CHORUS: Designing Human–AI Multi-Agent Collaboration for Professional Translators

George Xi Wang*
xw3617@nyu.edu
New York University
Brooklyn, New York, United States
Stony Brook University
Stony Brook, New York, United States

Jiaqian Hu*
jiaqianh@middlebury.edu
Translation and Localization Management
Middlebury Institute of International Studies at Monterey
Monterey, California, United States

Guande Wu
guandewu@nyu.edu
Tandon School of Engineering, New York University
New York City, New York, United States

Jing Qian
jqian1590@tongji.edu.cn
College of Electronic and Information Engineering
Tongji University
Shanghai, China

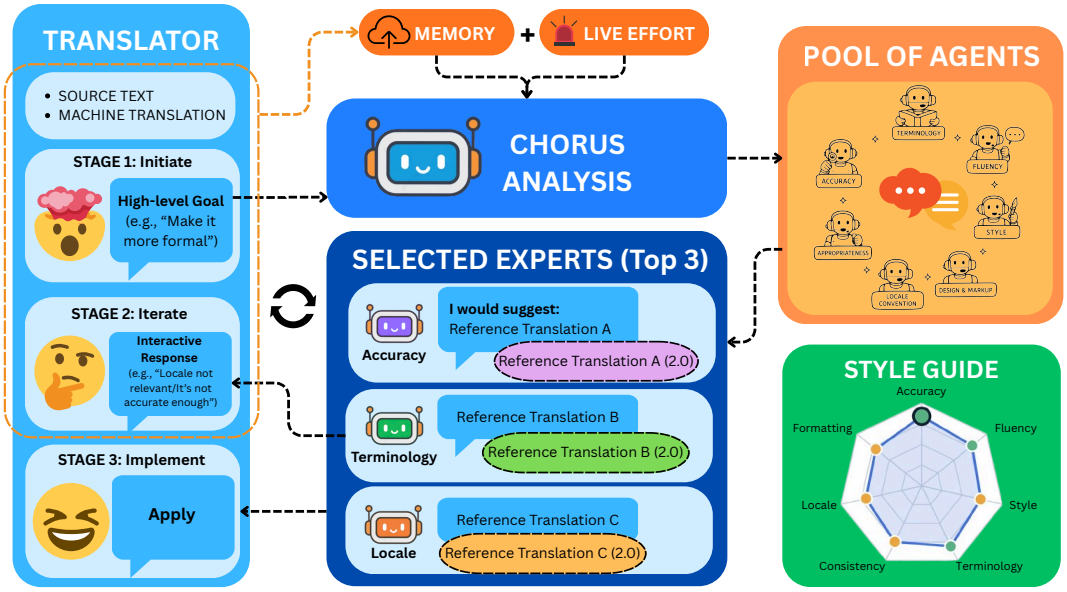

Figure 1: CHORUS is a multi-agent translation system for professional translators. With seven MQM-inspired agents, it provides concurrent editing suggestions while adapting to translators' editing habits from multi-modal inputs and editing histories.

Abstract

Despite the widespread use of automatic AI translation systems in daily language tasks, their limitations become apparent in professional translation contexts, where human expertise remains crucial. Professionals rarely rely on these systems in their practice due to a lack of detailed support for the translation process, matching professional styles, and accountability for the final outcome. To bridge this gap, we present CHORUS, a mixed-initiative system designed to support translators' workflow while preserving their personal style. The system adapts through behavioral interaction signals, such as keystrokes, pauses, and editing effort. A formative study found that incorporating MQM theory may be beneficial for professional translation, and the system should adapt to each individual translator's idiosyncratic traits. The final within-subject study with 30 licensed English–Chinese translators found that our system reduced completion time by 33.8%, lowered translators' cognitive effort, and improved final translation quality measured by the automatic metrics BLEU and COMET. Participants also reported that the system made translation issues easier to inspect, reduced repeated prompting compared to a chat-interface LLM, and offered reflections on their habits and traits. Our findings illustrate

*These authors contributed equally to this work.

