally, PE is typically categorized into two levels: light post-editing (LPE) and full post-editing (FPE). LPE emphasizes semantic correctness and basic comprehensibility, tolerating some grammatical or syntactic inaccuracies as long as the core information remains intact. The output may retain traces of machine-generated style, with imperfect grammar or unnatural phrasing, but is expected to be accurate in content. In contrast, FPE aims for grammatically, syntactically, and semantically correct translations that are also stylistically appropriate. While the final product may not match the stylistic nuance of a native-speaker human translator, it is expected to meet professional translation standards in terms of fluency, coherence, and linguistic correctness (Nitzke and Gros 2020).

As Pym (2012) observed, "Statistical-based MT, along with its many hybrids, is destined to turn most translators into post-editors one day, perhaps soon." This prediction has materialized in part with the rise of LLMs, which have shifted translation from a linear, one-directional process to a dynamic, interactive, and increasingly intelligent collaboration between humans and AI systems. Translators now routinely engage with AI-driven tools to improve productivity and output quality (Wang et al. 2023).

Trained on multilingual corpora, LLMs, such as ChatGPT, are capable of managing complex linguistic patterns and generating fluent, contextually appropriate translations in PE tasks. Their strength lies in handling a wide range of general-domain content with coherence and high speed. However, limitations persist, particularly in relation to domain-specific texts or distant language pairs, where LLMs' outputs often fail to match human-level precision or stylistic appropriateness (Bhattacharyya et al. 2023; Li and Li 2025). In these cases, human intervention through PE remains indispensable. The integration of human expertise and AI assistance continues to be a promising and complementary approach in PE workflows.

Recent studies have examined PE from multiple perspectives, including strategies for improving translation quality and reducing cognitive load (Fan and Yang

2024; Geng 2024; Shin and Chon 2023), comparisons between human translation and MT output (Jia and Sun 2022; Wang et al. 2024), the relationship between PE and translator effort (Lu and Sun 2018; Wang and Wang 2024; Zhong et al. 2024), and the application of PE in translation education (Feng and Liu 2018; Wang and Wang 2023; Zhong and Shu 2020). While the emergence of LLMs has opened a new avenue for human–AI collaborative PE (Geng and Hu 2023), empirical studies in this area remain limited (Khasawneh and Khasawneh 2023). Understanding the assistance roles of different tools in PE – particularly their effectiveness in assisting students during practices – is of significant importance for both the improvement of translation competence and the advancement of translation technologies.

To address this gap, the present study adopts a mixed-methods approach, incorporating keystroke logging, error annotation, screen recording, and questionnaire surveys, and recruits MTI students from a comprehensive university in Central China to participate in the study. They conducted full post-editing tasks on C–E MT output, using both NMT and LLM tools – specifically, DeepL, ChatGPT-4o, and DeepSeek-V3. The investigation focuses on three dimensions: workload, translation quality, and user perception. By evaluating student performance and perceptions across these dimensions, the study aims to shed light on the assistance role of different tools in supporting student translators' PE. Ultimately, the findings are expected to contribute to the optimization of PE strategies, the development of student translators' PE competence, and the innovation of AI-integrated translation pedagogy.

3 Research design

This study employed a quasi-experimental design with the intent of minimizing the impact of potential variables on experimental results as much as possible. It aims to investigate the assistance role of LLMs and NMT in student translators' PE in terms of workload, translation quality and user perception by answering the following research questions (RQs):

RQ1: Do student translators experience different workloads when assisted by LLMs and NMT during PE tasks?

RQ2: Does the post-edited translation quality by student translators differ when assisted by different tools?

RQ3: How do student translators perceive DeepL, ChatGPT-4o and DeepSeek-V3 in supporting PE?

3.1 Participants

This study recruited 30 MTI students as participants. Among them, 5 were male and 25 were female, with an average age of 23.6 years (SD = 0.93). A total of 100 % had passed the Test for English Majors Band 8 (TEM-8), indicating that their language and translation proficiency met the general standard for professional translators. All participants were native Chinese speakers with English as their second language. The pre-experiment questionnaire confirmed that all participants were skilled at typing and had not previously read the experimental texts. Informed consent was obtained from all participants before the experiment, and they were compensated for their participation.

3.2 Materials

3.2.1 Complexity measurement

Given that the complexity of source texts can significantly affect the difficulty level of PE tasks (Jia and Zheng 2022), and overly lengthy passages may lead to participants' fatigue, while overly short sentence lengths may leave few opportunities for PE (Daems et al. 2017a, 2017b), this study considered both lexical and syntactic complexity in text selection.

Lexical complexity was assessed using the total number of characters and the root type-token ratio (RTTR). Syntactic complexity was evaluated through three commonly used indicators: mean length of T-unit (MLTU), mean length of clause (MLC), and mean length of sentence (MLS). These five indices were calculated using the L2C-Rater, an automated Chinese text analysis tool developed by Beijing Normal University (Wang and Hu 2021). This approach ensured a comprehensive and objective measurement of source text complexity.

3.2.2 Source texts selection

Since the study focused on C–E PE and all participants were native speakers of Chinese, four source texts in Chinese were selected, each representing a distinct genre to reflect a variety of linguistic and stylistic features commonly encountered in real translation practices. They were excerpts from the bilingual introduction of Lu Ban Lock in the China Science and Technology Museum, *Modern Chinese Essays Volume I* ("英译中国现代散文选 (一)" in Chinese), *The Backstage Clan* ("装台" in

Table 1: Source text information.

Indices	Text1 Popular science	Text2 Prose	Text3 Novel	Text4 Academic work	Mean	SD
Character Num	162.00	164.00	176.00	160.00	165.50	7.19
Lexical RTTR	8.06	7.77	8.42	7.76	8.00	0.31
MLTU	16.20	11.71	11.00	10.00	12.23	2.74
MLC	13.50	7.81	9.78	7.27	9.59	2.82
MLS	54.00	41.00	58.67	53.33	51.75	7.55

Chinese), and *Fundamentals of Chinese Culture* ("中国文化要义" in Chinese) respectively (Chen 2019, 2023; Liang 2018, 2021; Zhang 2007). Each text had similar levels of complexity (see Table 1). Some examples from the source texts, and their English translations were shown in Appendix A.

3.2.3 Translation quality evaluation rules

Building on prior research in error typology and annotation rules (Li 2022; Secară 2005; Zhang and Zhao 2024; Zhu and Shao 2024), this study developed a set of error-based human evaluation rules, adapted to the linguistic and genre-specific features of the experimental texts (see Appendix B). To ensure the reliability and objectivity of the evaluation, error annotation was conducted by four postgraduate students with extensive experience in translation practice, each holding Level II certification in the China Accreditation Test for Translators and Interpreters (CATTI). The process was guided by a senior professor in translation studies.

Each post-edited translation was systematically annotated for errors and scored according to the established rules. In detail, all culture type errors were classified as major errors due to their potential to disrupt intercultural communication. For other error types, a two-level severity scale was adopted:

Major errors (5 points): Errors that distorted meaning, misled readers, or significantly hindered comprehension.

Minor errors (1 point): Errors that affected fluency, style, or appropriateness without obstructing understanding.

