Journal of Literature & Language, 2026, 2(1), 21-34
https://10.71204/d7beya25

AI-assisted Post-editing and the Development of Critical Technology Awareness:A Study of Chinese-Japanese Political Text Translation Pedagogy

Qiushi Gu [1, *], Shiyan Wang [2], Xinyu Ji [1], Xinyao Ren [1]

[1] School of Japanese Studies, Beijing International Studies University, Beijing 100024, China

[2] Japanese Department, University of International Relations, Beijing, 100091 China

*** Correspondence:**

Qiushi Gu

guqiushi@bisu.edu.cn

Received: 9 March 2026/ Accepted: 19 April 2026/ Published online: 21 May 2026

Abstract

Generative artificial intelligence and neural machine translation are changing the conditions under which translation is taught and practiced. For graduate students in translation, the challenge is no longer simply how to use machine translation and large language models, but how to identify, explain, and revise semantic shifts, terminological inconsistency, and register problems in AI-generated output. This study reports a one-group pre-test/post-test quasi-experiment conducted with 15 first-year MTI students in a Japanese translation program at a Chinese university. The instructional intervention focused on Chinese-Japanese post-editing of political texts. Three recent Chinese speeches from United Nations Security Council contexts were used as parallel materials: one for the pre-test, one for the post-test, and one for classroom training. The study drew on anonymized three-rater scoring, MQM-informed error annotation, an AI critical literacy questionnaire, post-editing logs, and student reflections. Wilcoxon signed-rank tests showed significant gains in overall post-editing quality, machine-translation error identification, and AI critical literacy. The strongest improvements appeared in revision rationale and evidence use, terminology and proper-name handling, and political-diplomatic register. Student logs and interview comments further suggest a shift from accepting fluent AI output to checking, justifying, and critically revising it. The paper argues that post-editing pedagogy in the AI era should move beyond tool operation and be designed as a competence loop linking technology use, error diagnosis, discourse reconstruction, and critical reflection.

Keywords: AI-Assisted Translation; Machine Translation Post-Editing; Translation Pedagogy; Political Texts; AI Critical Literacy

1. Introduction

The growing availability of machine translation systems and large language models has changed a basic premise of translator education. In graduate-level translation courses, it is no longer sufficient to ask whether students can use a tool. A more pressing question is whether they can judge when an AI-generated translation is reliable, where it is risky, and how a human translator should intervene responsibly. This question is especially important in the translation of political texts. Such texts are dense with proper names, institutional titles, resolution numbers, modality, stance markers, and formulaic diplomatic wording. A successful translation must therefore be accurate, terminologically consistent, stylistically appropriate, and discourse-sensitive.

AI systems can quickly produce readable target texts, but readability may conceal problems that matter in political communication. Errors may occur in the rendering of names, institutional terms, modal strength, diplomatic politeness, stance attribution, or culturally embedded assumptions. In Chinese-Japanese translation, even seemingly small shifts in register or modality can change the force of a political statement. For this reason, political texts provide a productive site for training students to evaluate AI output rather than merely polish it.

This study uses post-editing as the point of entry and addresses three questions in the teaching of Chinese-Japanese political text translation. First, can students identify politically consequential errors that may be hidden by the surface fluency of AI-generated Japanese? Second, can systematic post-editing instruction improve both translation quality and error-diagnosis ability? Third, how can teachers assess students' critical technology awareness and cross-cultural sensitivity through a workable classroom rubric?

The study was guided by the following research questions. RQ1: Does AI-assisted post-editing instruction improve graduate students' Chinese-Japanese post-editing quality in political text translation? RQ2: Does the intervention improve students' ability to identify errors in machine-translated output, particularly in terminology, proper names, political-diplomatic register, and AI-related critical reflection? RQ3: Are changes in AI critical literacy and error-identification ability associated with gains in post-editing quality?

