International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Human-AI Collaborative Feedback in Translator Training: A Mixed-Methods Study of Translation Quality, Revision Behavior, and Learner Perceptions

Shiyue Chen
School of Humanities and Foreign Languages, Zhejiang Shuren University, Hangzhou, China
Corresponding Author: Shiyue Chen, **E-mail**: shiyue@zjsru.edu.cn

| ABSTRACT

The integration of large language models (LLMs) into language education has prompted renewed interest in AI-assisted feedback, yet purely automated feedback remains vulnerable to contextual misalignment, cultural misreading, and reliability concerns that are particularly consequential in translation training. A human-AI collaborative feedback model, in which an instructor curates, corrects, and supplements LLM-generated commentary before students revise, offers a theoretically motivated alternative, yet its pedagogical effects in translator education remain empirically underexplored. This mixed-methods study examines the impact of such a hybrid feedback approach on undergraduate Chinese-to-English student translators. Forty senior undergraduates translated a 1,500-word cultural heritage text and received ChatGPT-4o-generated feedback subsequently reviewed and annotated by an experienced instructor using a color-coded transparency system. Quantitative analysis using a Multidimensional Quality Metrics (MQM) rubric revealed significant pre-to-post gains across all measured dimensions (overall MQM composite: Δ +1.20 on a 5-point scale, $p < .001$), with the largest improvements in terminology (Δ +1.47) and accuracy (Δ +1.32) and meaningful gains in cohesion, cultural adaptation, register, language conventions, and format (all $p < .001$). Think-aloud protocols revealed a consistent two-stage revision pattern and active source evaluation behavior, with students demonstrating greater decisiveness when AI and instructor annotations converged and deeper deliberation when they diverged. Student perception surveys indicated high ratings across clarity, trustworthiness, usefulness, and pedagogical value, with no significant differences between high- and low-performing students. Instructors reported meaningful workload relief on routine corrections while retaining pedagogical authority over higher-order feedback. These findings suggest the potential of a human-in-the-loop feedback framework for translator training in which AI handles systematic error detection while instructors validate, contextualize, and model evaluative judgment, a model that warrants further controlled investigation as a means of enhancing translation competence without displacing the pedagogical depth that professional translator development requires.

| KEYWORDS

human-AI collaborative feedback; translation pedagogy; translator training; feedback literacy; translation quality assessment

| ARTICLE INFORMATION

ACCEPTED: 30 April 2026 **PUBLISHED:** 06 May 2026 **DOI:** 10.32996/ijllt.2026.5.1.5

1. Introduction

Feedback is a cornerstone of translation education, guiding learners to refine accuracy, idiomatic expression, and cultural adaptation across multiple competence dimensions (Washbourne, 2014; Yu et al., 2020). In professional translator training, detailed instructor commentary remains the gold standard for supporting learners' revision and self-assessment processes. However, the manual provision of individualized, multi-dimensional feedback is resource-intensive: it does not scale easily to large cohorts, and the turnaround time between submission and feedback often limits students' ability to act on commentary while their translational decisions remain cognitively accessible (Jiao et al., 2025; S. Xu et al., 2025). This structural tension, between the pedagogical value of rich feedback and the practical constraints of delivering it, constitutes a longstanding challenge in translation program design.

The rapid development of large language models (LLMs) such as ChatGPT has introduced a potentially transformative supplement to this equation. Recent studies in writing instruction report that LLM-generated feedback can significantly improve revision quality

and student motivation(Mahapatra, 2024; Steiss et al., 2024), and preliminary evidence from translation contexts suggests comparable potential (X. Xu et al., 2025). LLMs can process full-length texts quickly, flag errors systematically across multiple dimensions, and generate explanatory commentary at scale, capabilities that directly address the bottleneck of manual feedback provision. Yet AI feedback is not without risk (Ma et al., 2024). LLMs may generate inaccurate suggestions, overlook pragmatic and cultural context, or apply generic corrections that misrepresent the target text's communicative demands. In translation specifically, where meaning, register, and cultural equivalence must be conveyed faithfully across languages, such errors could actively mislead learners rather than support them. The question, then, is not simply whether LLMs can generate feedback, but under what conditions that feedback is pedagogically reliable and trustworthy.

One response to this challenge is a human-in-the-loop (HITL) model (Wu et al., 2022; Mosqueira-Rey et al., 2023), in which instructor oversight is built into the feedback process rather than displaced by it. In professional translation technology, interactive machine translation already embodies this logic: systems generate draft outputs that human translators post-edit for accuracy, fluency, and contextual appropriateness (Koponen, 2016). Analogous collaborative architectures have shown promise in other educational domains, for instance, in programming education, where teacher-AI systems allow instructors to select and rewrite AI-generated hints, yielding more pedagogically effective support (Orak, 2025). Despite this convergent interest in hybrid feedback models, their application to translation pedagogy remains underexplored. Existing studies have largely examined AI-only or human-only feedback conditions, leaving open the question of how structured human-AI collaboration, specifically, instructor curation of LLM-generated commentary, affects translation quality, student revision behavior, and learner and instructor perceptions of the feedback process. This gap motivates the present study.

We investigate a collaborative feedback workflow in which ChatGPT-4o first generates comments on a student's draft translation, and an instructor then reviews, annotates, and supplements those comments before the student revises. The instructor distinguishes endorsed AI comments from modified or instructor-added remarks using a color-coded annotation system, making the source and reliability of each piece of feedback transparent to the learner. This design operationalizes the HITL principle in a classroom-feasible format and raises three research questions:

RQ1: To what extent does human-AI collaborative feedback improve the translation quality of undergraduate student translators across multiple dimensions, including accuracy, terminology, style, and cultural adaptation?
RQ2: How do students engage cognitively with the hybrid feedback during the revision process, and what patterns of revision behavior emerge from their think-aloud protocols?
RQ3: How do learners and instructors evaluate the clarity, trustworthiness, usefulness, and pedagogical value of the human-AI collaborative feedback model?

Addressing these questions contributes to a growing body of evidence on AI integration in language education while extending that evidence to the translation context, a domain with distinctive demands on cultural and pragmatic judgment that general writing feedback studies do not fully address. By focusing on the collaborative condition, the study examines the role of instructor oversight within an AI-assisted environment, with implications for how translation programs can harness the efficiency of LLMs without sacrificing the pedagogical depth that developing translator competence requires.

2. Literature Review

FeedbackFeedback occupies a central role in translator education, functioning not merely as error correction but as a mechanism through which learners internalize decision-making strategies for register, cultural equivalence, and target-audience awareness (Neunzig & Tanqueiro, 2005; Kiraly, 2015). Washbourne (2014) observed that iterative feedback enables students to adopt a reader's perspective and cultivate self-assessment capacities, a finding consistent with Mellinger's (2019) demonstration that metacognitive self-assessment abilities develop progressively through targeted translation coursework, with trainees shifting over time from dictionary-dependent micro-level approaches toward more contextually aware problem recognition. Empirical studies have confirmed that well-designed corrective feedback measurably improves translation quality. Chen et al. (2021) found that a combination of written feedback strategies, particularly corrective and explanatory comments, significantly enhanced revision accuracy in Chinese-to-English classrooms, and Zheng et al. (2020) further showed that students' engagement with teacher translation feedback operates across affective, behavioral, and cognitive dimensions, with negative affect potentially undermining the cognitive processing necessary for genuine learning. The type and explicitness of feedback also matter. Li et al. (2023) demonstrated that the availability of explicit solutions and the clarity of feedback strategy significantly predicted students' uptake during peer assessment in university translation courses, suggesting that how feedback is delivered shapes engagement as much as its content.