ICMI '26, Napoli, Italy

ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831126

how multi-agent AI systems can be designed to support expert workflows and their potential for professional use.

CCS Concepts

• **Human-centered computing → Interaction design**; • **Computing methodologies → Multi-agent systems**.

Keywords

translation, multi-agent, Multidimensional Quality Metrics

ACM Reference Format:

George Xi Wang, Jiaqian Hu, Guande Wu, and Jing Qian. 2026. CHORUS: Designing Human–AI Multi-Agent Collaboration for Professional Translators. In *Proceedings of INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3776574.3831126

1 Introduction

Professional translation is a crucial part of high-stakes scenarios, such as medical, legal, and business communications, where errors are much less tolerated, and quality and style are more prominent than in daily translations. Large Language Models (LLMs) have become increasingly popular, but they often lack accurate meaning preservation, cultural nuance, domain specificity, and personal style [7, 28, 49, 60, 79, 82] for professionals. Moreover, professional translation needs to meet specific requirements imposed by clients, and such requirements are rarely available in public domains for LLMs to learn. Translators often need to discuss and resolve nuanced details with clients of different cultural backgrounds, as they are solely responsible for the final outcomes. This responsibility is particularly critical in high-stakes domains where human assurance and accountability are still essential. Plus, most existing systems use a single LLM interface, which often blends accuracy, terminology, and style into a one-shot translation. This makes it difficult for translators to revise any of these dimensions individually, resulting in additional time during post-editing. Consequently, existing systems fail to provide the in-progress granularity needed to support professional translation, and guidance in the literature on how to build such systems is largely missing. As a result, we ask: how can we harness the power of LLMs to design an efficient tool to support professional translation work?

Through a formative evaluation with 6 professional translators, we gather design insights around how to better scaffold professional work by incorporating Multidimensional Quality Metrics (MQM), a framework widely used to describe translation quality across dimensions such as accuracy, terminology, fluency, and style [16, 37, 40]. Rather than treating MQM as an evaluation taxonomy, we use it to structure support during revision and enable translators to inspect, adjust, and review each dimension.

We present CHORUS, a mixed-initiative translation system that adopts an MQM-inspired multi-agent design. We develop a live effort algorithm that uses multi-modal input and editing history to adjust agents' focus during the working process. A Live Style Guide summarizes revision patterns and provides personalized feedback for idiosyncratic traits. This design goal raises three research questions: **RQ1:** whether CHORUS reduces translators' editing effort and perceived workload; **RQ2:** whether it improves final translation quality; and **RQ3:** how CHORUS supports professional translators in revision, reflection, and preference formation.

We evaluated CHORUS in a within-subject study with 30 licensed English–Chinese translators using WMT24 translation tasks [11]. Compared with a chat-interface LLM baseline (GPT-5.3 web), CHORUS reduced completion time by 33.8%, significantly lowered cognitive workload, and improved translation quality. Participants also found that CHORUS made error inspection clear, which was useful for self-improvement and assessment. This work contributes: (1) formative results on why current LLM translation tools fall short for professionals and insights for improvement; (2) CHORUS, an MQM-inspired multi-agent system that dynamically translates and adapts through multi-modal behavioral signals while scaffolding the translation process and style; and (3) empirical evidence that a multi-agent translation system reduces effort, improves translation quality and speed, and supports more accountable revision. The CHORUS system, including all agent prompts, is publicly available at https://github.com/iamgeorge/CHORUS-OSS.