2. Literature Review

2.1. Post-editing and Translator Competence

ISO 18587:2017 defines post-editing as human editing of machine translation output and stresses that full post-editing requires linguistic, cultural, domain, research, technical, and post-editing competences. Koponen (2016) notes that the value of post-editing depends on machine translation quality, text type, and editing effort; when the initial output is weak, post-editing can increase rather than reduce cognitive burden. O'Brien (2011) also shows that automatic quality indicators may be related to post-editing effort, but they cannot replace an understanding of the human editing process. In a teaching context, then, post-editing should not be reduced to making

a machine translation sound smoother. It should be used to cultivate error awareness, evidence awareness, and technological judgment.

Models of translation competence offer a further basis for this view. The PACTE Group (2003) emphasizes the interaction among bilingual, extra-linguistic, instrumental, strategic, and psycho-physiological sub-competences. The European Master's in Translation competence framework (European Commission, 2022) likewise treats technological competence, service provision, and intercultural language competence as integral to translator education. In the AI era, post-editing pedagogy extends these models into a setting where students must make decisions among tool output, textual norms, audience expectations, and professional responsibility.

2.2. AI, Translation Quality Assessment, and Critical AI Literacy

Large language models have opened new possibilities for translation and translation quality assessment. Jiao et al. (2023) report that ChatGPT, especially when powered by GPT-4, can perform competitively in several high-resource translation settings, while limitations remain in robustness, domain-specific translation, and language pairs with greater linguistic distance. Kocmi and Federmann (2023) propose GPT-based systems as translation quality evaluators and report strong system-level correlations with MQM human labels. Yet AI-based evaluation still raises concerns about transparency, reproducibility, and reliability at fine-grained levels. In classroom settings, AI-generated feedback may be useful as formative support, but it should not replace teacher assessment or student reflection.

The Multidimensional Quality Metrics framework (MQM; Lommel, Burchardt, & Uszkoreit, 2014) provides a flexible way to describe translation quality problems by error type. Unlike a single holistic score, MQM encourages evaluators to distinguish among accuracy, terminology, fluency, style, locale convention, and other issue categories. For post-editing pedagogy, this is valuable because it turns students' intuitive judgments into an explicit diagnostic process: students must say not only whether a translation is good or bad, but where the problem lies, what kind of problem it is, and why revision is needed. The present study does not adopt the entire MQM taxonomy. Instead, it uses MQM as a heuristic framework and adapts it to the needs of Chinese-Japanese political translation, focusing on semantic accuracy, terminology and proper names, Japanese fluency, and political-diplomatic register.

Research on AI literacy also suggests that translation teaching needs a critical dimension. UNESCO’s AI competency framework for students (Miao, Shiohira, & Lao, 2024) identifies human-centered AI use, AI ethics, AI techniques and applications, and AI system design as key areas of competence. Veldhuis et al. (2025) further argue that critical AI literacy requires learners to question the social, political, cultural, and ethical implications of AI systems. Post-editing can therefore serve not only as translation practice, but also as a concrete site for cultivating critical AI literacy.

2.3. Political Text Translation as a Pedagogical Site

Political discourse is institutional, intertextual, and culturally situated. Schäffner (2004) argues that political discourse analysis and translation studies can be brought together through questions of political effect, intercultural transfer, textual function, and translation process. Chinese-

Japanese political translation involves diplomatic wording, terminology from international organizations, references to resolutions, state positions, and levels of register. Students must therefore judge whether the Japanese text conveys the institutional context and force of the Chinese source.

For example, Chinese expressions such as yingdang (应当, “should”), bixu zhichu (必须指出, “it must be pointed out”), duncu (敦促, “urge”), and yanfang (严防, “strictly prevent”) cannot be mapped mechanically onto everyday Japanese. These choices involve not only language but also register, institutional voice, and stance. Political texts are thus well suited to training students to evaluate AI-generated translations critically.