Peer feedback, while beneficial for learner reflection, carries well-documented limitations in the translation context. Lin et al. (2021) found that peer reviewers in advanced Chinese-to-English translation courses predominantly offered direct surface corrections,

rarely addressing idiomatic or culturally subtle issues, a pattern that constrains the depth of revision peer feedback can drive. Conventional automated evaluation tools such as BLEU and TER metrics similarly lack the generativity and pedagogical explanation that learning requires (Chauhan & Daniel, 2023; C. Han & Lu, 2023). These limitations have created pressure to explore alternative feedback sources, and the emergence of LLMs has reopened this question considerably. Before discussing AI-generated feedback, it is worth noting the assessment framework underpinning the present study. We adopt the Multidimensional Quality Metrics (MQM) rubric (Lommel et al., 2014; Mariana et al., 2015), which evaluates translation across dimensions including accuracy, terminology, fluency, style, and cultural adaptation. The PACTE Group's et al. (2018) extensive experimental research has further established that translation competence is multi-componential, comprising linguistic, cultural, instrumental, and strategic sub-competencies, with strategic competence playing a central coordinating role across training stages. This multi-dimensional understanding of what translation competence entails frames our analysis of feedback effects across different quality dimensions. The deployment of LLMs in educational feedback contexts has grown rapidly, with a body of evidence now emerging on both their promise and their limitations. Wang et al. (2024) demonstrated in a randomized trial that ChatGPT-3.5-generated feedback significantly increased secondary students' revision performance and boosted motivation. More recently, Steiss et al. (2024) compared the quality of human and ChatGPT feedback on student writing, finding that while trained human evaluators provided higher quality feedback in most categories, particularly accuracy and supportive tone, ChatGPT excelled at criteria-based commentary, suggesting its greatest value may lie in systematic, rubric-aligned feedback on early drafts. Escalante et al. (2023) found no significant difference in writing outcomes between students receiving ChatGPT-4 feedback and those receiving human tutor feedback over six weeks, with students nearly evenly split in their source preference, a result that challenges simple hierarchies of feedback quality. Mahapatra (2024) similarly reported significant positive impacts of ChatGPT as a formative feedback tool on ESL students' writing skills across successive assessments, with strong student satisfaction. However, LLM feedback is not without pitfalls. It can be overly generic, contextually misaligned, or overconfident, particularly in domains requiring cultural and pragmatic judgment. Dai et al. (2023) found that while LLM-generated feedback was often more detailed and process-oriented than instructor feedback, its reliability was lower on certain evaluative dimensions, highlighting the need for quality control mechanisms. Er et al. (2025) further observed that students perceived instructor feedback as significantly more useful than AI feedback in terms of actionability, even when AI feedback was comparably detailed, a utility gap with direct implications for how AI feedback tools are introduced and framed in classroom settings.

Trust in AI feedback is also mediated by contextual transparency (Carless, 2012). Cheng et al. (2023) found that students' trust ratings shifted markedly when the source of feedback was made explicit, with provider information playing a decisive role in learners' willingness to act on suggestions. Banihashem et al. (2024) extended this line of inquiry by comparing peer-generated and AI-generated feedback on argumentative essays, finding that ChatGPT produced more descriptive and detailed feedback while peer reviewers identified more specific problems, suggesting the two sources function as complements rather than substitutes, each offering distinct epistemic value. These findings collectively motivate the human-in-the-loop design adopted in the present study: rather than treating AI and human feedback as alternatives, a collaborative model may capture the systematic strengths of each while mitigating their respective weaknesses.

The concept of human-in-the-loop (HITL) in AI refers to workflows in which human expertise supervises or augments machine outputs, preserving contextual judgment and mitigating errors (Kumar et al., 2024). Molenaar (2022) proposes a six-level automation framework, ranging from full teacher control to full technology control, and argues that hybrid human-AI technologies positioned in the middle range offer the most educationally productive configurations, augmenting rather than replacing human intelligence. Holstein et al. (2020) similarly propose a conceptual framework for human-AI hybrid adaptivity in education, identifying multiple dimensions along which teachers and AI systems can mutually augment each other's capabilities, noting that teachers possess knowledge of students' emotional states and social dynamics to which AI systems are typically blind. Memarian & Doleck (2024), in a systematic review of HITL studies in educational AI, found that most existing implementations involve insufficiently deliberate design of human-AI relationships, pointing to significant opportunity for more principled hybrid feedback architectures. In the domain of translation specifically, Koponen (2016) demonstrated that post-editing high-quality machine translation increases productivity, but that effort varies substantially by error type, with word-order errors and mistranslated idioms proving most cognitively demanding, findings with direct implications for how human oversight should be allocated within a collaborative feedback system. Wiboolyasarin et al. (2024) further demonstrated in an L2 writing context that integrating ChatGPT corrective feedback into a structured three-step collaborative process significantly improved writing proficiency, validating the principle that structured human-AI collaboration designs can outperform unmediated AI feedback alone. In our study, this logic is extended to translation feedback: ChatGPT-4o generates an initial commentary, and an instructor then curates, corrects, and contextualizes it, treating the AI output as a first draft subject to human editorial judgment.

The effectiveness of any feedback intervention ultimately depends on how students engage with it cognitively, affectively, and behaviorally (Sato & Loewen, 2018). Carless & Boud (2018) provide a foundational framework for student feedback literacy

comprising four features, appreciating feedback, making judgments, managing affect, and taking action, arguing that genuine cognitive engagement requires learners to develop these capacities before they can productively use feedback to enhance their work. Y. Han & Xu (2021) found in a study of Chinese undergraduates that feedback literacy comprises cognitive, social, and affective components, and that unbalanced development across these three dimensions frequently limits students' behavioral engagement with written corrective feedback. In translation specifically, think-aloud research has shown that students use feedback to trigger both corrections and broader reflections on linguistic choices. Kim & Bowles (2019) demonstrated through think-aloud protocols that learners processed different error types at markedly different depths depending on feedback type, with sentential and paragraph-level errors receiving deeper processing under reformulation feedback, evidence that the relationship between feedback design and cognitive engagement is neither simple nor uniform. Tabari et al. (2023) further found that combining think-aloud protocols with written corrective feedback had differential effects depending on feedback type and linguistic dimension, with think-aloud verbalization not universally triggering deeper processing. Chen & Zhou (2026) observed that students engaging with ChatGPT-generated translation feedback invested considerable cognitive effort, re-reading comments, consulting corpora, and double-checking suggestions, even when the feedback was ostensibly understandable. In a hybrid feedback setting, where students must navigate between AI and instructor commentary of potentially differing quality and focus, this negotiation process is likely to be more complex. Winstone et al. (2019) caution that lower-achieving students, potentially those who would benefit most from rich feedback, are least likely to engage with feedback voluntarily, underscoring the importance of structuring the feedback environment to scaffold engagement rather than assuming it will occur spontaneously. Our study uses think-aloud protocols and semi-structured interviews to examine how students navigate this dual-source feedback environment, and whether the transparency afforded by explicit source labeling supports the kind of critical, evaluative engagement that feedback literacy frameworks consider essential.

3. Methodology

3.1 Participants

Forty senior undergraduates enrolled in a Chinese-to-English translation program at a major Chinese university participated. All had passed the national TEM-4 proficiency test but had no prior experience with AI-assisted translation tools. Four experienced translation instructors (5+ years of teaching) also participated as feedback editors and interviewees. Ethics approval was obtained and all participants provided written informed consent.

3.2 Materials and Procedures

The translation task consisted of a 1,500-word Chinese text on cultural heritage (Appendix A), selected for its demands on lexical choice, cultural adaptation, and register. Students translated the text individually during a 90-minute session. GPT-4o (accessed via ChatGPT, September 2025) was selected as the AI feedback generation tool given its multimodal capabilities and status as the default flagship model available at the time of data collection.