To prevent inflated scores, repeated occurrences of the same error within a single sentence were only counted once. Cross-evaluation among the four raters was employed to mitigate potential bias and enhance inter-rater consistency in assessing translation quality.

3.3 Experimental procedures

The 30 participants were randomly assigned to three groups.
Group 1 (G1) used the traditional NMT tool DeepL, with the option to consult online resources such as Baidu.com for reference when needed.
Group 2 (G2) was assisted by ChatGPT-4o developed by OpenAI.
Group 3 (G3) worked with DeepSeek-V3 developed by Hangzhou DeepSeek Artificial Intelligence Co., Ltd.

Prior to the formal experiment, each group received targeted technical training sessions, aiming to reduce disparities in participants' familiarity with technological tools and PE-related knowledge, thus minimizing their potential impact on experimental outcomes.

As Youdao Translate has relatively stable performance across various kinds of texts (Dai and Liu 2024), it was employed to generate the raw machine translations of the source texts and some examples were shown in Appendix A. In Translog-II, the source texts and their MTs were displayed as Figure 1. Participants needed to complete the PE tasks of all four texts on Translog-II User as quickly as possible.

After completing the tasks, participants' PE files were collected from the three groups, yielding 120 translation products along with their editing process records, and then, participants were asked to complete a brief questionnaire designed to collect their subjective evaluations of the error correcting capabilities of each tool, as well as of the tools' advantages and disadvantages in supporting their PE tasks.

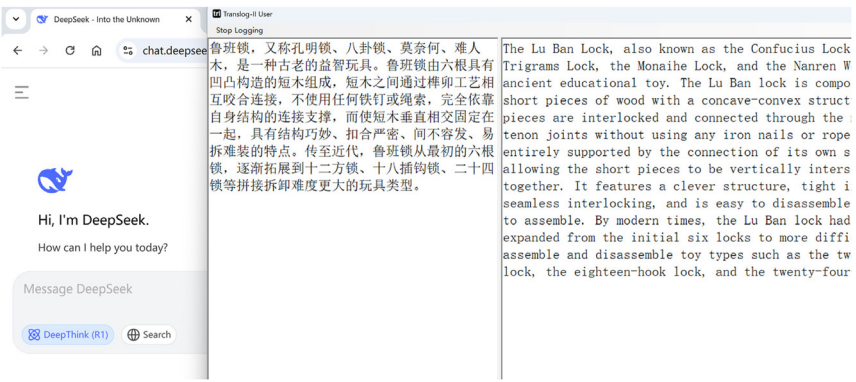

Figure 1: An example of the student translator's post-editing interface.

4 Results

After collecting the experiment data from each group, we conducted a comparative analysis from the three perspectives: workload, translation quality, and user perception based on keystroke logging, screen recording, error annotation and questionnaire results.

4.1 Workload

Krings (2001) classified the workload of PE into temporal, technical, and cognitive efforts. Temporal effort refers to the time needed for PE, technical effort relates to deletions, insertions, and other mechanical operations, while cognitive effort involves the mental processing activated during PE. In this study, to measure these three types of efforts, we adopted task duration, keyboard events, and pause length, which were recorded and processed using Translog-II and CRITT TPR-DB (Carl et al. 2016).

First, the normality of the three types of effort data was examined using the Shapiro–Wilk test in SPSS 27.0 for Windows. A p-value higher than 0.05 indicated that the data followed a normal distribution, while a p-value lower than or equal to 0.05 suggested a non-normal distribution. Subsequently, a one-way ANOVA was conducted for normally distributed data, and the Kruskal–Wallis test was employed for non-normally distributed data. When a significant difference was found, post hoc comparisons were conducted using the Bonferroni test for parametric data and the Mann–Whitney U test for non-parametric data, in order to determine whether the use of NMT or LLMs had a significant impact on student translators' performance.

Task duration refers to the time needed to complete PE tasks, and a longer task duration indicates more temporal effort. Data analysis revealed that there was no significant difference ($p = 0.228$) in total task duration among the three groups when text type was not taken into account. Similarly, when text type was considered, there were also no differences in task duration across different groups for popular science ($p = 0.147$), prose ($p = 0.543$), novel ($p = 0.256$), or academic work text ($p = 0.738$) tasks.

Keyboard events refer to translators' keyboard operations, and a higher number of events indicates greater technical efforts. Based on comparative statistics, there was no significant difference in the total number of keyboard events among the three groups ($p = 0.079$).

However, when analyzed by text type, a significant difference ($p = 0.008$) was observed in the number of events for the Text 1 task. Based on Table 2, the post hoc comparison result indicated that, students of $G2$ and $G3$ required fewer keyboard

Table 2: Comparison of total keyboard events of *G*1, *G*2 and *G*3 in Text 1 task.

*G*1			*G*2			*G*3			Sig.	Post-hoc
Sum	***M***	**SD**	**Sum**	***M***	**SD**	**Sum**	***M***	**SD**		
43	4.30	0.48	30	3.00	1.15	27	2.70	1.16	0.008*	1 > 2** 1 > 3* 2 = 3

Note: *$p < 0.05$, **$p < 0.01$.

events than that of *G*1 in this task. Nevertheless, for Text 2 ($p = 0.225$), Text 3 ($p = 0.098$), and Text 4 ($p = 0.338$), no significant differences in the number of keyboard events were observed among the groups.

In terms of cognitive effort, a threshold of 1,000 ms was adopted to filter out shorter pause lengths, and only pause lengths equal to or longer than this threshold were included in the analysis. The longer the pause length, the greater the cognitive effort required during PE tasks (Kumpulainen 2015).

Our results showed that, when text type was not taken into consideration, there was no significant difference ($p = 0.965$) in total pause lengths among the three groups. Similarly, when examining student translators' cognitive effort across different text types, no significant differences in pause lengths were found among the groups for Text 1 ($p = 0.995$), Text 2 ($p = 0.804$), Text 3 ($p = 0.924$), or Text 4 ($p = 0.886$).

In summary, from the perspective of PE workload, the findings suggest that when student translators engage in PE tasks involving texts of similar complexity, the use of either LLMs or NMT does not lead to significant differences in total temporal, technical, or cognitive efforts. The only notable exception was found in the PE task involving the popular science text, where students assisted by ChatGPT-4o and DeepSeek-V3 completed the task with fewer keyboard events compared to those using DeepL. This may be attributed to the fact that popular science texts – unlike prose, novels, or academic work texts – are generally easier for LLMs to process.

4.2 Translation quality

Although translation quality assessment involves both automatic and human evaluation (Yang 2012), our prior experimental experience and the results of automatic translation evaluation in this study suggest that automatic metrics such as BLEU, TER, and METEOR – while capable of reflecting quality differences – are insufficient for capturing the PE process or identifying specific error types (Daems et al. 2017a, 2017b). Therefore, this study only reports the results of human evaluation results.

4.2.1 Error counts

First, we calculated the error counts in student translators' outputs based on the Human Evaluation Rules (see Appendix B) and adopted the same data processing methods to compare the three groups' data using SPSS 27.0 for Windows.