Previous studies provide important foundations in post-editing competence, machine translation quality assessment, and AI literacy. However, two gaps remain. First, much of the discussion focuses on tool use or post-editing efficiency, while less attention has been paid to how students identify and correct register shifts in AI-rendered political texts. Second, teaching reform often emphasizes AI application without integrating error diagnosis, evidence-based editing, and translator agency into a single evaluation framework. This study addresses these gaps through a classroom-based quasi-experiment using Chinese-Japanese political texts.

3. Research Design

3.1. Overall Design

The study adopted a one-group pre-test/post-test quasi-experimental design. This design was chosen for three reasons. First, the participants were enrolled in the same graduate course, and random assignment would have been impractical and potentially unfair. Second, the course itself was designed for competence development, making a pre-test/post-test structure appropriate for the teaching context. Third, anonymized multi-rater scoring, process data, questionnaire data, and effect-size reporting can strengthen the credibility of a small-scale teaching study even without a control group.

3.2. Participants

The participants were 15 first-year MTI students majoring in Japanese translation at a Chinese university. To be included in the analysis, students had to be enrolled in the MTI program, have a foundation in Chinese-Japanese translation, be able to complete Chinese-Japanese political text translation or post-editing tasks independently, and agree that anonymized coursework could be used for teaching research. Students who did not complete either the pre-test or the post-test, did not submit editing logs, or did not consent to the use of anonymized data were excluded. All 15 students had studied Chinese-Japanese translation and had prior exposure to CAT tools such as Trados and to MTPE tasks. Thirteen had passed the Japanese-Language Proficiency Test N1. The group therefore had a solid Japanese-language foundation and sufficient initial experience with translation technology for an AI-assisted post-editing experiment.

3.3. Materials

The materials were drawn from three Chinese speeches delivered in United Nations Security Council contexts in 2025. The texts were excerpts from speeches on practicing multilateralism and improving global governance (9861st meeting), artificial intelligence and international peace and security (10005th meeting), and the future of the United Nations (10024th meeting). Each text contained between 2,500 and 2,700 Chinese characters. The three texts were selected to be broadly comparable in length, paragraph structure, terminological density, and register difficulty. All displayed features typical of political-diplomatic discourse, including international-organization terminology, proper names, stance expressions, modality, and long sentences.

The three texts were assigned to three sets: Text A for the pre-test, Text B for the post-test, and Text C for classroom training. Each text included proper names, institutional terms, modality, causal or stance-marking expressions, and long-sentence structures. To control the source of machine translation output, the researcher generated all Japanese drafts with the same AI system, model version, and prompt before the experiment. The prompt was: 'Please translate the following Chinese political-diplomatic text into natural, formal Japanese that conforms to the style of Japanese political news and diplomatic statements. Please keep proper names, institutional names, resolution numbers, and political positions accurate, and do not add information not found in the source text.' All AI-generated Japanese drafts were used as the initial texts for student post-editing. Questionnaires, logs, and post-edited texts were anonymized with stable student codes.

Table 1. Experimental Materials and Data Collection Tools

Material	Function	Size/Description
Pre-test Text A	Chinese political text translated into Japanese by the same AI system; students post-edited the output.	2,530 Chinese characters
Post-test Text B	Parallel political text; students completed the same post-editing procedure.	2,600 Chinese characters
Training Text C	Used for classroom explanation, error annotation, group discussion, and teacher demonstration.	2,680 Chinese characters
Questionnaires and logs	AI-use background questionnaire, AI critical literacy questionnaire, and post-editing reflections.	AI-use questionnaire before the course; 12-item AI critical literacy questionnaire; logs after each task.

3.4. Instructional Intervention

The intervention was implemented in the middle part of the course and lasted approximately four to six weeks. It followed a sequence of conceptual input, error diagnosis, editing practice, and reflective transfer. AI tool use, MQM-informed error annotation, political-diplomatic register analysis, and teacher feedback were integrated throughout the module. The goal was to prevent post-editing instruction from becoming a purely technical exercise and to guide students toward error identification, evidence retrieval, and critical revision.