The study followed a convergent mixed-methods design (Creswell & Plano Clark, 2018) with three stages.
Stage 1 - Initial Translation. Students produced a draft translation under exam conditions.

Stage 2 - Human-AI Collaborative Feedback. Each student received an integrated feedback report developed in two steps. First, ChatGPT-4o was prompted via a standardized instruction to comment on the draft across four dimensions: accuracy, terminology, style, and cultural equivalence. These four prompt dimensions were designed to broadly correspond to the seven MQM rubric categories used in quality assessment, with style encompassing register and cohesion, and cultural equivalence capturing cultural adaptation. Second, an instructor reviewed the AI-generated comments using a color-coded annotation system: AI comments judged accurate were endorsed (green); flawed suggestions were corrected (blue); gaps in the AI feedback were supplemented with instructor remarks. The final feedback sheet thus made the origin and reliability of each comment transparent to the student. Figure 1 illustrates a sample feedback sheet produced through this annotation procedure, showing all three comment types as they would appear to a student.

Stage 3 - Revision, Submission, and Reflection. Students revised their translations in a 30-45-minute session while thinking aloud; verbal protocols were audio-recorded. After submission, students completed a perception questionnaire and a subset (n = 10, purposively sampled across performance levels) participated in 10–20-minute semi-structured interviews. Instructors took part in two 90-minute focus group sessions.

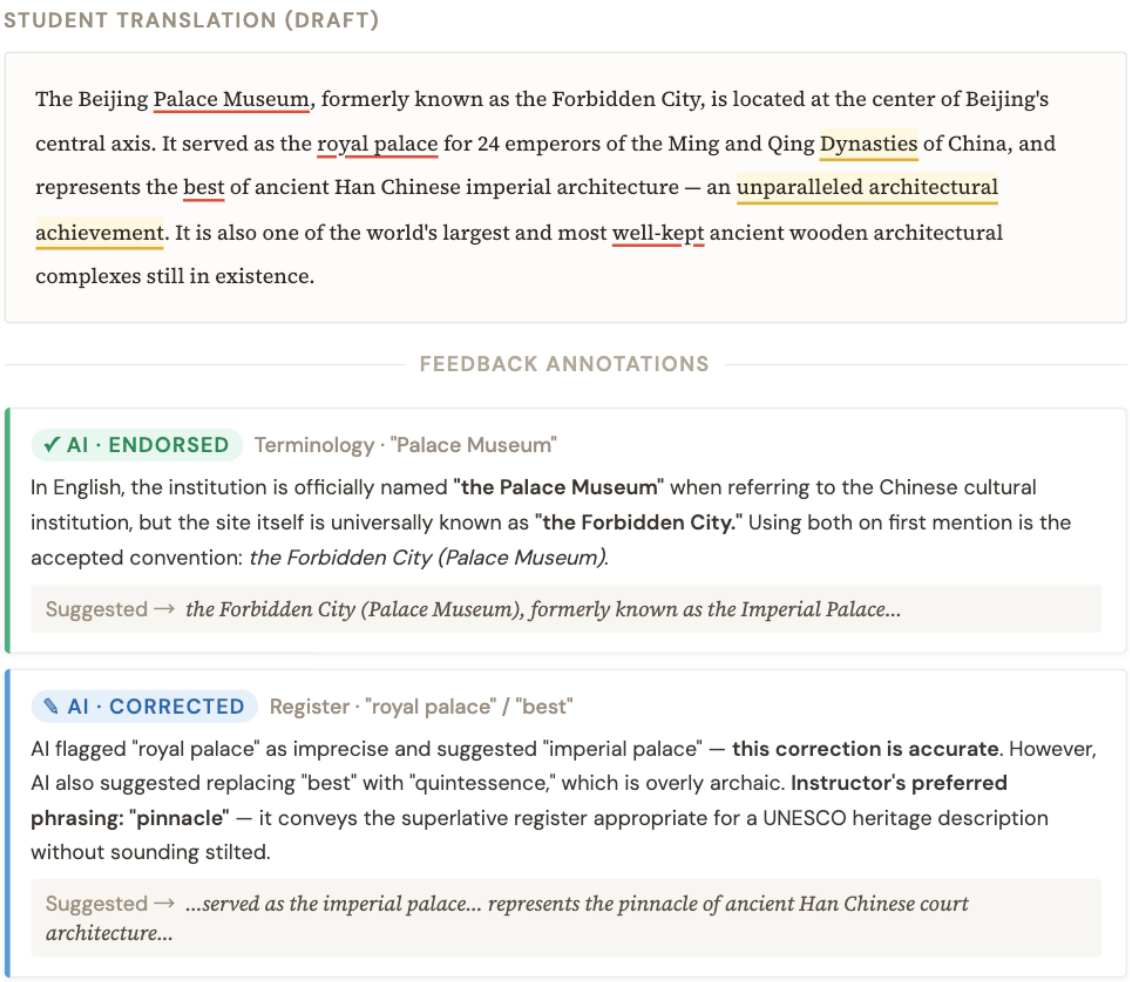

Figure 1 Sample human-AI collaborative feedback sheet demonstrating the color-coded annotation system

3.3 Translation Quality Assessment and Data Analysis

Pre- and post-revision translations were scored by two certified translators using an analytic rubric adapted from the Multidimensional Quality Metrics (MQM) framework across seven dimensions: accuracy, terminology, language conventions, cohesion, cultural adaptation, register, and format (each 0-5). Raters were blind to student identity and revision order; inter-rater reliability was high (Cohen's $\kappa = 0.86$).

Think-aloud transcripts were subjected to inductive content analysis to identify patterns in students' revision orientations. An initial coding scheme was developed from a subset of transcripts and iteratively refined through discussion between two researchers until consensus was reached. Two researchers independently coded 20% of the transcripts as a reliability check ($\kappa = 0.84$), indicating strong inter-rater agreement. Interview and focus group data were analyzed thematically following Braun & Clarke (2006), with themes generated inductively. A 20-item Likert-scale questionnaire (1-5) measured four dimensions (Appendix B): clarity, trustworthiness, usefulness, and pedagogical value of the hybrid feedback. ANOVAs examined whether perceptions varied by initial performance level. Students were classified as high- or low-performing based on a median split of their initial draft MQM composite scores, yielding two groups of equal size ($n = 20$ each).

4. Findings

4.1 RQ1

Revised translations showed consistent improvements across all measured dimensions. The overall MQM composite score rose from a pre-revision mean of 2.92 ($SD = 0.61$) to a post-revision mean of 4.12 ($SD = 0.54$), yielding a mean gain of $\Delta +1.20$ ($t(39) = 13.49$, $p < .001$, Cohen's $d = 2.13$). Subscore gains were largest in terminology ($\Delta +1.47$, $t(39) = 16.03$, $p < .001$, $d = 2.53$) and accuracy ($\Delta +1.32$, $t(39) = 15.17$, $p < .001$, $d = 2.40$), with language conventions ($\Delta +1.28$, $t(39) = 15.27$, $p < .001$, $d = 2.42$), cultural adaptation ($\Delta +1.12$, $t(39) = 12.43$, $p < .001$, $d = 1.96$), register ($\Delta +1.09$, $t(39) = 12.53$, $p < .001$, $d = 1.98$), cohesion ($\Delta +1.07$, $t(39) = 12.53$, $p < .001$, $d = 1.98$), and format ($\Delta +1.07$, $t(39) = 13.28$, $p < .001$, $d = 2.10$) also improving significantly. Effect sizes were

uniformly large across all dimensions (Cohen's d range: 1.96–2.53), indicating that the observed gains were not only statistically significant but also of substantial practical magnitude. No student's score declined. Full details are presented in Table 1.