2 Related Work

2.1 LLM-Based Translation and Post-Editing

LLM-based translation work studies prompting, adaptation, and post-editing strategies, including zero-shot prompting, few-shot learning, fine-tuning, prompt design, demonstration quality, and example selection [44, 45, 56, 70, 84]. These methods improve benchmarks and affect document-level behavior, but translation quality remains context-sensitive, and perturbing demonstrations can substantially degrade output [56, 84]. Although LLMs can outperform supervised baselines in some directions, commercial accountability remains a concern [13, 87], motivating WMT evaluations with professional translators and human evaluation beyond reference-overlap metrics [11, 16, 17]. LLM-based post-editing can improve MT scores and perceived trustworthiness, and MQM-derived annotations can improve automatic quality metrics such as TER (Translation Edit Rate, number of edits needed to match a reference translation) [65], BLEU (n-gram overlap with reference translations) [50], and COMET (a neural metric trained on human quality judgments) [14, 29, 41, 58]. Yet hallucinated edits threaten high-stakes deployment [56, 75], and it remains unclear how annotations should guide concrete revision decisions [29]. This motivates human-in-the-loop workflows where translators retain oversight over quality and accountability [16, 72]. Iterative and multi-agent LLM work also shows both promise and limits: self-refinement and iterative translation refinement can improve outputs, fluency, and naturalness, but can complicate metric interpretation and amplify model self-bias unless external feedback is introduced [10, 42, 77]. Multi-agent systems use specialized roles for translation production or evaluation [6, 76, 81, 83], and interactive MT emphasizes translator control [15, 35]; however, prior systems often focus on agent specialization, agent-to-agent refinement, or traditional interactive MT rather than close human coordination during multi-dimensional revision [1, 26, 31]. CHORUS builds on this space by combining LLM revision, specialized agent support, and translator oversight.

Table 1: Participant profile for the formative study.

Expert	Side	Domain	Experience	Gender
E1	Client	Game Localization	12+	Female
E2	Client	Marketing	6+	Male
E3	Client	Government	8+	Female
E4	Vendor	Medical	10+	Female
E5	Vendor	Game Localization	7+	Male
E6	Vendor	Chip Design	11+	Female

2.2 Quality Evaluation in Translation

Translation quality is assessed through automatic metrics and human evaluation, but evaluation can be misleading without explicit error analysis, especially for strong MT systems where differences are subtle [22, 36, 43, 67, 85]. Prior work therefore argues for structured error types and severities [16, 37]. Automatic metrics include BLEU, which depends on reference similarity [8, 50, 59]; TER, which estimates revision effort [55, 65, 66]; and COMET, trained on human judgments such as Direct Assessment, HTER/TER-style judgments, and MQM annotations [4, 40, 58, 65]. Human evaluation protocols and quality frameworks include Direct Assessment, continuous scoring, the Dynamic Quality Framework, SAE J2450, the LISA QA Model, and Multidimensional Quality Metrics (MQM) [4, 9, 20, 21, 40, 61, 63]. Following prior work, we use MQM because it provides a fine-grained taxonomy of translation errors and is widely used in professional, research, operational quality-control, and agent-based evaluation settings [16, 18, 38, 39, 41]. In CHORUS, MQM defines the quality concerns available in the interface and the seven dimensions summarized in Table 2.

2.3 Adaptive Translation Systems

Adaptive translation systems treat translation as an interactive process: the system proposes translations, humans correct or rate them, and the system updates to reduce future effort [16, 48, 52, 74]. Mixed-initiative and interactive MT have long offered alternatives to pure post-editing through target-text-mediated interaction, completions compatible with translator input, links between human effort and machine learnability, and feedback datastores for later translations [15, 23, 24, 35, 72]. Recent adaptive work collects fine-grained edits, prefix constraints, accept/reject actions, segment-level post-edits, and structured error tags to reduce editing cost and make feedback reusable [12, 19, 29, 30, 34, 73, 78, 80]. LLM-enabled systems can produce useful edits but still require validation and human oversight [57]; iterative self-refinement can improve fluency and naturalness while complicating metric interpretation [10, 42, 54]. This points to a design gap for human-centered multi-agent translation: adaptation should respond not only to user feedback, but also to the effort and commitment behind that feedback.

3 Formative Study

3.1 Participants and Data Collection

We interviewed six professional translators (E1–E6), evenly split between client-side and vendor-side roles. Each participant had at least 6 years of professional practice across diverse translation domains, including game localization, marketing, government, medical translation, and chip design (Table 1). All interviews were conducted remotely via Zoom, and sessions were recorded and transcribed. Interviews averaged 42 minutes. The study was reviewed by our institution's IRB and granted Exempt status, and all participants provided informed consent prior to the interview.