Table 2. Instructional Intervention

Week	Focus	Classroom Activities	After-class Task
Week 1	Post-editing and ISO 18587	Explaining full and light post-editing, post-editing goals, and risks in AI output.	Complete pre-test Text A and AI-use background questionnaire.
Week 2	Terminology and proper names in political texts	Building a terminology list.	Revise the terminology list and provide sources.
Week 3	MQM-informed error annotation	Annotating semantic, terminological, fluency, and register errors in Text C.	Submit an error annotation sheet and revision reasons.
Week 4	Critical reading of AI output and prompt adjustment	Comparing machine translation, student versions, and teacher references; discussing when AI output can or cannot be trusted.	Complete a post-editing task with an editing log.
Week 5	Reconstructing political-diplomatic register	Group revision of modality, stance boundaries, sentence compression, and Japanese norms.	Submit a second draft and self-evaluation.
Week 6	Post-test and reflection	Complete post-test Text B and AI critical literacy questionnaire.	Submit learning reflections and interview prompts.

3.5. Instruments

The post-editing quality rubric was developed from three sources. First, ISO 18587:2017 was used to incorporate accuracy, terminology, target-language expression, and post-editor responsibility. Second, MQM categories such as accuracy, terminology, fluency, and style were adapted to describe common problems in machine-translated output. Third, the teaching goals of Chinese-Japanese political translation led to the addition of political-diplomatic register, revision rationale and evidence, and critical reflection on AI.

The rubric had a total score of 100 and consisted of six dimensions: semantic accuracy, terminology and proper names, Japanese fluency, political-diplomatic register, revision rationale and evidence, and critical reflection on AI. The first three dimensions measured general translation quality, while the last three captured requirements specific to political text post-editing and AI-era translator competence.

The second instrument was an error-identification task. Two or three teachers with experience in translation teaching or political text translation independently annotated errors in the pre-test and post-test AI outputs using an MQM-informed classification. They then discussed the annotations and established a standard answer key specifying the location, type, and explanation of each error. Student annotations were compared with this key. A student annotation was counted as a true positive if it broadly matched the location and type of an error in the key, a false positive if it identified a problem not included in the key, and a false negative if a key error was missed. Precision, recall, and F1 were then calculated.

The third instrument was a 12-item AI critical literacy questionnaire on a five-point Likert scale. It included four dimensions: tool understanding, verification behavior, discourse/bias

sensitivity, and agency and ethics. Sample items included: 'I check whether AI-generated translations of political terms have official or authoritative sources'; 'I can judge whether fluency in an AI translation conceals a semantic shift'; 'When post-editing, I can explain which revisions are evidence-based rather than based only on intuition'; and 'I believe that the translator remains responsible for the final translation and cannot shift responsibility to AI.'

Table 3. Post-editing Quality Rubric

Dimension	Score	Description
Semantic accuracy	30	Accurately conveys facts, logical relations, agency, and stance without adding or omitting key information.
Terminology and proper names	20	Uses stable and evidence-based translations for international organizations, places, names, resolutions, and political terms.
Japanese fluency	15	Uses grammatically natural and coherent Japanese appropriate for political news or diplomatic discourse.
Political-diplomatic register	15	Controls modal force, politeness, stance boundaries, and institutional register appropriately.
Revision rationale and evidence	10	Explains revisions with reference to official terminology, parallel texts, context, or terminology resources.
Critical reflection on AI	10	Identifies limitations in AI output and explains why machine-generated wording is retained, revised, or rejected.

3.6. Rating Procedure and Reliability Control

Three teachers rated the student translations. Rater A was the course instructor and was familiar with the teaching goals. Rater B was a Japanese translation teacher who did not teach the class. Rater C had experience in political text translation, international communication, or translation quality assessment. Before formal rating, the three raters discussed the rubric and anchor samples using five practice segments. During formal scoring, student texts were anonymized and randomly ordered. All raters scored without seeing student identities; Raters B and C did not know whether a text came from the pre-test or the post-test, and Rater A scored the mixed files without viewing the pre/post labels. Inter-rater reliability was calculated using the intraclass correlation coefficient. Because the three raters were designated for this study, a two-way mixed-effects, absolute-agreement, average-measures ICC model was used.