Table 1. Pre- and post-revision Translation Quality (N = 40)

Dimension	Pre-M (SD)	Post-M (SD)	Δ	SD_diff	t(39)	p	Cohen's d
Overall	2.92 (0.61)	4.12 (0.54)	+1.20	0.52	13.49	<.001	2.13
Accuracy	2.76 (0.65)	4.08 (0.57)	+1.32	0.55	15.17	<.001	2.40
Terminology	2.65 (0.68)	4.12 (0.59)	+1.47	0.58	16.03	<.001	2.53
Language Conventions	3.10 (0.58)	4.38 (0.49)	+1.28	0.53	15.27	<.001	2.42
Cohesion	2.95 (0.62)	4.02 (0.55)	+1.07	0.54	12.53	<.001	1.98
Cultural Adaptation	2.80 (0.66)	3.92 (0.58)	+1.12	0.57	12.43	<.001	1.96
Register	2.88 (0.63)	3.97 (0.56)	+1.09	0.55	12.53	<.001	1.98
Format	3.05 (0.57)	4.12 (0.50)	+1.07	0.51	13.28	<.001	2.10

Note. Scores are rated on a 5-point scale (0 = lowest, 5 = highest). Δ = post-M − pre-M. SD_diff = standard deviation of paired differences. All paired-samples t-tests: df = 39. Cohen's d computed as Δ / SD_diff. Inter-rater reliability was strong (Cohen's κ = .86). All p-values reflect two-tailed tests.

The pattern of gains across dimensions is notable. Terminology and accuracy, areas where GPT-4o's systematic scanning capability is well-suited to detection, showed the steepest improvements, together accounting for the largest share of the total composite gain. Cohesion and cultural adaptation, which depend more heavily on contextual and pragmatic judgment, showed comparatively smaller but still statistically significant and large-magnitude gains, suggesting that instructor annotation was effective in addressing higher-order issues that AI feedback alone is less likely to resolve reliably. Raters' blind scoring and high inter-rater reliability (Cohen's κ = .86) support the validity of these quality assessments.

4.2 RQ2

Think-aloud transcripts revealed a consistent two-stage revision pattern. Students first addressed surface-level issues (vocabulary substitution, grammar correction), then shifted to deeper restructuring (sentence syntax, register, cultural equivalence). Representative of the first stage, one student narrated: "*First I went through and fixed the sentence structures and word order as the feedback flagged. Once those were done, I tackled the more complex phrasing and cultural nuance, which took a bit more thought.*" The second stage was characterized by longer deliberative episodes, during which students frequently re-read the source text, compared the flagged passage against their own draft, and weighed the AI annotation against the instructor's revised comment before committing to a revision. Students did not passively apply suggestions; they actively evaluated the source of each comment. When AI and instructor annotations diverged, students typically re-read the source text or consulted supplementary resources before deciding. One student explained: "*GPT flagged this term, but the teacher added a note, I trust the teacher's context here, so I'll use the refined version.*" Another noted: "*It helped a lot to see both viewpoints, but I had to pause and think each time which one was correct.*"

A third pattern also emerged: in cases where the AI and instructor annotations aligned, students reported proceeding more quickly and with greater confidence. As one participant put it: "*When both the AI and the teacher pointed to the same problem, I didn't need to think twice, I just fixed it.*" This suggests that convergent feedback from the two sources functioned as a reliability signal, reducing students' evaluative burden and accelerating the revision process for unambiguous corrections. Such metacognitive behavior was observed across performance levels, suggesting that the transparency of the dual-source feedback sheet prompted deliberate, critical engagement rather than automatic compliance. Notably, even lower-performing students demonstrated active source evaluation, indicating that the hybrid format supported critical engagement regardless of baseline proficiency.

4.3 RQ3

Students rated the hybrid feedback highly on all four dimensions (Table 2). Usefulness received the highest mean (M = 4.7, SD = 0.5), followed by trustworthiness (M = 4.6, SD = 0.4), clarity (M = 4.4, SD = 0.5), and pedagogical value (M = 4.3, SD = 0.6). Internal consistency was strong across all four subscales (Usefulness α = .88, Trustworthiness α = .85, Clarity α = .83, Pedagogical Value α = .81). Students were classified as high- or low-performing based on a median split of their initial draft MQM composite scores (median = 2.90), yielding two groups of equal size (n = 20 each). One-way ANOVAs revealed no significant differences between groups on any subscale (Usefulness: F(1, 38) = 1.24, p = .27; Trustworthiness: F(1, 38) = 0.87, p = .36; Clarity: F(1, 38) = 1.76, p = .19; Pedagogical Value: F(1, 38) = 2.14, p = .15), indicating broad acceptance across ability levels. This uniformity across performance groups is noteworthy: it suggests that the hybrid model was not perceived as benefiting only stronger students who

might more readily evaluate AI suggestions but was valued equally by students who entered the task with lower draft quality scores.

Interview data corroborated these ratings: students consistently attributed their confidence in AI suggestions to the visible instructor endorsement. As one participant noted: "*Knowing the teacher had checked it made me trust the suggestions more.*" Others elaborated on the value of the dual-source structure itself: "*Having both the AI comment and the teacher's note side by side made it easier to see which issues were really important, if the teacher flagged it too, I knew it was something I had to fix.*" Several students also commented on the pedagogical value dimension specifically, noting that the feedback process had made them more aware of systematic patterns in their translation errors: "*I realized I keep struggling with the same type of sentence structure. Seeing it flagged repeatedly made me think about why, not just fix it each time.*"

Table 2. Student Perceptions of Hybrid Feedback (N = 40)

Dimension	M	SD	Range	% ≥4	α	F(1,38)	p
Usefulness	4.7	0.5	3–5	95%	.88	1.24	.27
Trustworthiness	4.6	0.4	3–5	93%	.85	0.87	.36
Clarity	4.4	0.5	3–5	88%	.83	1.76	.19
Pedagogical Value	4.3	0.6	2–5	85%	.81	2.14	.15

Note. Items were rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Cronbach's α is reported separately for each subscale; overall scale internal consistency was α = .87. % ≥ 4 indicates the proportion of students who selected Agree or Strongly Agree as their mean subscale response. ANOVA revealed no significant differences between high- and low-performing student groups on any subscale (all $p > .10$).

Instructors reported that the model substantially reduced their corrective workload. On average, instructors estimated modifying or supplementing approximately 20% of AI-generated comments, primarily in areas involving cultural context and idiomatic register, precisely the domains where LLM outputs were least reliable. One instructor observed: "*GPT caught routine errors, I can skip straight to discussing style issues.*" Another highlighted the importance of maintaining pedagogical authority: "*I was always the final arbiter of what feedback they got.*" Beyond workload reduction, instructors described a qualitative shift in how they engaged with student work. Rather than spending most of their annotation time on surface corrections, they reported being able to focus almost entirely on higher-order issues, register choices, cultural equivalence, and argumentative coherence in extended passages. As one instructor noted: "*The AI handles the mechanical side, which frees me to have a real conversation with the student's translation choices rather than just marking errors.*" Another observed that reviewing and annotating AI-generated comments had itself become a productive reflective exercise: "*Going through what GPT flagged made me think more explicitly about what I value in a good translation, it was clarifying, even for me.*" All four instructors acknowledged a learning curve associated with prompt design and AI output evaluation, suggesting that instructor preparation constitutes a necessary condition for effective implementation of the model. Three of the four also expressed interest in continuing to use the hybrid format in future courses, contingent on further refinement of the prompting process and clearer institutional guidance on how to communicate the role of AI tools to students.