3.2 Procedure

We conducted semi-structured interviews consisting of three parts: **Background and tool usage:** Participants were asked about the AI tools they use and their satisfaction with these tools (including reasons). **Think-aloud translation task:** Participants were given a sentence to translate in real time while verbalizing their thought process and explaining their edits. **Open-ended design reflection:** Participants were asked to imagine and describe their ideal translation tool while thinking aloud. The full interview protocol is provided in Appendix.

We analyzed the interview transcripts and notes using thematic analysis [5]. Two authors conducted open coding to identify recurring issues and breakdowns in current translation workflows. The authors then compared and discussed their codes, resolved disagreements through discussion, and grouped related codes into higher-level themes. This analysis led to three recurring challenges that informed the design of CHORUS.

3.3 Challenges in Current AI Translation Tools

C1. Single-LLM rewrites obscure distinct translation quality dimensions. Participants described professional revision as a process of balancing multiple quality dimensions within the same sentence, including accuracy, terminology, fluency, and style (E1, E2, E5). However, current AI tools often return a single broad rewrite, making it difficult for translators to see which quality dimension the system addressed and whether the revision introduced trade-offs elsewhere. As one participant explained, "It often rewrites everything at once. I cannot tell if it is fixing terminology or just changing the style, so I end up going back and rechecking" (E2). Across interviews, all six experts reported general-purpose AI chatbots (GPT web) as their primary day-to-day AI aid, while CAT tools were reserved for projects with established translation memories.

C2. Current AI tools do not carry forward translators' quality adjustments. Participants emphasized that translation decisions unfold through revision: a sentence may evolve through several versions, some segments may be repeatedly reconsidered, and translators may concentrate more attention on one quality concern than another (E1, E2). These process traces matter because not all edits carry the same weight. A terminology correction in medical content may reflect a domain constraint, while a fluency revision may matter more in consumer-facing text (E2, E6). However, current AI tools rarely preserve these signals across the workflow. Translators must repeatedly tell the system what to prioritize, as in E3's example: "This part has been revised three times. Don't change it, but you can still improve the surrounding sentences."

C3. Current AI tools provide little support for reviewing and justifying translation decisions. Finally, participants wanted more structured guidance for checking what had been

addressed during revision (E1, E5). They described needs such as reviewing style guides (E1), identifying critical issues (E1, E3), checking domain-specific constraints (E4), and understanding why a suggestion was made (E5). These needs were tied to professional accountability: translators often have to justify and defend decisions to supervisors, clients, or other stakeholders (E3, E4, E5). Current AI tools offer suggestions, but provide limited support for seeing which quality dimensions have been checked, which concerns may still need attention, and how a final decision can be explained.

3.4 Design Rationale

D1. Operationalize MQM as separable agent roles. Because LLM-based translation support is context-sensitive, a single-LLM interface can merge several quality goals into one opaque rewrite and become difficult for translators to inspect. All participants acknowledged MQM as an authoritative framework for multi-dimensional thinking. We therefore decompose revision support into MQM-aligned agents, each focused on one quality lens, so translators can inspect dimension-specific suggestions and coordinate trade-offs before deciding on the final translation.

D2. Capture and use editing history to reduce repeated MQM-specific quality adjustments. Professional translators often need to steer the balance among quality dimensions across many segments, such as preserving terminology, adjusting fluency, or maintaining a client-specific style. In current LLM workflows, these preferences often have to be restated through repeated prompts, which adds effort to the revision process. We should use the translator's interaction history to ease this work. We need to introduce a novel mechanism with confirmed edits and revision traces to help the system recognize which quality adjustments the translator has already made and carry them forward into later suggestions.

D3. Support structured review through revision history and quality coverage.

We should help translators review their own revision process without turning reflection into a separate task. The interface should surface useful traces of the work already done: recurring edits and preferences from revision history, how attention has been distributed across MQM dimensions, and which quality concerns may still deserve review. This gives translators material for checking coverage and explaining decisions while keeping final judgment with the user.