3.7. Data Analysis

The quantitative analysis had four parts. First, descriptive statistics were calculated for total and dimensional post-editing scores, including means, standard deviations, medians, and interquartile ranges. Second, Shapiro-Wilk tests were used to examine the normality of pre-test/post-test difference scores. Given the small sample size (N = 15) and the fact that some difference scores did not stably meet the normality assumption, Wilcoxon signed-rank tests were used for pre-test/post-test comparisons. Third, effect size r was reported for Wilcoxon results and calculated as $|Z|/\sqrt{N}$. Fourth, precision, recall, and F1 were calculated for the error-identification task, followed by pre-test/post-test comparisons.

To answer RQ3, Spearman's rho was used to examine associations among changes in AI critical literacy, changes in error-identification F1, and gains in post-editing quality. Because of the small sample size, correlation results were interpreted as associations rather than causal evidence. Qualitative data from editing logs, revision explanations, classroom discussion, and interview comments were analyzed thematically, with attention to students' trust in AI output, evidence use, awareness of political register, and sense of translator responsibility.

Table 4. Variables, Data Sources, and Analysis Methods

Variable	Indicator	Data Source	Analysis Method
Post-editing quality	Total score and six dimensional scores	Three-rater scoring	ICC, descriptive statistics, Wilcoxon signed-rank test, effect size
Error-identification ability	Precision, recall, F1	Error annotation task	Descriptive statistics, Wilcoxon signed-rank test, effect size, error-type analysis
AI critical literacy	Total and four dimensional questionnaire scores	Pre/post questionnaire	Cronbach's alpha, descriptive statistics, Wilcoxon signed-rank test, effect size
Editing process	Number of revisions, types of rationale, evidence sources	Post-editing logs	Descriptive statistics, thematic analysis
Learning experience	Trust in tools, translator agency, perceived difficulties	Interviews/reflections	Thematic analysis
Variable relationships	Changes in AI critical literacy, F1, and post-editing quality	Pre/post difference scores	Spearman correlation

3.8. Figures Ethics and Data Handling

The study was conducted within a regular course. All tasks were part of normal learning activities. Before analysis, students were informed of the research purpose, the intended use of classroom data, and the principle of anonymization. Post-edited texts, questionnaires, logs, and interview materials were coded and did not include names, student numbers, or other identifying information. The research use of these data was intended for teaching improvement and academic reporting and did not affect course grades.

4. Figures

4.1. Inter-rater Reliability

The three raters scored the pre-test and post-test translations. The ICC for the total score was 0.84, with a 95% confidence interval of [0.69, 0.93], indicating good agreement. Dimensional ICCs ranged from 0.79 to 0.88. The highest agreement was found for terminology and proper

names, while political-diplomatic register showed the lowest, though still acceptable, agreement. This lower agreement is understandable because judging modal strength and diplomatic register involves finer interpretive decisions.

Table 5. Inter-rater Reliability

Scoring Dimension	ICC	95% CI	Interpretation
Total score	0.84	[0.69, 0.93]	Good
Semantic accuracy	0.85	[0.70, 0.93]	Good
Terminology and proper names	0.88	[0.75, 0.95]	Good; highest among dimensions
Japanese fluency	0.83	[0.66, 0.92]	Good
Political-diplomatic register	0.79	[0.58, 0.90]	Good, but relatively lower
Revision rationale and evidence	0.82	[0.64, 0.91]	Good
Critical reflection on AI	0.81	[0.62, 0.91]	Good

4.2. Changes in Post-editing Quality

As shown in Table 6, students' total post-editing scores increased from 72.40 (SD = 6.74) in the pre-test to 83.40 (SD = 8.15) in the post-test. The Wilcoxon signed-rank test showed that this gain was statistically significant, $Z = -2.17$, $p = .030$, $r = .56$, indicating a medium-to-large instructional effect.