5. Discussion

The quantitative gains in translation quality are consistent with the core premise of the human-AI collaborative model among senior undergraduate translators with no prior AI feedback experience. All dimensions of the MQM rubric improved substantially after revision, with the most pronounced increases in terminology and grammatical accuracy, areas where ChatGPT-4o's systematic scanning capability is well-suited to detecting surface errors. Importantly, style and cultural adaptation also improved, dimensions that purely automated feedback studies have found more resistant to change (S. Chen & Zhou, 2024). This pattern suggests that the instructor's curatorial role may have been essential rather than supplementary, potentially filling the contextual and cultural gaps that AI output alone tends to leave. However, without a control condition isolating AI-only or instructor-only feedback, it is not possible to attribute these gains exclusively to the collaborative structure; future controlled designs will be necessary to establish the relative contribution of each component. These findings extend preliminary evidence of LLM-driven quality gains from general writing to the translation context (Mohammed, 2025), while raising questions about the view that AI feedback is sufficient for higher-order language development (Derakhshan & Taghizadeh, 2025); our data suggest that human oversight warrants further investigation as a potentially necessary condition when the feedback target involves culturally embedded meaning.

Beyond quality outcomes, the think-aloud data offer insight into how students processed the hybrid feedback, adding a process-level dimension that purely quantitative studies cannot capture. Rather than applying suggestions passively, students acted as active evaluators of AI advice, accepting clear-cut corrections readily but pausing to deliberate over style and cultural suggestions.

This pattern replicates Lau et al.'s (2025) observation that learners frequently re-evaluate ChatGPT suggestions and consult external resources before deciding and corroborates the broader argument that AI feedback engagement involves active metacognitive processing. A nuance that prior work has not fully addressed, however, emerged clearly in our data: when AI and instructor annotations aligned, students proceeded with confidence; when they diverged, they invested substantially more deliberative effort. The dual-source structure of the feedback sheet thus did more than signal reliability. It actively organized the revision process by indicating which suggestions warranted deeper scrutiny. Students' reports of noticing recurring error patterns (e.g., consistent difficulties with passive constructions) suggest that this engagement extended beyond the immediate task toward durable skill development, though whether such awareness translates into sustained behavioral change across subsequent tasks remains an open question that longitudinal designs would be better positioned to address.

The trust dimension of these findings deserves particular attention, as it both supports and complicates existing accounts. Students' confidence in applying AI suggestions was closely tied to the visible presence of instructor endorsement, directly supporting Cheng et al. 's (2023) finding that explicit provider information significantly increases trust in LLM-generated feedback. The color-coded annotation system operationalized this transparency in a practical, low-infrastructure format, offering a replicable classroom mechanism for a principle that prior research had identified theoretically but not yet implemented procedurally. At the same time, trust was not uniformly high: when AI and instructor comments diverged, some students reported uncertainty and tended toward conservative, surface-level choices rather than engaging deeply with the conflict. This points to a gap in feedback literacy that the hybrid model alone cannot resolve and suggests that building on Carless & Boud's (2018) framework, instructors should explicitly teach students how to interpret and reconcile competing feedback sources rather than assuming this capacity develops spontaneously. These trust dynamics, shaped in part by students' unfamiliarity with AI feedback tools, may differ in populations with greater prior experience of AI-assisted workflows, and future research should examine whether the instructor endorsement effect persists as AI feedback becomes more normalized in translator education.

From the teachers' perspective, the model produced a qualitative shift in professional role that carries its own implications. Instructors self-reported modifying or supplementing approximately 20% of AI-generated comments, primarily in areas of cultural context and idiomatic register, consistent with evidence that AI can absorb routine correction workload in language courses (Ranalli, 2023). More significantly, the instructors did not report feeling displaced; instead, they described their role as shifting from error-detector to pedagogical curator. This finding directly challenges the concern, raised in AI-in-education critiques (e.g., Selwyn, 2022), that automated feedback tools risk marginalizing teacher expertise. When the model is structured to require instructor validation, teacher judgment becomes more visible and more targeted rather than less relevant. That said, all four instructors acknowledged a learning curve in prompt design and output evaluation, a practical reminder that effective implementation depends on adequate professional development rather than ad hoc adoption. Whether this learning curve diminishes with sustained use, and how instructor annotation behavior may differ in graduate-level or professionally oriented translation programs, are questions that the present sample cannot address.

Taken together, these findings contribute to a growing body of evidence that effective AI integration in language education is a matter of complementarity rather than substitution (Tseng et al., 2023). The present study's specific contribution is to demonstrate that this complementarity can be operationalized through a replicable, classroom-ready procedure, structured color-coded annotation, that preserves pedagogical control without requiring specialized technical infrastructure. The human-in-the-loop feedback framework explored here positions ChatGPT as an intelligent assistant responsible for systematic error detection, while the instructor curates, contextualizes, and models evaluative judgment (Inkpen et al., 2023; Guo et al., 2024; X. Xu et al., 2025). For the undergraduate Chinese-to-English translation students in this study, this approach yielded rich, transparent, and actionable guidance; for teachers, it offered meaningful workload relief without sacrificing the professional authority that defines their pedagogical role. Future work should explore how instructor intervention level can be progressively reduced as students' feedback literacy develops, and whether the framework generalizes across language pairs and text genres beyond the cultural heritage domain examined here.

6. Implications

The findings reported here carry practical implications for how translation programs design feedback environments, prepare instructors, and orient future inquiry. Most immediately, the evidence that instructor-validated AI feedback was associated with substantial gains in cultural adaptation and style, dimensions that purely automated feedback studies have found more resistant to improvement, has direct curricular relevance (Wilson & Czik, 2016; Cao & Zhou, 2025). Translation courses should not simply introduce AI tools as efficiency aids; rather, they should build structured opportunities for students to encounter, evaluate, and reconcile AI and instructor perspectives as a pedagogical goal. In practical terms, this means designing early-stage assignments around the hybrid feedback format introduced here, and progressively reducing instructor annotation density as students' capacity for independent feedback evaluation grows. The meta-cognitive gains observed in this study, students' growing awareness of

recurring error patterns, suggest that such a staged approach could yield cumulative benefits across a program rather than isolated task-level improvements.

The trust and transparency dynamics observed in this study further suggest that feedback system design warrants careful attention. Students' confidence in applying AI suggestions was contingent on the visible presence of instructor endorsement; when this signal was absent or ambiguous, engagement became more conservative and less productive (Guo & Wang, 2024). This implies that any institutional implementation of AI feedback tools should prioritize making feedback provenance explicit, clearly distinguishing AI-generated from instructor-validated commentary, rather than presenting hybrid output as a seamless whole. Software interfaces that allow instructors to annotate, endorse, or correct AI suggestions in a single workflow, and that surface this editorial layer to students, would directly support the kind of critical engagement documented here. The color-coded annotation method used in this study offers a low-technology model for this principle that requires no specialized infrastructure.

A parallel implication concerns instructor preparation. The learning curve reported by participating teachers, particularly around prompt design and the identification of AI misreadings of cultural references, signals that effective implementation of human-in-the-loop feedback is not self-evident (Kinder et al., 2025). Professional development programs should address prompt engineering for translation-specific feedback tasks, strategies for efficiently reviewing AI output for cultural and contextual reliability, and approaches to modeling feedback literacy explicitly in the classroom. The instructors in this study described their role as shifting from error-detector to pedagogical curator; preparing teachers for this reorientation should be treated as a substantive training need rather than an incidental adjustment.