According to Tables 3 and 4, the total error counts in students' post-edited translations decreased across all three groups compared with the raw MT outputs. However, statistically significant differences were observed in total error counts between *G*1 and *G*2, as well as between *G*1 and *G*3. Specifically, the total number of errors in *G*2 or *G*3's translations was less than that of *G*1, with a more pronounced difference between *G*1 and *G*3. This suggested that assistance from LLMs – particularly DeepSeek-V3 – led to fewer translation error counts across the four PE tasks compared to assistance from the NMT tool.

In terms of the five error types, differences were also observed across the groups. For completeness, accuracy, and style errors, students in *G*1 produced more errors than those in *G*2 and *G*3. In the case of culture errors, with the assistance of ChatGPT-4o, students produced fewer errors than when using DeepL. For language errors, no significant differences were found among the three groups, indicating that all tools were comparably helpful in addressing grammatical, spelling, and incorrect word choice errors.

Given that the PE tasks involved four different types of texts, the effectiveness of assistance also varied by text type. As shown in Table 5, significant differences in translation error counts were found between *G*1 and *G*2, *G*1 and *G*3 across all texts, with *G*1 consistently producing more errors. Further comparative analysis on error types revealed that, in the case of Text 1, student translators assisted by both ChatGPT-4o ($p_{\text{completeness}}$ = 0.039; p_{accuracy} = 0.003) and DeepSeek-V3 ($p_{\text{completeness}}$ = 0.012; p_{accuracy} = 0.03) produced less completeness and accuracy types of errors

Table 3: Error counts and scores of MT outputs.

Error type	Text 1		Text 2		Text 3		Text 4		Total	
	Sum	Score	Sum	Score	Sum	Score	Sum	Score	Sum	Score
Completeness	2	10	0	0	3	15	0	0	5	25
Accuracy	6	26	6	30	6	22	6	26	24	104
Language	2	10	0	0	2	10	2	10	6	30
Culture	0	0	2	10	4	20	0	0	5	30
Style	1	5	3	3	1	5	3	11	8	24
Total	11	51	11	43	16	61	11	47	48	232

Table 4: Error counts in the five error types.

Error type	*G*1			*G*2			*G*3			Sig.	Post-hoc
	Sum	*M*	SD	Sum	*M*	SD	Sum	*M*	SD		
Completeness	43	4.30	0.48	30	3.00	1.15	27	2.70	1.16	0.003*	1 > 2* 1 > 3* 2 = 3
Accuracy	184	18.40	2.41	135	13.50	2.80	130	13.00	1.34	0.001***	1 > 2** 1 > 3*** 2 = 3
Language	65	6.50	1.51	54	5.40	2.46	47	4.70	1.49	0.070	
Culture	40	4.00	1.05	28	2.80	0.79	30	3.00	0.94	0.030*	1 > 2* 2 = 3
Style	57	5.70	1.16	27	2.70	1.34	24	2.40	2.01	0.001***	1 > 2*** 1 > 3*** 2 = 3
Total	389	38.90	4.79	274	27.40	6.50	258	25.80	6.34	0.001***	1 > 2** 1 > 3*** 2 = 3

Note: *p < 0.05, **p < 0.01, ***p < 0.001.

Table 5: Error counts in the four texts.

Text type	*G*1			*G*2			*G*3			Sig.	Post-hoc
	Sum	*M*	SD	Sum	*M*	SD	Sum	*M*	SD		
Text1	88	8.80	1.81	65	6.50	1.08	61	6.10	2.33	0.005**	1 > 2* 1 > 3** 2 = 3
Text2	82	8.20	1.75	67	6.70	0.95	54	5.40	2.46	0.008**	1 > 2* 1 > 3*
Text3	106	10.60	2.01	68	6.80	3.88	73	7.30	3.27	0.018*	1 > 2* 1 > 3**
Text4	102	10.20	1.32	64	6.40	2.76	58	5.80	2.20	0.001***	1 > 2** 1 > 3***

Note: *p < 0.05, **p < 0.01, ***p < 0.001.

compared with those assisted by DeepL. For Text 2, ChatGPT-4o (p_{style} = 0.001) and DeepSeek-V3 (p_{style} = 0.008) both helped students reduce style type errors, and DeepSeek-V3 (p_{accurary} = 0.01) was slightly more effective in the accuracy respect. In Text 3, ChatGPT-4o significantly helped them reduce errors in accuracy and culture (p_{accuracy} = 0.031; p_{culture} = 0.014) aspects, while DeepSeek-V3 contributed to reducing accuracy errors (p_{accuracy} = 0.04). In Text 4, both ChatGPT-4o (p_{accuracy} = 0.002) and DeepSeek-V3 (p_{accuracy} = 0.004) effectively helped reduce accuracy errors.

4.2.2 Error scores

Secondly, we calculated each student's error scores. According to Tables 3 and 6, three groups of students' error scores were lower than those of MTs, and no significant difference was found between the total error scores of *G*2 and *G*3, while both groups differed from *G*1. The differences of total error scores appeared between *G*1 and *G*2, *G*1 and *G*3, indicating that with LLM assistance, student translators

Table 6: Error scores in the five error types.

Error type	*G*1			*G*2			*G*3			Sig.	Post-hoc
	Sum	*M*	SD	Sum	*M*	SD	Sum	*M*	SD		
Completeness	195	19.5	3.54	152	15.2	4.61	131	13.1	4.89	0.005**	1 > 3**
Accuracy	736	73.6	9.96	608	60.8	10.36	523	52.3	11.99	0.001***	1 > 2* 1 > 3***
Language	286	28.6	7.89	235	23.5	8.97	213	21.3	6.7	0.063	
Culture	200	20	5.27	155	15.5	3.69	165	16.5	5.3	0.121	
Style	133	13.3	4.42	63	6.3	4.16	56	5.6	4.06	0.001***	1 > 2** 1 > 3** 2 = 3
Total	1,550	155	23.59	1,213	121.3	18.55	1,088	108.8	23.56	0.001***	1 > 2** 1 > 3***

produced lower error scores in total than with NMT. Furthermore, the absence of differences between *G*2 and *G*3 suggested that both LLMs were similarly effective. When text type was not considered, LLMs consistently helped reduce accuracy and style error scores, while DeepSeek-V3 was particularly effective in reducing completeness error scores.

Table 7 further demonstrates that, when assisted by LLMs, students produced lower error scores in PE tasks involving popular science, novels, and academic work text tasks compared to when assisted by NMT. Specifically, for the popular science text, students with DeepSeek-V3 produced lower completeness ($p_{\text{completeness}} = 0.014$) and accuracy ($p_{\text{accuracy}} = 0.014$) error scores. In the prose text, ChatGPT-4o ($p_{\text{style}} = 0.001$) helped them reduce style error scores, while DeepSeek-V3 ($p_{\text{accuracy}} = 0.01$) outperformed ChatGPT-4o in helping reduce accuracy error scores. In the novel text, ChatGPT-4o ($p_{\text{culture}} = 0.021$) contributed to a reduction in culture error scores. For the academic work text, both ChatGPT-4o ($p_{\text{accuracy}} = 0.004$; $p_{\text{style}} = 0.002$)

Table 7: Error scores in the four texts.