Table 6. Pre-test/Post-test Differences in Post-editing Quality

Indicator	Pre-test M (SD)	Post-test M SD)	Statistic	p	Effect size r
Total score	72.40 (6.74)	83.40 (8.15)	Z = -2.17	.030	.56
Semantic accuracy	21.00 (2.82)	24.20 (3.36)	Z = -2.05	.041	.53
Terminology and proper names	14.80 (1.95)	17.20 (2.24)	Z = -2.42	.016	.62
Japanese fluency	11.20 (1.76)	11.90 (2.01)	Z = -1.54	.124	.40
Political-diplomatic register	10.80 (2.05)	12.90 (2.32)	Z = -2.33	.020	.60
Revision rationale and evidence	7.60 (1.38)	8.80 (1.55)	Z = -2.61	.009	.67
Critical reflection on AI	7.00 (1.40)	8.40 (1.68)	Z = -1.99	.047	.51

**Note. N = 15. The total score is the sum of the six dimensions. Wilcoxon signed-rank tests were used; effect size r was calculated as $|Z|/\sqrt{N}$.*

At the dimensional level, the largest gain appeared in revision rationale and evidence, which increased from 7.60 to 8.80, $Z = -2.61$, $p = .009$, $r = .67$. This suggests that students became better able to justify revisions with reference to the source text, context, and textual function rather than relying only on intuition. Terminology and proper names, as well as political-diplomatic register,

also improved significantly, indicating that the intervention helped students handle standardized political terminology and formal register more effectively.

Japanese fluency improved numerically but did not reach statistical significance, Z = -1.54, p = .124, r = .40. This pattern suggests that short-term instruction had a stronger effect on terminology awareness, register awareness, and evidence-based editing than on general target-language naturalness, which is likely to require longer-term linguistic development. The post-test standard deviations were also generally larger than the pre-test standard deviations, suggesting that students differed in how quickly they transferred the trained criteria to actual post-editing tasks.

4.3. Changes in Error-identification Ability

The error-identification task also showed improvement (Table 7). Precision increased from 0.62 to 0.76, recall from 0.49 to 0.68, and F1 from 0.55 to 0.71. All three changes were statistically significant. The larger gain in recall suggests that students missed fewer real errors after the intervention, while the gain in precision indicates that they also became more accurate in deciding what counted as an error.

Table 7. Pre-test/Post-test Differences in Error-identification Ability

Indicator	Pre-test M (SD)	Post-test M SD)	Statistic	p	Effect size r
Indicator	P	0.76 (0.15)	Z = -2.38	.017	.61
Precision	0.62 (0.11)	0.76 (0.15)	Z = -2.38	.017	.61
Recall	0.49 (0.13)	0.68 (0.18)	Z = -2.56	.011	.66
F1	0.55 (0.12)	0.71 (0.16)	Z = -2.48	.013	.64

Error-type analysis further clarifies the pattern (Table 8). The largest improvement occurred in terminology and proper-name errors, followed by political-diplomatic register errors and AI critical reflection-related errors. Grammar and fluency errors improved only slightly. This pattern is consistent with the nature of political texts: standardized terminology can be improved relatively quickly through terminology lists and parallel-text checking, whereas register and critical interpretation require more sustained discourse-level training.