7. Discussion Limitations and Future Directions

Several limitations of the present study warrant acknowledgment. First, the single-task, cross-sectional design constrains the conclusions that can be drawn about learning over time. The pre-to-post gains documented here reflect improvement within one feedback cycle; whether repeated exposure to hybrid feedback produces cumulative development in translation competence or feedback literacy remains an open question. A single session is also insufficient to determine whether the metacognitive gains suggested by the think-aloud data, students' awareness of recurring error patterns, translate into sustained behavioral change in subsequent translation tasks. Furthermore, the pre-to-post design introduces potential confounds that the current study cannot fully disentangle. Because students revised the same text they had originally translated, score gains may partially reflect increased familiarity with the source text rather than feedback-driven learning per se. Repeated engagement with a 1,500-word passage affords students additional exposure to its lexical and structural demands, and the act of revision itself constitutes a form of practice that may independently contribute to quality improvement. These text familiarity and practice effects represent plausible alternative explanations for a portion of the observed gains, and future designs using parallel texts for pre- and post-assessments would help isolate feedback effects more cleanly. It is also worth noting that the uniformity of gains, with no student's score declining, may partly reflect a ceiling effect specific to this task and text. Students who entered with higher baseline scores had less room for improvement on a 5-point scale, and the absence of any score decline could reflect the relatively bounded nature of a single revision cycle rather than a universally effective intervention. Future studies might consider wider scoring ranges or more complex tasks to better capture variance in revision outcomes.

Second, the absence of a control condition limits the causal attributions available from this dataset. Without AI-only or human-only comparison arms, it is not possible to precisely quantify how much of the observed improvement is attributable to the collaborative structure specifically, as opposed to the feedback intervention more generally. Future studies employing randomized or quasi-experimental designs with clearly delineated conditions would allow more fine-grained claims about the added value of instructor curation over and above AI feedback alone.

Third, the sample was relatively homogeneous: all participants were senior undergraduates from a single institution, sharing similar proficiency backgrounds and L1. The extent to which the framework generalizes across different learner profiles, including students with lower baseline proficiency, different L1 backgrounds, or experience with AI tools, remains to be established. Similarly, the translation task was confined to a single genre (cultural heritage text), which was chosen precisely for its demands on cultural adaptation and register. Whether the hybrid model produces comparable benefits for more technical genres, such as legal or scientific translation, is a question the current data cannot answer.

Finally, the instructor variable was not systematically controlled. The four participating teachers differed in their familiarity with AI tools and their approaches to annotation, and these differences were not formally analyzed. Individual variation in how instructors curate AI output may be a meaningful moderator of student outcomes, and future research should examine what constitutes effective instructor intervention within this framework, both in terms of quantity and type of annotation. Related to this, the figure reported in the findings, that instructors modified or supplemented approximately 20% of AI-generated comments, should be

interpreted with caution, as it was derived from participants' self-reported estimates rather than systematic tracking of annotation behavior. It is possible that actual modification rates varied meaningfully across the four instructors and across different sections of the source text, and that retrospective recall introduced inaccuracies. Future studies should incorporate objective logging of instructor annotation activity to yield more reliable and granular data on how human oversight is distributed across feedback dimensions and text types.

These limitations point toward a productive agenda for future inquiry. Longitudinal designs tracking student development across a full semester of hybrid feedback cycles would provide evidence on whether the framework supports progressive growth in both translation quality and independent feedback evaluation. Such designs might also incorporate parallel-text assessments to control for text familiarity effects. Comparative trials with AI-only and human-only conditions would sharpen the causal picture. Research examining individual difference variables, including learner attitudes toward AI, prior feedback experience, and developing feedback literacy, would support the kind of adaptive implementation envisaged in the implications. Finally, extending the framework to other language pairs, text genres, and institutional contexts would test its generalizability and lay the groundwork for evidence-based guidelines on AI integration in translator training more broadly.

8. Conclusion

This study set out to examine whether a structured human-AI collaborative feedback model could improve translation quality and support meaningful learning in undergraduate translator training. The results suggest that it can and offer preliminary insight into the processes through which this may occur. The significant pre-to-post gains across all MQM dimensions, including the more resistant dimensions of style and cultural adaptation, are consistent with a feedback process that may exceed what either AI or instructor feedback achieves in isolation, however, the absence of a control condition means that causal claims about the collaborative structure specifically must remain tentative pending controlled replication. The complementarity was not incidental: it was the structural requirement for instructor validation that appeared to convert AI efficiency into pedagogical depth, prompting students to engage critically with competing sources of guidance rather than defaulting to either uncritically.

What the data suggest is that the value of the hybrid model lies less in the technology itself than in the epistemic situation it creates for learners. Students in this study did not simply receive better feedback; they were positioned to evaluate, compare, and negotiate between two sources of commentary with different strengths and limitations. This active arbitration, evidenced consistently in the think-aloud protocols, was associated with both immediate quality improvements and the kind of reflective awareness of recurring error patterns that may support durable skill development. The transparency of the feedback format, which made the origin and reliability of each comment visible, appeared to be a necessary condition for this engagement: without it, the collaborative structure would have been invisible to the learner.

For translation pedagogy more broadly, these findings point toward a reframing of how AI integration should be conceptualized. The human-in-the-loop framework explored here is not primarily a workload management strategy, though instructors did report meaningful relief from routine correction tasks. It is, more fundamentally, a candidate pedagogical design that aims to preserve the conditions under which deep learning occurs, conditions that require human judgment to remain visible, authoritative, and instructionally purposeful. As AI tools become more capable and more pervasive in language education, the question facing educators is not whether to use them, but how to structure their use so that they amplify rather than circumvent the relational and reflective dimensions of learning. This study offers one empirically informed starting point for that question, and points toward a broader research agenda examining how the balance of human and AI input can be adapted as learner competence and feedback literacy develop over time.

Funding: This research was funded by Zhejiang Shuren University, grant number 2026SK004.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD: 0000-0003-1170-1948

References

[1] Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

[2] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[3] Cao, S., & Zhou, T. (2025). Exploring the Efficacy of ChatGPT-Based Feedback Compared With Teacher Feedback and Self-Feedback: Evidence From Chinese-English Translation. *SAGE Open*, *15*(3), 21582440251369204. https://doi.org/10.1177/21582440251369204

[4] Carless, D. (2012). Trust and its role in facilitating dialogic feedback. In *Feedback in Higher and Professional Education*. Routledge.

[5] Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

[6] Chauhan, S., & Daniel, P. (2023). A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics. *Neural Processing Letters*, *55*(9), 12663–12717. https://doi.org/10.1007/s11063-022-10835-4

[7] Chen, J., Zhang, L. J., Wang, X., & Zhang, T. (2021). Corrigendum: Impacts of Self-Regulated Strategy Development-Based Revision Instruction on English-as-a-Foreign-Language Students' Self-Efficacy for Text Revision: A Mixed-Methods Study. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.747252

[8] Chen, S., & Zhou, T. (2024). Culturally based semantic losses in Lonely Planet's travel guides translations for Beijing, Shanghai, and Sichuan. *Frontiers in Communication*, *9*. https://doi.org/10.3389/fcomm.2024.1343784

[9] Chen, S., & Zhou, T. (2026). Prompt-induced cultural mediation and its limits: A micro-level analysis of LLM translation of Chinese tourism texts. *Cogent Arts & Humanities*, *13*(1), 2631304. https://doi.org/10.1080/23311983.2026.2631304

[10] Cheng, L., Li, Y., Su, Y., & Gao, L. (2023). Effect of regulation scripts for dialogic peer assessment on feedback quality, critical thinking and climate of trust. *Assessment & Evaluation in Higher Education*, *48*(4), 451–463. https://doi.org/10.1080/02602938.2022.2092068