Text type	*G*1			*G*2			*G*3			Sig.	Post-hoc
	Sum	*M*	SD	Sum	*M*	SD	Sum	*M*	SD		
Text1	455	45.50	7.58	351	35.10	7.17	312	31.20	10.57	0.003**	1 > 2* 1 > 3** 2 = 3
Text2	298	29.80	9.07	279	27.90	3.14	222	22.20	8.65	0.078	
Text3	431	43.10	8.56	339	33.90	5.69	332	33.20	9.19	0.016*	1 > 2* 1 > 3* 2 = 3
Text4	366	36.60	7.63	244	24.40	9.73	222	22.20	8.77	0.005**	1 > 2* 1 > 3**

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

and DeepSeek-V3 ($p_{accuracy}$ = 0.004; p_{style} = 0.005) were effective in reducing error scores of accuracy and style types.

In summary, both in terms of error counts and scores, student translators produced higher-quality post-edited translations when assisted by LLMs compared to NMT. In addition, ChatGPT-4o and DeepSeek-V3 demonstrated different strengths in addressing various error types and supporting PE tasks across different text types.

4.3 User perception

User perception, including users' positive and negative attitude or feedback towards technologies, is invaluable for providing insights into tools and workflows and revealing issues that would not be evident from the translations or process data (Bundgaard 2017). After finishing the PE tasks, the participants were asked to complete a post-task questionnaire to evaluate the tools they used.

The brief questionnaire focused on participants' user experiences by evaluating each tool's error-correcting capabilities across the five error categories (completeness, accuracy, language, culture, and style), as well as their reflections on the advantages and disadvantages of LLMs and NMT in supporting PE.

Students first assessed the error-correcting capabilities of each tool across various error types, which demonstrated their clear preferences. According to the results in Figure 2, DeepL, ChatGPT-4o, and DeepSeek-V3 received 39, 56, and 42 votes respectively. The highest number of votes was obtained by ChatGPT-4o, suggesting that student translators generally perceived it as the most effective tool in terms of error-correcting capability during C-E PE practices.

From the perspective of the five error categories, ChatGPT-4o received the highest number of votes in the dimensions of completeness, language, and style, with

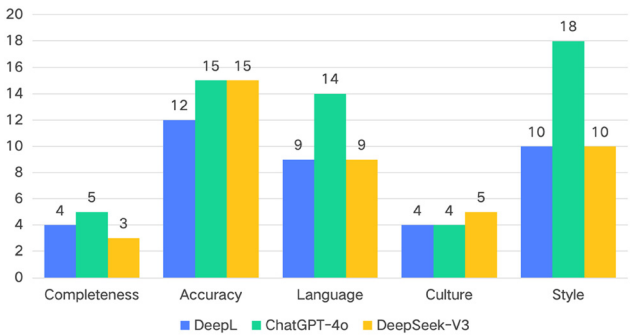

Figure 2: Number of votes for each tool's error-correcting ability.

particularly higher votes in the latter two. This suggests that students perceived ChatGPT-4o as more capable of assisting them in dealing with language and style types of errors during PE. In the dimension of accuracy, both ChatGPT-4o and DeepSeek-V3 received higher votes than DeepL, indicating that LLMs' assistance was helpful for resolving accuracy errors. Last, in the culture dimension, DeepSeek-V3 received slightly more votes than the other two tools, implying that students viewed it as more effective in correcting culture errors.

In addition to voting for the three tools' error-correcting ability in the five error types, students shared qualitative evaluations, mainly focusing on the advantages and disadvantages of each tool. Based on their evaluations, we calculated the theme distribution and frequency, and sorted them into five dimensions including assistance efficiency, information accuracy, user experience, translation quality, and context comprehension (Figure 3) to visually present student translators' evaluation of advantages and disadvantages towards DeepL, ChatGPT-4o and DeepSeek-V3 in assisting PE practices.

From the chart, we found that, in terms of assistance efficiency, ChatGPT-4o demonstrated the most notable advantage, followed by DeepSeek-V3, while DeepL was comparatively less efficient. Some students noted that when using DeepL to assist PE, the process was cumbersome, and the task flow would be easily disrupted when they had to search information online that DeepL couldn't support. In the information accuracy aspect, LLMs did not outperform DeepL, primarily because they offered a wide range of information – some of which may be inaccurate – leaving the translator to make judgments. As for user experience, student translators generally found LLM-assisted PE to be highly convenient, and the

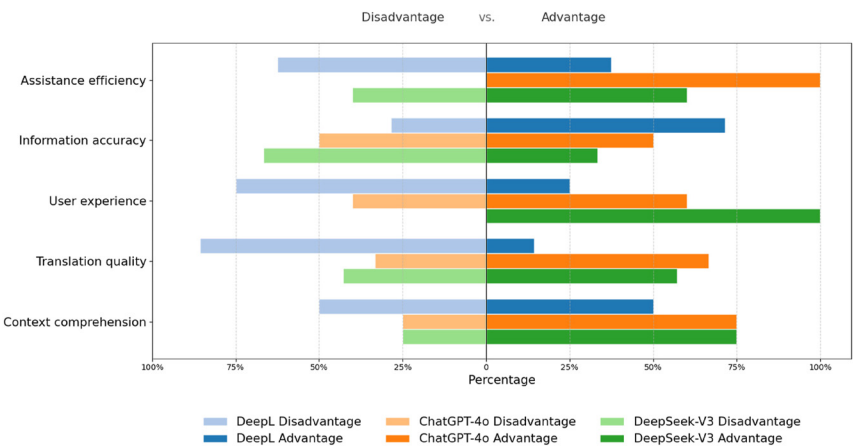

Figure 3: Frequency distribution of students' evaluations of the tools' advantages and disadvantages.

information provided often inspired new ideas and perspectives. In terms of translation quality, DeepL was outperformed by the LLMs, with ChatGPT-4o showing a particularly clear advantage. In terms of context comprehension, DeepL lagged behind the LLMs as it was unable to accurately understand the context of source texts.

Overall, the data from total task duration, keyboard events, and pause length told us that there was no significant difference in three groups of student translators' PE workload – encompassing total temporal, technical, and cognitive efforts. The only exception was observed in the popular science text tasks, where students in G1, assisted by NMT, exhibited a noticeably higher number of keyboard events compared to the other two groups.

In contrast, the data on error counts and error scores showed a clear advantage for LLM-assisted PE. Students in the G2 and G3 produced translations with fewer errors and lower error scores in total than those in the G1. This suggested that LLMs can effectively support students in generating higher-quality translations during PE tasks.

From the perspective of user perception, students rated ChatGPT-4o highest in terms of error correcting capability and assistance efficiency. DeepSeek-V3 was recognized for offering the best user experience, while DeepL was regarded as the most reliable in terms of information accuracy. These findings highlighted the importance of considering both objective performance and subjective user experience when integrating AI tools into PE practices.

5 Discussion

5.1 Guidelines for C-E PE

This experiment selected four different types of Chinese source texts, with their MTs containing various types of errors, including completeness, accuracy, language, culture, and style. These errors posed considerable challenges for student translators in the PE process. Although they produced fewer errors with lower scores in PE tasks assisted by LLMs compared to those aided by NMT, some errors still persisted. By comprehensively analyzing the source texts, MTs, and screen recordings of the students' translation processes, we identified three key aspects in which the assistance role of LLMs should be further leveraged to enhance the PE quality.