4 CHORUS System

4.1 Human-AI Multi-Agent Collaboration

4.1.1 Synchronization among multiple agents. Separating quality dimensions introduces a coordination problem: suggestions from multiple agents must remain anchored to the same current draft. CHORUS therefore synchronizes agent outputs after each user edit. Each agent's response is represented as a token-level difference against the latest draft using the Longest Common Subsequence algorithm [27, 33, 71]. CHORUS applies the minimal patch and immediately re-renders agent suggestions (see Fig. 2), keeping dimension-specific feedback aligned with the translator's current text.

4.1.2 Ranking agents. Since several dimensions may be relevant at once, CHORUS ranks agents to reduce attentional load while preserving access to the full MQM space. The system obtains an initial relevance score for each agent from the translator's high-level goal and current context. Subsequent interactions update each agent's score, so agents that better match the translator's current focus become more visible over time. This serves as an attention-management mechanism: the most relevant agents are ranked according to the current context while keeping other MQM dimensions always available.

4.1.3 Error handling. Hallucinations are known to cause errors in LLMs' responses [56, 75]. CHORUS employs a list of "bad examples" whenever the user identifies an error in LLM output. During regeneration, CHORUS commands the LLM to avoid similar mistakes cached in the bad example list. Additionally, users can regenerate any single agent's output if unsatisfactory.

4.2 Effort-Aware Memory

To help the system personalize toward users' habits, we introduce an effort-aware memory mechanism that converts users' editing behavior into weighted memory for prompt adaptation. Live Effort estimates which edits required more translator effort, Micro-Edits preserve what was changed and where the change occurred, and Memory uses these effort-weighted edit records to generate prompt guidance for the AI agents. Together, these components allow CHORUS to remember which edits matter the most and adjust future prompts accordingly.

Live Effort Following Krings and Stasimioti [32, 68], CHORUS models live effort at three levels: temporal, technical, and cognitive. *Temporal effort* is defined by two metrics: an initial pause that reflects the time spent reading and understanding the source text and a total edit duration that captures the time to refine. *Technical effort* is measured through keystroke counts (deletion and cursor movement). *Cognitive effort* captures the deeper reasoning processes, such as the number of redos and stress levels, which are inferred by ChatGPT-5.3 using appraisal theories of emotion (difficulty, ambiguity, risk, and controllability) [53, 62]. To combine these three metrics into one live effort score, we use a linear effort model [2, 68, 69] and join these metrics into the final score.

Micro-Edits capture deletions or replacements during editing. This design allows the system to preserve not only *what* the translator changed, but also *where* the change occurred. It can also be used to infer *why* the change was made and, when combined with the system's memory component, *how much effort* it required.

Memory CHORUS uses a weighted memory to reduce the need for repeated edits. Inspired by importance-aware retrieval [51, 86], CHORUS uses seven memory buckets to match up with seven AI agents' editing history. The memory contains a snapshot of micro-edits, target translation, and user prompt whenever a change is made. Non-agent edits such as modifying the output sentence directly are stored in a general memory bucket. To personalize CHORUS, we use the stored memory to adjust the prompt template to fit the professional translator's idiosyncratic traits. This begins with ranking micro-edits using the live effort algorithm. After ranking, CHORUS injects the top-five micro-edits as few-shot examples into a prompt template to generate a weighted memory, personalizing

Table 2: Seven MQM-aligned dimensions operationalized in CHORUS, adapted from MQM error typologies [37].

Dimension	Primary focus	What the expert checks
Accuracy	Meaning preservation	If the translation conveys source meaning without omissions, additions, or distortions.
Terminology	Lexical consistency	If domain-specific terms and glossary entries are used correctly and consistently.
Fluency	Linguistic well-formedness	If the sentence is grammatical, natural, and readable in the target language.
Style	Tone and register	If the wording matches the intended voice, formality, and conventions of the task.
Audience Appropriateness	Reader fit	If the translation is suitable for the intended audience and expertise level.
Locale Convention	Regional adaptation	If the translation follows culturally specific conventions, formats, and templates.
Design and Markup	Structural integrity	If formatting, line breaks, placeholders, and layout-sensitive elements are correct.