[11] Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325. https://ieeexplore.ieee.org/abstract/document/10260740/

[12] Derakhshan, A., & Taghizadeh, M. S. (n.d.). *Does Artificial Intelligence (AI) Nurture or Hinder Language Learners' Higher- Order Thinking Skills (HOTS)? A Phenomenological Study on L2 Learners' Perspectives and Lived Experiences*. https://doi.org/10.1111/ijal.12824

[13] Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., & Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI -generated feedback. *British Journal of Educational Technology*, *56*(3), 1074–1091. https://doi.org/10.1111/bjet.13558

[14] Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

[15] Group, P., Hurtado Albir (principal investigator), A., Galán-Mañas, A., Kuznik, A., Olalla-Soler, C., Rodríguez-Inés, P., & Romero (research team, in alphabetical order), Lupe. (2018). Competence levels in translation: Working towards a European framework. *The Interpreter and Translator Trainer*, *12*(2), 111–131. https://doi.org/10.1080/1750399X.2018.1466093

[16] Guo, K., Chen, X., & Qiao, S. (2024). Exploring a Collaborative Approach to Peer Feedback in EFL Writing: How Do Students Participate? *RELC Journal*, *55*(3), 658–672. https://doi.org/10.1177/00336882221143192

[17] Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

[18] Han, C., & Lu, X. (2023). Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? *Computer Assisted Language Learning*, *36*(5–6), 1064–1087. https://doi.org/10.1080/09588221.2021.1968915

[19] Han, Y., & Xu, Y. (2021). Student Feedback Literacy and Engagement with Feedback: A Case Study of Chinese Undergraduate Students. *Teaching in Higher Education*, *26*(2), 181–196. https://doi.org/10.1080/13562517.2019.1648410

[20] Holstein, K., Aleven, V., & Rummel, N. (2020). *A Conceptual Framework for Human–AI Hybrid Adaptivity in Education* (pp. 240–254). https://doi.org/10.1007/978-3-030-52237-7_20

[21] Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L. H., & Quinn, G. (2023). Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Transactions on Computer-Human Interaction*, *30*(5), 1–29. https://doi.org/10.1145/3534561

[22] Jiao, H., Hu, W., & Zhang, X. (2025). *To eat or to feed: Can large language models provide useful feedback in translation education?*

[23] Kim, H. R., & Bowles, M. (2019). How Deeply Do Second Language Learners Process Written Corrective Feedback? Insights Gained From Think-Alouds. *TESOL Quarterly*, *53*(4), 913–938. https://doi.org/10.1002/tesq.522

[24] Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, *8*, 100349. https://doi.org/10.1016/j.caeai.2024.100349

[25] Kiraly, D. (Ed.). (2015). *Towards Authentic Experiential Learning in Translator Education* (1st ed.). V&R Unipress. https://doi.org/10.14220/9783737004954

[26] 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*(2), 131–148.

[27] Kumar, S., Datta, S., Singh, V., Datta, D., Kumar Singh, S., & Sharma, R. (2024). Applications, Challenges, and Future Directions of Human-in-the-Loop Learning. *IEEE Access*, *12*, 75735–75760. https://doi.org/10.1109/ACCESS.2024.3401547

[28] Lau, G. R., Low, W. Y., Tay, L., Guevarra, Y., Gašević, D., & Hartanto, A. (2025). *Understanding Critical Thinking in Generative Artificial Intelligence Use: Development, Validation, and Correlates of the Critical Thinking in AI Use Scale* (arXiv:2512.12413). arXiv. https://doi.org/10.48550/arXiv.2512.12413

[29] Li, M., Yu, S., Mak, P., & Liu, C. (2023). Exploring the efficacy of peer assessment in university translation classrooms. *The Interpreter and Translator Trainer*, *17*(4), 585–609. https://doi.org/10.1080/1750399X.2023.2236920

[30] Lin, Z., Song, X., Guo, J., & Wang, F. (2021). Peer Feedback in Translation Training: A Quasi-Experiment in an Advanced Chinese–English Translation Course. *Frontiers in Psychology*, *12*, 631898. https://doi.org/10.3389/fpsyg.2021.631898

[31] Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica Tecnologies de La Traducció*, (12), 455–463. https://doi.org/10.5565/rev/tradumatica.77

[32] Ma, H., Ismail, L., & Han, W. (2024). A bibliometric analysis of artificial intelligence in language teaching and learning (1990–2023): Evolution, trends and future directions. *Education and Information Technologies*, *29*(18), 25211–25235. https://doi.org/10.1007/s10639-024-12848-z

[33] Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

[34] Mariana, V., Cox, T., & Melby, A. (2015). The Multidimensional Quality Metrics (MQM) Framework: A new framework for translation quality assessment. *The Journal of Specialised Translation*, 137–161. https://doi.org/10.26034/cm.jostrans.2015.343

[35] Mellinger, C. D. (2019). Metacognition and self-assessment in specialized translation education: Task awareness and metacognitive bundling. *Perspectives*, *27*(4), 604–621. https://doi.org/10.1080/0907676X.2019.1566390

[36] Memarian, B., & Doleck, T. (2024). Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100053. https://doi.org/10.1016/j.chbah.2024.100053

[37] Mohammed, T. A. S. (2025). Evaluating Translation Quality: A Qualitative and Quantitative Assessment of Machine and LLM-Driven Arabic–English Translations. *Information*, *16*(6), 440. https://doi.org/10.3390/info16060440

[38] Molenaar, I. (2022). Towards Hybrid Human-AI Learning Technologies. *European Journal of Education*, *57*(4), 632–645. https://doi.org/10.1111/ejed.12527

[39] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

[40] Neunzig, W., & Tanqueiro, H. (2005). Teacher Feedback in Online Education for Trainee Translators. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *50*(4). https://doi.org/10.7202/019873ar

[41] Orak, S. D. (2025). Turkish EFL teachers' perspectives on AI-generated feedback: Negotiating trust, control, and pedagogical adaptation in writing instruction. *Applied Linguistics: Research, Measurement and Practice*, *1*(1), 74–94. https://doi.org/10.65334/n40kdg46

[42] Sato, M., & Loewen, S. (2018). Metacognitive Instruction Enhances the Effectiveness of Corrective Feedback: Variable Effects of Feedback Types and Linguistic Targets. *Language Learning*, *68*(2), 507–545. https://doi.org/10.1111/lang.12283

[43] Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

[44] Tabari, M. A., Sato, M., & Wang, Y. (2023). Engagement with written corrective feedback: Examination of feedback types and think-aloud protocol as pedagogical interventions. *Language Teaching Research*, 13621688231202574. https://doi.org/10.1177/13621688231202574

[45] Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, *51*, 101440.