First, ensuring semantic accuracy. Machine translation systems often failed to accurately translate expressions containing culture-specific concepts. For example, in Text 3, the cultural term "叼空" is a regional colloquialism meaning "to squeeze time to do something," yet the Youdao translated it simply as "manage," which meant

to succeed in doing something without the implication of time management, thereby omitting the cultural nuance of the source text. Similarly, "风水书" does not refer to an ordinary book, but rather to a type of Chinese almanac developed through centuries of Chinese people's practices to guide the selection of auspicious dates. The reference translation used "fengshui almanac," which accurately conveyed the concept, but if it was translated as "fengshui book" as Youdao may mislead readers into thinking that it refers to a general book about fengshui theory, being inconsistent with the context.

Second, maintaining logical consistency. Machine translation systems also frequently failed to deal with referential relationships. In Text 2, for instance, "梁任公," "梁启超," and "任公" all referred to the same historical figure. However, Youdao inconsistently translated these as "Mr. Liang Rengong," "Liang Qichao," or "Mr. Rengong," thereby disrupting referential coherence. Similarly, in Text 1, "(鲁班锁)完全依靠自身结构的连接支撑" emphasized that the connection and supporting functions depend on the mechanism of Luban lock. However, the machine translation system translated it into "It is supported by the connection of its own structure", which distorted the original logic.

Third, following the language rules. In Text 4, the sentence "所谓纪律习惯, 盖指多人聚集场面, 无待一条一条宣布" was translated as "The so-called discipline and habit refer to the fact that in a scene where many people gather, there is no need to announce one by one." This translation contained two clauses without a clear syntactic connection, violating English cohesion norms. Furthermore, the subject in the second clause was also vague.

5.2 Methods for PE with LLM assistance

Experimental data suggested that LLM-assisted PE achieved higher translation quality than with NMT. Experimental observations and questionnaire results indicated some methods for effective LLMs integration in PE practices.

First, it is essential to fully leverage the respective strengths of different LLMs. Although the experimental results demonstrated that PE assisted by ChatGPT-4o and DeepSeek-V3 led to higher translation quality, certain differences between the two tools remained. In C–E PE tasks, student translators may choose tools based on text types, quality expectations, and personal preferences. For example: (1) When text type is not taken into consideration, both ChatGPT-4o and DeepSeek-V3 can effectively assist in reducing error counts and scores in PE tasks. However, these two tools also exhibit certain differences. For instance, compared to using DeepL, ChatGPT-4o is more effective in helping students reduce language type errors, whereas DeepSeek-V3 performs better in assisting the reduction of error scores in

completeness aspect; (2) For texts requiring higher translation quality – whether in the popular science, prose, or academic work texts – DeepSeek-V3 helped student translators reduce accuracy errors (including mistranslation, repetition, ambiguity, and omission), while ChatGPT-4o contributed more to reducing style errors (including stylistic inconsistency and overly literal translation) in prose and academic work texts. (3) According to students' evaluations of tool performance, ChatGPT-4o was considered the most efficient in terms of assistance efficiency, while DeepSeek-V3 provided the best user experience. Therefore, student translators may also take into account user perception and preference when selecting LLMs for supporting PE.

Second, it is important for translators to proofread the text themselves. In Text 2, the "前辈的学者" in the source text referred to elder scholars as opposed to the scholars of younger generation. However, the machine rendered it as "senior scholars", which typically referred to scholars with higher academic rank or seniority, failing to match the intended contrast with "后生" (the younger generation), thus introducing semantic inaccuracy. Notably, among the students using ChatGPT-4o or DeepSeek-V3, only one student identified and corrected this error. Likewise, the LLMs themselves seldom found or addressed this issue during their interaction with students. In contrast, 80 % of the students who used DeepL successfully corrected the error. This suggested that effective use of LLMs not only involved crafting appropriate prompts but also needed translators to proofread their post-edited work in parallel with the source text. Blind reliance on LLMs' revision suggestions may lead to missed errors in translations.

5.3 Implications for PE education

Translation students generally acknowledged the effectiveness of LLMs in PE, indicating the necessity of incorporating AI technologies into translation teaching to enhance student PE competence. Previous research defined PE competence as the knowledge and cognitive literacy required to revise machine-generated outputs based on task-specific objectives, encompassing MT knowledge, terminology management, discourse knowledge, documentation skills, intercultural awareness, error identification, and editing efficacy (Feng and Liu 2018; Koponen 2015; O'Brien 2002). Our findings further emphasized the importance of error correcting and critical thinking skills for translator training in the LLM era.

From the screen recordings of students' translation processes, we observed that among the multiple translation suggestions provided by LLMs, at least one or more often contained relatively inaccurate expressions. For example, in Text 3, the phrase "鬼使神差" was rendered in the machine translation as "by some forces", which was

overly simplistic and resulted in cultural loss. The translation failed to capture the literary and stylistic information of the original expression. During student–LLM interaction, DeepSeek-V3 proposed more culturally sensitive or stylistically appropriate expression suggestions, such as "As if guided by ghosts and gods", which retained cultural imagery, or the idiomatic English phrase "It is a twist of fate". However, some students ultimately adopted a less appropriate revision like "driven by some inexplicable reasons", which may obscure the intended meaning and diminish the stylistic effect.

At the same time, LLMs may also provide misleading suggestions. In Text 3, ChatGPT-4o advised the student to revise the translation of "无所为而为" as "a kind of effortless action – what the Daoists would call 'doing without striving'". This reinterpretation deviated significantly from the original meaning, which in context referred to Liang Qichao's pursuit of learning driven by internal motivation rather than Daoist philosophy. Nevertheless, the student adopted the suggestion without critical thinking. Similarly, in Text 1, DeepSeek-V3 recommended translating "间不容发" as "fail-proof interlocking." Yet, the term "fail-proof" typically denoted resistance to failure, which distorted the original text. These cases highlighted the risk of over-relying on LLM outputs without sufficient error correcting and critical thinking abilities.

Furthermore, some students expressed in the post-experiment questionnaire that, while neural NMT lacked the richness and immediacy of interaction that LLMs can offer, they appreciated the greater cognitive space provided for independent thinking. This observation underscores the value of pedagogical designs that balance AI assistance with opportunities for autonomous decision-making in PE tasks, and translation instruction should critically evaluate the role of AI, utilizing it as an auxiliary tool while maintaining human agency and critical awareness.

Moreover, prompt quality directly affects the quality of LLM-generated contents. Understanding prompt variation enhances ChatGPT interaction (Ekin 2023), and effective prompt design hinges on comprehension of LLM mechanisms (Polverini and Gregorcic 2024). Translation teaching should therefore incorporate additional knowledge of other areas, such as prompt engineering, enabling students to refine their "prompting intelligence" through PE practice.

6 Conclusions

This study investigated student translators' workload, translation quality, and user perception when performing C-E PE practices assisted by NMT and LLMs. The results indicated: (1) In workload, student translators' total task duration, keyboard events, and pause lengths during the PE process were not affected by the assistance tools

they chose, except in the popular science text task, students assisted by NMT had more keyboard events than those assisted by LLMs. (2) In translation quality, student translators assisted by LLMs produced higher-quality translation than those assisted by NMT with less error counts and lower error scores. (3) In user perception, student translators regarded ChatGPT-4o as the most powerful error correcting tool with the highest assistance efficiency, appreciated DeepSeek-V3 for its good user experience, and DeepL for its accurate information.