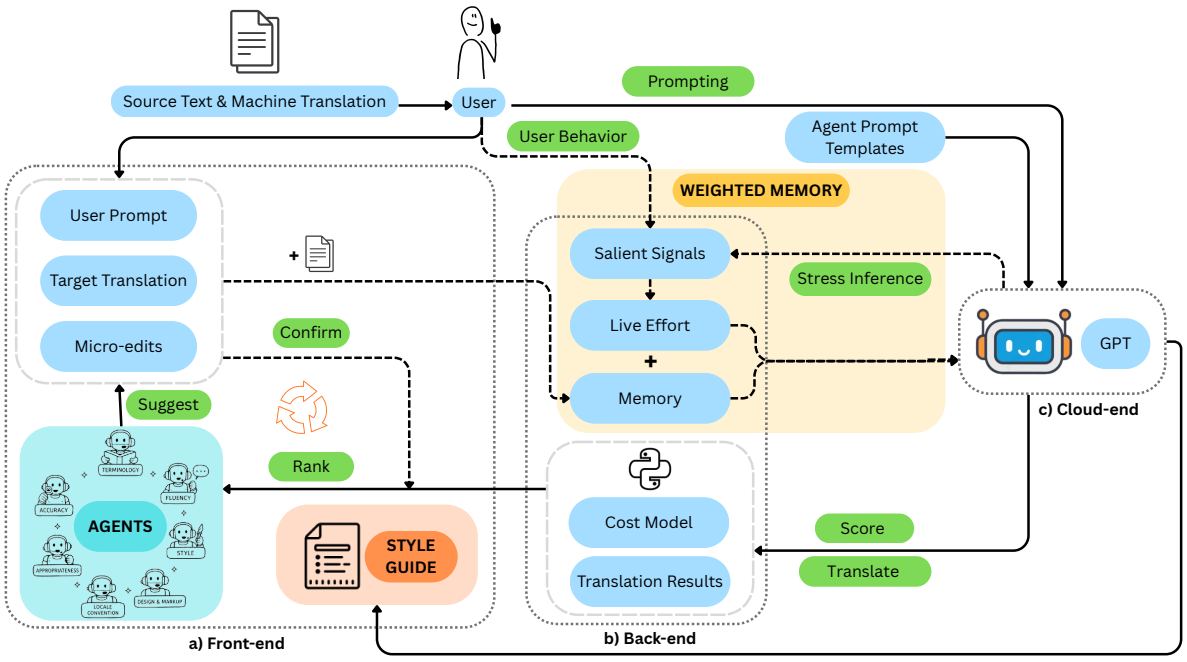

Figure 2: System flow chart of CHORUS workflow integrating context, draft, goals, agents, revisions, effort, and memory.

what matters most to users without restating their needs. Appendix shows a complete example and prompt templates.

4.3 Translational Scaffolding

We instrument *Live Style Guide* to show translational scaffolding by surfacing the learned preferences, recurring correction patterns, and dimension-level strengths as a spider graph using the weighted memory. This helps users to see which of the MQM dimensions are more frequently used and their personalized feedback on their performance. This feedback is created based on the LLM's response using the MQM website's suggestions [46]. This way, the system helps professional translators to recognize their blind spots over time and reflect on their strengths.

5 Evaluation

5.1 Experiment Design

The evaluation used English-to-Chinese sentence-level data from WMT24 [11]. We sampled 10 sentences from each of four domains: *literary*, *news*, *social*, and *speech*. WMT24 reference translations were retained for quality evaluation. In the **Baseline** condition, participants revised translations while using the GPT-5.3 web [47] as AI support. Chat history was cleared between trials. In the **CHORUS** condition, they worked with the CHORUS interface (see Figure 3), which contains the source text, an editable machine-translated draft, and access to seven AI agents. In both conditions, initial drafts were produced by GPT-5.3. Condition order was counterbalanced using a pre-generated table, and to avoid learning effects, no identical sentence appeared across conditions or trials.