Table 8. Changes in Error-type Recognition Rates

Error Type	Pre-test Recognition Rate	Post-test Recognition Rate	Change
Semantic errors	.58	.70	+.12
Terminology and proper-name errors	.55	.78	+.23
Grammar and fluency errors	.64	.72	+.08
Political-diplomatic register errors	.42	.58	+.16
AI critical reflection-related errors	.38	.52	+.14

4.4. AI Critical Literacy and Process Evidence

The AI critical literacy questionnaire showed good internal consistency, with Cronbach's alpha = 0.84. The total score increased from 3.21 (SD = 0.46) to 3.78 (SD = 0.58), Z = -2.47, p = .013, r = .64. The strongest gains were found in verification behavior and agency/ethics. Verification behavior increased from 3.05 to 3.86, and agency/ethics from 3.18 to 3.82. These results suggest that students became more willing to check AI-generated translations against authoritative sources and more aware that the translator remains responsible for the final text.

Correlation analysis showed that changes in AI critical literacy were positively associated with gains in post-editing quality, rho = .61, p = .016. Changes in error-identification F1 were also positively associated with gains in the terminology and proper-name dimension, rho = .58, p = .023. These associations suggest that students who became more critical and better able to detect AI errors also tended to improve more in post-editing performance.

Student logs and interview comments support this interpretation. One student (S03) noted that earlier she tended to accept AI translations if the Japanese sounded fluent, but later she first checked whether the source text's stance and key terms had been preserved. Another student (S11) wrote that when revising terminology, she no longer simply asked AI which expression sounded natural, but checked fixed expressions on official websites and in Japanese news texts. Such comments indicate a shift from accepting fluent AI output to checking it, justifying revisions, and making evidence-based decisions.

Table 9. Pre-test/Post-test Differences in AI Critical Literacy

Indicator	Pre-test M (SD)	Post-test M SD)	Statistic	p	Effect size r
Total AI critical literacy	3.21 (0.46)	3.78 (0.58)	Z = -2.47	.013	.64
Tool understanding	3.30 (0.50)	3.62 (0.57)	Z = -1.86	.063	.48
Verification behavior	3.05 (0.52)	3.86 (0.63)	Z = -2.73	.006	.71
Discourse/bias sensitivity	3.31 (0.54)	3.73 (0.61)	Z = -2.05	.041	.53
Agency and ethics	3.18 (0.49)	3.82 (0.61)	Z = -2.55	.011	.66

5. Discussion

The findings answer RQ1 by showing that a six-week AI-assisted post-editing intervention significantly improved students' overall post-editing quality. The most visible gains were not in general fluency but in revision rationale, terminology, and political-diplomatic register. This matters pedagogically. In political-text translation, post-editing is not simply language polishing. It is evidence-based decision-making across source-text facts, target-language norms, institutional register, and AI-generated alternatives.

RQ2 is addressed by the gains in error-identification ability. Students improved in precision, recall, and F1, with the largest increase in terminology and proper-name error recognition. This

suggests that error annotation, terminology verification, and comparison with teacher reference versions can quickly strengthen students' ability to locate concrete problems in AI output. However, political-diplomatic register and AI critical reflection remained more difficult, which indicates the need for longer-term discourse-level training.

RQ3 is addressed by the positive associations between AI critical literacy, error identification, and post-editing improvement. Students who became more willing to verify AI output and more aware of translator responsibility tended to show stronger gains in translation performance. Although the small sample size means that these correlations should not be interpreted as causal evidence, the quantitative patterns and student reflections point in the same direction: critical engagement with AI is closely related to better post-editing practice.

From the perspective of teaching reform, the study supports a three-part framework: technology use, technological criticism, and cross-cultural sensitivity. Technology use appears in students' ability to generate draft translations, build terminology lists, and record editing processes. Technological criticism appears in their ability to identify errors, explain their causes, and justify revisions. Cross-cultural sensitivity appears in their handling of Chinese-Japanese differences in register, stance expression, and institutional discourse.

The study also cautions against treating AI evaluation as a final authority. Large language models can help students notice possible problems, generate alternatives, and compare versions. Yet the final quality of a political translation must still be verified through teachers, students, and authoritative sources. This is especially true when the text involves state positions, international organizations, policy terms, or politically sensitive expressions. The goal of AI-assisted post-editing pedagogy should therefore not be to make students use AI faster, but to help them remain alert, evidence-oriented, and professionally responsible when working with AI output.