This study suggests that LLMs can assist student translators in C-E PE, particularly in improving translation quality by helping students with reducing error counts and error scores. The experiment also offered insights for clarifying error correcting abilities and advantages and disadvantages of different NMT and LLMs in supporting PE, provided guidelines for C-E PE, methods for PE with LLMs' assistance, and implications for PE pedagogy in the AI era.

Nonetheless, this study has certain limitations. First, this experiment only selected C-E texts as PE materials and participants from a single institution, and future studies could incorporate texts of multiple language pairs and diverse levels of difficulty, and recruit participants from different institutions to enhance the representativeness of experimental results. Second, as translation process data, including task duration, keyboard events and pause length, shows almost no differences, further research on the process of LLM-assisted PE needs to explore more factors to comprehensively evaluate student translators' performance within this translation mode. Third, with adequate equipment foundations, tools such as eye-tracking devices could also be used in the future to further explore human-AI collaborative PE.

Note: The Test for English Majors-Band 8 (TEM-8) is based on the highest level of standard for English major students in China and is taken in the eighth term. TEM-8 comprehensively evaluates students' English ability in listening, reading, writing and translating abilities. The China Accreditation Test for Translators and Interpreters (CATTI) is a state-level vocational qualification examination for translators and interpreters to demonstrate that they show certain aptitudes required by the industry.

Research funding: This paper is supported by the 14th Five-Year Plan of Education Sciences in Hunan Province (Project No.: ND228199; Project Approval No.: XJK22BGD012) and the Scientific Innovation Fund for Post-graduates of Central South University of Forestry and Technology (Project No.: 2024CX02102).

Appendix A: Examples from the source texts and their references and MT outputs

Text	Source text	Reference	Machine translation
1	鲁班锁由六根具有凹凸构造的短木组成, 短木之间通过榫卯工艺相互咬合连接,不使用任何铁钉或绳索,完全依靠自身结构的连接支撑	The Lu Ban Lock consisted of six short battens with concave-convex construction, occlusion connection between short battens was made via mortise and tenon joint craft, connection and supporting was made totally depending on its structure.	The Lu Ban lock is composed of six short pieces of wood with a concave-convex structure. The short pieces are interlocked and connected through the mortise and tenon joints. It is entirely supported by the connection of its own structure.
	(鲁班锁)具有结构巧妙、扣合严密、间不容发、易拆难装的特点	It had features of ingenious structure, precise fastening, extremely small gap, and ease of disassembly but difficulty in assembly.	It features a clever structure, tight interlocking, seamless interlocking, and is easy to disassemble but difficult to assemble.
2	前辈的学者常以学问的趣味启迪后生, 因为他们自己实在是得到了学问的趣味,故不惜现身说法,诱导后学。	Scholars of the older generation often urge young people to develop interest in learning because they themselves have been enjoying the real pleasure of academic studies. And they are ever ready to cite their own example by way of advice.	Since they themselves have truly gained the interest of learning, they are willing to share their own experiences to guide the younger generation.
	梁任公先生就说过: “我是个主张趣味主义的人, 倘若用化学化分 ‘梁启超’ 这件东西,把里头所含一种原素名叫 ‘趣味’ 的抽出来, 只怕所剩下的仅有个零了。”	The distinguished scholar Liang Qichao once said wittily, “‘I always stand for interest-ism. If you broke down Liang Qichao’s stuff into its component parts, there would be nothing left except an element named ‘Interest’.”	Mr. Liang Rengong once said, “I am a person who advocates the doctrine of taste. If we were to chemically decompose the substance ‘Liang Qichao’ and extract an element called ‘taste’ from it, I’m afraid there would be only zero left.”
3	这几天给话剧团装台, 忙得两头儿不见天, 但顺子还是叼空, 把第三个老婆娶回来了。	Smooth Diao had been so busy lately assembling the stage for the modern drama troupe that he didn’t catch the sunlight at either end of the day. Still, he did manage to squeeze in enough time to see that his new wife - the third - was fetched home.	These days, he was busy setting up the stage for the drama troupe and was so busy that he couldn’t see the sky at all. But Shunzi still managed to get his third wife back.
	可神使鬼差的, 好像不娶都不行了, 他也就自己从风水书上翻看了日子, 没	The ghosts seemed to be piloting his course, and so he consulted the feng shui almanac to divine	But by some strange force, it seemed that he had no choice but to marry her. So he looked up the date in a feng shui book

(continued)

Text	Source text	Reference	Machine translation
	带一个人, 打辆出租车, 就去把人接回来了。	an auspicious date, before hiring a taxi to shuttle her over.	by himself, didn't bring anyone along, took a taxi and went to pick her up.
4	所谓纪律习惯, 盖指多人聚集场面, 无待一条一条宣布, 而群众早已习惯成自然的纪律。	This is the type of discipline that need not be declared on occasions when many people gather together, and which the populace has observed so long that it is second nature.	The so-called discipline and habit refer to the fact that in a scene where many people gather, there is no need to announce one by one, but the masses have already become accustomed to the natural discipline.
	无论消极积极, 扼要一句话: 必求集体行动起来, 敏捷顺利, 效率要高不因人多而牵扰费时。	Whether a practice is prohibited or encouraged, one sentence well sums up the key point: to make things smooth and effective without causing disturbance or consuming time, collective actions are required.	Whether positive or negative, in a nutshell: One must act collectively, nimbly and smoothly, with high efficiency and not be disturbed or time-consuming by a large number of people.

Appendix B: Human evaluation rules

Error type		Description	Example
Completeness	Omission	The translation fails to include information from the source text.	ST: 他忙得两头儿不见天。 TT: He was so busy that he couldn't see the sky at all. (The phrase "两头儿" is omitted in the translation.)
Accuracy	Mistranslation	The translation misinterprets information of the source text.	ST: 风水书。 TT: A fengshui book. (Here, "书" refers to a calendar but not a book.)
	Ambiguity	The source text is clear, but the translation introduces ambiguity.	ST: 顺子把第三个老婆娶回来了。 TT: Shunzi got his third wife back. ("Got back" may misleadingly suggest he retrieved a lost wife, rather than married a new one.)
	Redundancy	The translation includes unnecessary repetition.	ST: 扣合严密、间不容发。 TT: Tight interlocking and seamless interlocking. (The repeated use of "interlocking" is redundant; it could be

(continued)

Error type		Description	Example
			simplified to “tight and seamless interlocking.”)
	Addition	The translation adds information not presented in the source text.	ST: 前辈的学者常以学问的趣味启迪后生。 TT: Senior scholars, like gentle gardeners, often inspire the younger generation with the interest of learning. (“Like gentle gardeners” is an added metaphor not found in the source.)
Language	Grammar error		ST: 所谓纪律习惯, 盖指多人聚集场面, 无待一条一条宣布。 TT: The so-called discipline and habit refer to the fact that in a scene where many people gather, there is no need to announce one by one. (The translation contains two clauses without a clear syntactic connection.)
	Spelling error		A student misspelled “crash” as “crashh.”
	Incorrect word choice		ST: 秋菊盆景。 TT: autumn chrysanthemum bonsai (“Bonsai” refers to a small tree that is grown in a pot, which is not suitable for chrysanthemum.)
Culture	Cultural discrepancy	The cultural information in the translation is inconsistent with Chinese culture.	ST: 无所为而为。 TT: He merely acted without doing anything. (The translation fails to convey the cultural and philosophical connotations of the source text.)
Style	Inconsistent style	The translation’s style does not match the source text.	ST: 扼要一句话。 TT: In a nutshell. (The source text is from an academic work with rigorous logic, while the translation is informal and idiomatic.)
	Overly literal translation	The translation is stiff and shows signs of word-for-word rendering.	ST: 走进学问的大门。 TT: Enter the door of learning. (The literal translation misses the cultural and metaphorical richness of the source text.)