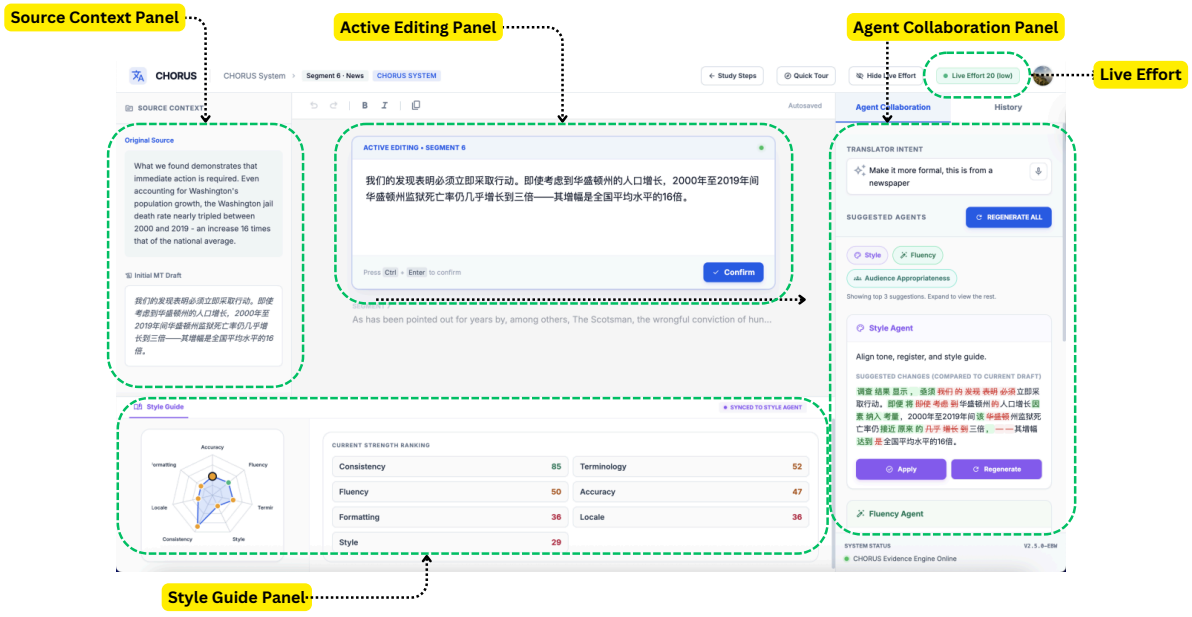

Figure 3: Overview of CHORUS. Source Context Panel (A) shows the source and initial translation. The Active Editing panel (B) is the main workspace for revision. The Agent Collaboration panel (C) provides suggestions from MQM agents. The Live Effort (D) indicates real-time translator effort. The Style Guide panel (E) provides learned preferences and user feedback.

CAT tools were not used because they rely on accumulated translation memories and termbases unavailable for our tasks. Other systems were also unsuitable as end-to-end baselines: CASMACAT [1] predates LLM-based assistance and was not evaluated on English–Chinese, TranSmart [26] is a closed commercial service, and MQM-APE and HiMATE [41, 83] provide evaluation methods rather than interactive translation systems.

5.2 Participants

We recruited 30 licensed professional English–Chinese translators who provided proof of certification. Participants were 21 to 50 years old (M=28.9, SD=6.1) and reported 3 to 21 years of translation experience (M=5.4, SD=4.9). They reported frequent use of AI translation tools (median = 4/5) and moderate-to-high familiarity with CAT tools (median = 3.5/5). Each participant was assigned two out of the four possible WMT24 domains and paired with conditions. For each domain, participants edited 10 sentences (10 trials), and each domain contained the two conditions. As a result, each participant conducted 10 sentences per condition, 20 per domain, and 40 in total. The overall frequency of domains was balanced across all participants. For example, we assigned P_n domains A and B and P_{n+1} domains C and D. This pattern was alternated.