[46] Washbourne, K. (2014). Beyond error marking: Written corrective feedback for a dialogic pedagogy in translator training. *The Interpreter and Translator Trainer*, *8*(2), 240–256. https://doi.org/10.1080/1750399X.2014.908554

[47] Wiboolyasarin, W., Wiboolyasarin, K., Suwanwihok, K., Jinowat, N., & Muenjanchoey, R. (2024). Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers and Education: Artificial Intelligence*, *6*, 100228. https://doi.org/10.1016/j.caeai.2024.100228

[48] Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

[49] Winstone, N. E., Mathlin, G., & Nash, R. A. (2019). Building Feedback Literacy: Students' Perceptions of the Developing Engagement With Feedback Toolkit. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00039
[50] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381.
[51] Xu, S., Su, Y., & Liu, K. (2025a). Investigating student engagement with AI-driven feedback in translation revision: A mixed-methods study. *Education and Information Technologies*, *30*(12), 16969–16995. https://doi.org/10.1007/s10639-025-13457-0
[52] Xu, S., Su, Y., & Liu, K. (2025b). Investigating student engagement with AI-driven feedback in translation revision: A mixed-methods study. *Education and Information Technologies*. https://doi.org/10.1007/s10639-025-13457-0
[53] Xu, X., Sun, F., & Hu, W. (2025). Integrating human expertise with GenAI: Insights into a collaborative feedback approach in translation education. *System*, *129*, 103600. https://doi.org/10.1016/j.system.2025.103600
[54] Yu, S., Zhang, Y., Zheng, Y., & Lin, Z. (2020). Written Corrective Feedback Strategies in English-Chinese Translation Classrooms. *The Asia-Pacific Education Researcher*, *29*(2), 101–111. https://doi.org/10.1007/s40299-019-00456-2
[55] Zheng, Y., Zhong, Q., Yu, S., & Li, X. (2020). Examining Students' Responses to Teacher Translation Feedback: Insights From the Perspective of Student Engagement. *Sage Open*, *10*(2), 2158244020932536. https://doi.org/10.1177/2158244020932536

Appendix A

Cultural Heritage Text

长城是中华民族的精神象征，是我国现存体量最大、分布最广的文化遗产。长城修筑的历史可上溯到西周时期，秦灭六国统一天下后，秦始皇连接和修缮战国长城，始有万里长城之称。明代是最后一个大规模修筑长城的朝代。按照中国国家文物局《长城保护报告》公布的数据：我国各时代长城资源分布于北京、天津、河北、山西、内蒙古、辽宁、吉林、黑龙江、山东、河南、陕西、甘肃、青海、宁夏、新疆 15 个省（自治区、直辖市）404 个县（市、区）。各类长城资源遗存总数 43721 处（座/段），其中墙体 10051 段，壕堑/界壕 1764 段，单体建筑 29510 座，关、堡 2211 座，其他遗存 185 处。墙壕遗存总长度 21196.18 千米。北京市范围内的长城始建于北齐，明代大规模修筑，东起平谷区，经密云区、怀柔区、延庆区、昌平区，西至门头沟区，长城墙体 461 段、全长 520.77 千米，关堡 147 座，单体建筑 1742 座，相关设施 6 处，核定公布为全国重点文物保护单位的遗存有 2328 处。长城于 1987 年 12 月列入世界文化遗产。

北京故宫，旧称为紫禁城，位于北京中轴线的中心，是中国明、清两代 24 位皇帝的皇家宫殿，是中国古代汉族宫廷建筑之精华，无与伦比的建筑杰作，也是世界上现存规模最大、保存最为完整的木质结构的古建筑群之一。北京故宫由明成祖朱棣于永乐四年（公元 1406 年）开始建设，到明代永乐十八年（公元 1420 年）建成，占地面积约为 72 万平方米，建筑面积约为 15 万平方米，宫殿建筑均是木结构、黄琉璃瓦顶、青白石底座。被誉为世界五大宫之首（北京故宫、法国凡尔赛宫、英国白金汉宫、美国白宫和俄罗斯克里姆林宫）。北京故宫是全国重点文物保护单位，于 1987 年 12 月列入世界文化遗产。

颐和园位于北京市西北郊，原名清漪园，始建于 1750 年，在 1860 年第二次鸦片战争中被英法联军烧毁，1886 年使用海军军费等款项重修，并于两年后改名颐和园。颐和园以万寿山、昆明湖构成其基本框架，占地 293 公顷，水面约占四分之三，园中有点景建筑物百余座、大小院落 20 余处，古建筑 3000 余间，面积 70000 多平方米，还有古树名木 1600 余株。其中佛香阁、长廊、石舫、苏州街、十七孔桥、谐趣园、大戏台等都是家喻户晓的代表性建筑。颐和园是全国重点文物保护单位，于 1998 年 11 月列入世界文化遗产。

天坛位于北京市东城区永定门内大街东侧，是明、清两代皇帝祭祀皇天上帝的场所，始建于明永乐十八年（1420 年），后经不断改扩建，至清乾隆年间建成。天坛占地 273 公顷，分为内外两坛，内坛由圜丘、祈谷坛两部分组成，外坛为林区、广植树木，有古松柏 3500 余株。天坛代表性建筑主要有祈年殿、圜丘、皇穹宇、斋宫、神乐署、牺牲所等。天坛是全国重点文物保护单位，于 1998 年 11 月列入世界文化遗产。

明十三陵位于北京市昌平区境内天寿山南麓，是明朝（1368 年-1644 年）十三位皇帝的陵寝建筑群，规模宏大、体系完备、保存较为完整。陵区面积约 120 余平方公里，有长陵、献陵、景陵、裕陵、茂陵、泰陵、康陵、永陵、昭陵、定陵、庆陵、德陵、思陵和相关陪葬陵及神路等附属设施。明十三陵是全国重点文物保护单位，于 2003 年 7 月列入世界文化遗产。

北京中轴线位于北京老城中心，纵贯老城南北，是统领整个老城规划格局的建筑与遗址的组合体，北端钟鼓楼，向南经万宁桥、景山，过故宫、端门、天安门、外金水桥、天安门广场及建筑群、正阳门、中轴线南段道路遗存，至南端永定门，太庙和社稷坛、天坛和先农坛分列东西两侧。北京中轴线始建于 13 世纪，形成于 16 世纪，此后不断完善，历经逾 7 个世纪，形成了由古代皇家宫苑建筑、古代皇家祭祀建筑、古代城市管理设施、国家礼仪和公共建筑、居中道路遗存共同构成的秩序井然、气势恢宏

的城市建筑群。北京中轴线规模宏大、规划格局均衡对称、城市景观井然有序，是中国传统都城中轴线发展至成熟阶段的杰出范例，也是中国现存最为完整的传统都城中轴线建筑群。

Appendix B

Learner Perception Questionnaire: Human-AI Collaborative Feedback

Please rate each statement on a scale of 1 to 5: 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree

Dimension 1: Clarity

1. The feedback I received was clearly written and easy to understand.
2. I could tell what specific problems each piece of feedback was pointing to.
3. The distinction between AI-generated and instructor-revised comments was clear to me.
4. The feedback helped me understand *why* my original choices were problematic, not just *that* they were problematic.
5. I knew what I was expected to do after reading the feedback.

Dimension 2: Trustworthiness

6. I felt confident that the feedback I received was accurate.
7. Knowing that an instructor had reviewed the AI-generated comments increased my trust in the feedback.
8. I was comfortable acting on the feedback without seeking additional verification.
9. The feedback reflected a reliable understanding of the source text and its cultural context.
10. I trusted the feedback even when it suggested changes I had not initially considered.

Dimension 3: Usefulness

11. The feedback guided me toward concrete and actionable improvements in my translation.
12. Following the feedback helped me produce a better translation than I could have achieved on my own.
13. The feedback addressed the aspects of my translation that most needed improvement.
14. The combination of AI and instructor comments gave me more useful guidance than either source alone would have.
15. The feedback was specific enough to apply directly during revision.

Dimension 4: Pedagogical Value

16. Engaging with this feedback made me more aware of my recurring translation errors.
17. The feedback process encouraged me to think more deeply about translation decisions, not just surface corrections.
18. I feel more confident in evaluating translation quality after going through this feedback process.
19. This experience has changed how I will approach feedback, whether from AI or instructors, in the future.
20. I believe this type of feedback supports my long-term development as a translator, not just task-level improvement.

Scoring note: Each dimension yields a subscale mean (1-5). An overall perception score can be computed as the mean of all 20 items. Item 3 is specific to the hybrid format and may be analyzed separately as a transparency indicator if needed.