References

Adawiyah, Azza Rabiatul, Lalu Ali Wardana Baharuddin & Santi Farmasari. 2023. Comparing post-editing translations by Google NMT and Yandex NMT. *TEKNOSASTIK* 21(1). 23–34.

Bhattacharyya, Pushpak, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri & Turchi Marco. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, 672–681. Pennsylvania: Association for Computational Linguistics.

Bowker, Lynne & Jairo Buitrago Ciro. 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Bingley: Emerald Publishing.

Bundgaard, Kristine. 2017. Translator attitudes towards translator-computer interaction – Findings from a workplace study. *Hermes-Journal of Language and Communication in Business* 56. 125–144.

Carl, Michael, Schaeffer Moritz & Bangalore Srinivas. 2016. The CRITT translation process research database. In Michael Carl, Srinivas Bangalore & Moritz Schaeffer (eds.), *New directions in empirical translation process research*, 13–54. Cham: Springer.

Chen, Yan. 2019. 装台 [*The backstage clan*]. Beijing: People's Literature Publishing House.

Chen, Yan. [2019] 2023. *The backstage clan*. Translated by Hu, Zongfeng & Robin Gilbank. West Sussex: ACA Publishing Limited.

Da, Jeff, Ronan Le Bras, Ximing Lu, Yejin Choi & Bosselut Antoine. 2021. Analyzing commonsense emergence in few-shot knowledge models. https://doi.org/10.48550/arXiv.2101.00297.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker & Lieve Macken. 2017a. Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta* 62(2). 245–270.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker & Lieve Macken. 2017b. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology* 8. 1–15.

Dai, Guangrong & Siqi Liu. 2024. Towards predicting post-editing effort with source text readability: An investigation for English-Chinese machine translation. *The Journal of Specialised Translation* 41. 206–229.

Dai, Ling, Xiaowei Zhao & Zhiting Zhu. 2023. 智慧问学:基于ChatGPT的对话式学习新模式 [A new inquiry learning: Conversational learning with ChatGPT]. *Open Education Research* 29(6). 42–51, 111.

Ekin, Sabit. 2023. Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. https://doi.org/10.36227/techrxiv.22683919.v29.

Fan, Zirui & Wendi Yang. 2024. 人机耦合时代机器翻译译后编辑原则与策略例析 [Principles and strategies of post-editing in the human-machine coupling era]. *Shanghai Journal of Translators* (4). 29–34.

Farghal, Mohammed & Ahmad S. Haider. 2024. Translating classical Arabic verse: Human translation vs. AI large language models (Gemini and ChatGPT). *Cogent Social Sciences* 10(1). 1–15.

Feng, Quangong & Qiliang Cui. 2016. 译后编辑研究:焦点透析与发展趋势 [Research focuses and trends in post-editing of machine translation]. *Shanghai Journal of Translators* (6). 67–74.

Feng, Quangong & Ming Liu. 2018. 译后编辑能力三维模型构建 [Towards the construction of a three-dimension model of post-editing competence]. *Foreign Language World* (3). 55–61.

Gao, Ruiyao, Yumeng Lin, Nan Zhao & Zhenguang G. Cai. 2024. Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications* 11(1). 1–10.

Geng, Xiaolong. 2024. Optimizing post-editing strategies in human-computer interaction: An empirical investigation of efficiency and cognitive load. *International Journal of Information and Communication Technology Education* 20(1). 1–15.

Geng, Fang & Jian Hu. 2023. 人工智能辅助译后编辑新方向——基于ChatGPT的翻译实例研究 [New direction for post-editing by artificial intelligence translation: A case study of ChatGPT translation]. *Foreign Languages in China* 20(3). 41–47.

Han, Ziman & Tongda Chai. 2024. 人工智能知识翻译能力探析——以文学掌故为例 [Knowledge translation competence of artificial intelligence: A case study of literary anecdotes]. *Technology Enhanced Foreign Language Education* (5). 3–10.

Heinzerling, Benjamin & Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Pennsylvania: Association for Computational Linguistics.

Hendy, Amr, Abdelrehim Mohamed, Amr Sharaf, Vikas Raunak, Gabr Mohamed, Hitokazu Matsushita, Jin Kim Young, Afify Mohamed & Hassan Awadalla Hany. 2023. How good are GPT models at machine translation? A comprehensive evaluation. Available at: https://arxiv.org/abs/2302.09210.

Hu, Kaibao & Juan Li. 2024. 大语言模型背景下的翻译人才培养:挑战与前景 [Cultivating translation talents in the era of large language models: Challenges and prospects]. *Technology Enhanced Foreign Language Education* (6). 3–7.

Huang, Youyi. 2022. 从"翻译世界"到"翻译中国" [*From "Translating the World" into "Translating China*"]. Beijing: Foreign Languages Press.

ISO. 2014. *Translation services – Post-editing of machine translation output – Requirements* (ISO 18587: 2014).

Jia, Yanfang & Sanjun Sun. 2022. Man or machine? Comparing the difficulty of human translation versus neural machine translation post-editing. *Perspectives* 15(5). 950–968.

Jia, Yanfang & Binghan Zheng. 2022. The interaction effect between source text complexity and machine translation quality on the task difficulty of NMT post-editing from English to Chinese: A multi-method study. *Across Languages and Cultures* 23(1). 36–55.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi & Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. https://doi.org/10.48550/arXiv.2301.08745.

Kauf, Carina, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko & Alessandro Lenci. 2023. Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science* 47(7). 1–20.

Khasawneh, Yusra Jadallah Abed & Mohamad Ahmad Saleem Khasawneh. 2023. The use of artificial intelligence in improving machine translation post-editing: Insights from translation editors. *Journal of Namibian Studies: History Politics Culture* 34. 7123–7146.

Koponen, Maarit. 2015. How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of the 4th Workshop on Post-editing Technology and Practice*. Miami, USA: Association for Machine Translation in the Americas.

Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Ohio: Kent State University Press.

Kumpulainen, Minna. 2015. On the operationalisation of 'pauses' in translation process research. Translation & interpreting. *The International Journal of Translation and Interpreting Research* 7(1). 47–58.

Li, Fengxi. 2022. 人工智能时代人机英汉翻译质量对比研究 [A comparative study on the quality of English-Chinese translation between translation learners and a machine translation system in the era of artificial intelligence]. *Foreign Language World* (4). 72–79.