6. Conclusion

This study designed and implemented a quasi-experimental teaching intervention for MTI students' Chinese-Japanese post-editing of political texts. Using authentic political-diplomatic texts and AI-generated Japanese drafts, the study combined pre-test/post-test tasks, three-rater scoring, MQM-informed error annotation, an AI critical literacy questionnaire, editing logs, and student reflections. The results indicate that students' post-editing quality, machine-translation error-identification ability, and AI critical literacy improved after the intervention, with particularly clear gains in terminology, evidence-based revision, and awareness of political-diplomatic register.

The practical value of this study lies in translating broad teaching-reform goals such as artificial intelligence, translation practice, technological criticism, and cross-cultural communication into classroom tasks that can be observed and assessed. The design is suitable for small graduate classes and offers a framework for future empirical research on AI-assisted translation pedagogy.

Several limitations should be acknowledged. The sample was small, and the one-group pre-test/post-test design did not include a strict control group. The materials were limited to political-diplomatic texts, so it remains to be seen whether the findings transfer to literary, journalistic, or business translation. The rubric and AI critical literacy questionnaire also require validation with larger samples. Future research may extend the design to Chinese-Japanese-English translation, compare different AI systems, and incorporate process data such as keystroke logging, screen recording, or eye tracking to better explain students' decision-making in AI-mediated translation.

Funding

This work was supported by a university-level general project on educational and teaching reform for graduate students “Translation Research and Translation Practice in the Era of Artificial Intelligence” (Grant No.111220264041) of Beijing International Studies University. It is also a phased research outcome of the 2025 Beijing Digital Education Research Project “Research on a Human–AI Collaborative Teaching Model in Translation Education in the Era of Artificial Intelligence” (Grant No. BDEC2025619139).

Author Contributions:

Qiushi Gu, as the first and corresponding author, led the conceptualization, methodology, research design, project supervision, and overall coordination of the study, and was responsible for the final revision of the manuscript. Shiyan Wang contributed to the theoretical framing, validation of the research design, and critical revision of the manuscript. Xinyu Ji contributed to data organization, statistical analysis, and interpretation of the empirical results. Xinyao Ren contributed to data collection, qualitative coding, and preparation of the preliminary draft. All authors have read and approved the final version of the manuscript.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

European Commission, Directorate-General for Translation. (2022). European Master's in Translation: Competence Framework 2022. Publications Office of the European Union. https://doi.org/10.2782/858200

International Organization for Standardization. (2017). ISO 18587:2017 Translation services — Post-editing of machine translation output — Requirements. ISO.

Jiao, W., Wang, W., Huang, J.-T., Wang, X., Shi, S., & Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Kenny, D., & Doherty, S. (2014). Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. The Interpreter and Translator Trainer, 8(2), 276–294.

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 193–203. https://aclanthology.org/2023.eamt-1.19/

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. The Journal of Specialised Translation, 25, 131–148.

Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica, 12, 455–463.

Miao, F., Shiohira, K., & Lao, N. (2024). AI competency framework for students. UNESCO. ISBN 978-92-3-100709-5.

O'Brien, S. (2011). Towards predicting post-editing productivity. Machine Translation, 25(3), 197–215.

PACTE Group. (2003). Building a translation competence model. In F. Alves (Ed.), Triangulating Translation: Perspectives in process oriented research (pp. 43–66). John Benjamins. https://doi.org/10.1075/btl.45.06pac

Schäffner, C. (2004). Political discourse analysis from the point of view of translation studies. Journal of Language and Politics, 3(1), 117–150.

Veldhuis, A., Lo, P. Y., Kenny, S., & Antle, A. N. (2025). Critical artificial intelligence literacy: A scoping review and framework synthesis. International Journal of Child-Computer Interaction, 43, 100708.