Li, Menglu & Dechao Li. 2025. Human expertise vs AI efficiency: A comparative analysis of student and ChatGPT post-editing. In Sanjun Sun, Kanglong Liu & Riccardo Moratto (eds.), *Translation studies in the age of artificial intelligence*, 150–171. London: Routledge.

Liang, Shuming. 2018. 中国文化要义 [*Fundamentals of Chinese culture*]. Shanghai: Shanghai People's Publishing House.

Liang, Shuming. [2018] 2021. *Fundamentals of Chinese culture*. Translated by Li, Ming. Amsterdam: Amsterdam University Press.

Liao, Hailong. 2025. DeepSeek large-scale model: Technical analysis and development prospect. *Journal of Computer Science and Electrical Engineering* 7(1). 33–37.

Lu, Daokun & Jiyu Chen. 2024. Sora:学校教育的"拯救者"还是"终结者" [Sora: "Rescuing" or "terminating" school education]. *Journal of Xinjiang Normal University* 45(6). 112–127.

Lu, Zhi & Juan Sun. 2018. 人工翻译和译后编辑中认知加工的眼动实验研究 [An eye-tracking study of cognitive processing in human translation and post-editing]. *Foreign Language Teaching and Research* 50(5). 760–769, 801.

Nitzke, Jean & Anne-Kathrin Gros. 2020. Preferential changes in revision and post-editing. In Maarit Koponen, Brian Mossop, Isabelle S. Robert & Giovanna Scocchera (eds.), *Translation revision and post-editing: Industry practices and cognitive processes*, 21–34. London: Routledge.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*. Manchester, England: European Association for Machine Translation.

Polverini, Giulia & Bor Gregorcic. 2024. How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics* 45(2). 1–36.

Pym, Anthony. 2012. Translation skill sets in a machine-translation age. *Meta* 58(3). 487–503.

Raiaan, Mohaimenul Azam Khan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, J. Ahmad, M. E. Ali & S. Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access* 12. 26839–26874.

Safavi, Tara & Danai Koutra. 2021. Relational world knowledge representation in contextual language models: A review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Pennsylvania: Association for Computational Linguistics.

Secară, Alina. 2005. Translation evaluation: A state of the art survey. In *Proceedings of the eCoLoRe/ MeLLANGE Workshop*. Leeds: Centre for Translation Studies, University of Leeds.

Shanahan, Murray. 2024. Talking about large language models. *Communications of the ACM* 67(2). 68–79.

Shin, Dongkawang & Yuah V. Chon. 2023. Second language learners' post-editing strategies for machine translation errors. *Language, Learning and Technology* 27(1). 1–25.

Wang, Shaoshuang. 2024. 技术赋能视域下翻译能力体系的反思与重构——现代译者的"知——思—行"翻译能力模型解析 [Rethinking and reconstructing translation competence framework from the perspective of technological empowerment: An explanation of the knowledge-thinking-action competence model for modern translators]. *English Studies* 21. 52–64.

Wang, Yupei & Renfen Hu. 2021. A prompt-independent and interpretable automated essay scoring method for Chinese second language writing. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Beijing: Chinese Information Processing Society of China.

Wang, Lyu & Xiangling Wang. 2023. ChatGPT时代机器翻译译后编辑能力培养模式研究 [The study of training model of machine translation post-editing competence in the era of ChatGPT]. *Technology Enhanced Foreign Language Education* (4). 16–23+115.

Wang, Jiayi & Liyang Wang. 2024. 机器翻译译后编辑认知努力研究进展 [Research progress of cognitive effort of post-editing in machine translation]. *Journal of Xi'an International Studies University* 32(2). 81–86.

Wang, Huashu & Chengzhi Zhang. 2025. GenAI时代的翻译实践模式: 技术迭代、业态变革与趋势展望 [Towards reshaping translation practice models in the GenAI era: Technological iteration, industry transformation, and development trends]. *Foreign Language Education* 46(1). 53–58.
Wang, Junsong, Weiqing Xiao & Qiliang Cui. 2023. 人工智能时代技术驱动的翻译模式:嬗变、动因及启示 [Technology-driven translation modes in the AI age: Evolution, causes and implications]. *Shanghai Journal of Translators* (4). 14–19.
Wang, Xiangling, Xiaoye Li & Guangjiao Chen. 2024. 人工译文修改与机器翻译译后编辑的对比研究——来自键盘记录、反省法与调查问卷的证据 [Comparing translation revision and machine post-editing: Evidence from keylogging, retrospection and questionnaire]. *Foreign Language Learning Theory and Practice* (5). 88–97.
Washbourne, Kelly. 2014. Beyond error marking: Written corrective feedback for a dialogic pedagogy in translator training. *The Interpreter and Translator Trainer* 8(2). 240–56.
Yang, Zhihong. 2012. 翻译质量量化评估: 模式、趋势与启示 [Quantitative assessment of translation quality: Models, trends and implication]. *Foreign Languages Research* (6). 65–69, 112.
Zhang, Peiji. 2007. 英译中国现代散文选(一) *[Selected modern Chinese essays: English translation, Volume 1]*. Shanghai: Shanghai Foreign Language Education Press.
Zhang, Wenyu & Bi Zhao. 2024. 生成式人工智能开创机器翻译的新纪元了吗? 一项质量对比研究及对翻译教育的思考 [Has generative AI opened a new era for machine translation? – A contrastive quality study and reflections on translation education]. *Journal of Beijing International Studies University* 46(1). 83–98.
Zhong, Wenming & Chao Shu. 2020. 译后编辑的能力结构与课程设置——基于国外译后编辑课程的前沿分析 [Competence structure and course design of post-editing: A frontier analysis based on foreign post-editing courses]. *Technology Enhanced Foreign Language Education* (6). 86–91.
Zhong, Wenming, Di Wang & Tian Sha. 2024. 生成式人工智能与神经网络机器翻译人工译后编辑效率对比研究 [An investigation into the post-editing efficiency between AI-generated content and neural machine translation]. *Translation Research and Teaching* (2). 96–105.
Zhu, Yue'e. & Xiong Shao. 2024. ChatGPT反馈辅助科技翻译质量提升的量化研究 [Quantitative research on improving the quality of sci-tech translation assisted by ChatGPT feedback]. *Language and Intelligence* (1). 87–106.
Zhu, Yue'e. & Xiong Shao. 2025. 交互式翻译实践中ChatGPT反馈对学生科技翻译质量的影响 [Impacts of ChatGPT feedback on students' sci-tech translation quality in interactive translation practice]. *Translation Horizons* (1). 98–115.

Bionotes

Xiong Shao
College of Foreign Languages, Central South University of Forestry and Technology, Changsha, China
1300913542@qq.com
https://orcid.org/0009-0001-9446-6295

Xiong Shao is a postgraduate student in the MTI program at the College of Foreign Languages, Central South University of Forestry and Technology. His main research interests include computer-assisted language learning and machine translation.

Yue'e Zhu
College of Foreign Languages, Central South University of Forestry and Technology, Changsha, China
445447701@qq.com

Yue'e Zhu is a professor at the College of Foreign Languages, Central South University of Forestry and Technology. Her main research interests include language education and translation theories.