5.3 Procedure and Measures

After participants signed the consent form, the experimenter explained the tasks and walked them through the interface and how to perform the trial. Participants had five minutes to practice. Once ready, they were assigned one of the conditions and switched to the other condition upon completing the first. The NASA-TLX survey [25] was used to assess cognitive load across conditions. At the end of the experiment, a semi-structured interview followed to collect qualitative feedback and their impressions of the two conditions. Sessions lasted approximately 100 minutes, and participants were compensated ¥100.

We collected interaction logs, outcome translations, and self-report data. Timed performance was computed from active editing duration, excluding idle periods. Perceived workload was measured with NASA-TLX on the original 0–100 scale, and open-ended responses were collected. To assess quality, we used COMET [58] and BLEU [50] automatic evaluation metrics as a means of automatic comparison as they come with "ground-truth" labeling. Further, we asked three professional translators with more than 6 years of experience to compare the translated sentence with the original.

6 Results

6.1 Efficiency, Effort, and Workload

A paired t-test on log-transformed completion time showed that CHORUS significantly reduced completion time compared with Baseline ($t(59) = -5.35$, $p < .001$). On the original scale, the geometric-mean time ratio of CHORUS over Baseline was 0.662, 95% CI [0.567, 0.772], indicating 33.8% faster task completion on average. Domain-level comparisons followed the same direction, with the largest reductions in literary and speech tasks (Fig. 4).

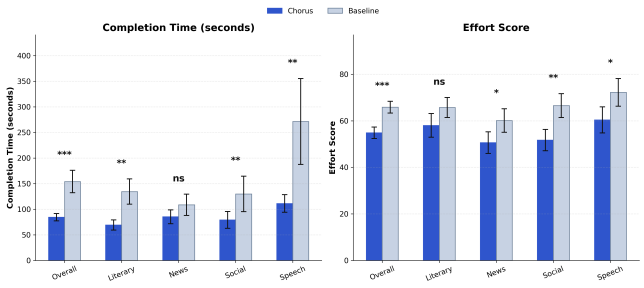

Figure 4: Completion time and live effort scores for CHORUS and Baseline across domains.

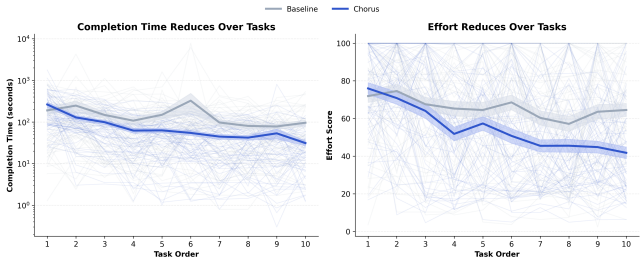

Figure 5: Task-order trajectories for completion time and live effort for two conditions. Completion time is log-scaled.

CHORUS also produced significantly lower Live Effort scores than Baseline ($t(59) = -5.81$, $p < .001$). The average effort score was 54.85 ($SD = 19.14$) for CHORUS and 65.84 ($SD = 19.61$) for Baseline, a reduction of about 11 points. Mixed-effects analyses found no significant omnibus domain effect for completion time, $\chi^2(3) = 5.45$, $p = .142$, or effort, $\chi^2(3) = 6.00$, $p = .112$, suggesting that the CHORUS advantage was not driven by a single domain.

Task-order analyses further suggested that participants settled into more efficient interaction patterns with CHORUS over time (Fig. 6). For completion time, the interaction between task order and CHORUS was significant, $b = -0.146$, $z = -6.66$, $p < .001$, with a significant mean slope difference against Baseline, -0.147, $t(59) = -6.12$, $p < .001$. Live Effort showed the same pattern: the CHORUS interaction was significant, $b = -2.444$, $z = -5.67$, $p < .001$, with a significant mean slope difference of -2.448, $t(59) = -5.14$, $p < .001$ (Fig. 5).

6.2 Translation Quality and Adaptivity

CHORUS improved final translation quality relative to Baseline. Measured against human-translated WMT references, mean BLEU increased from 34.90 under Baseline to 37.98 under CHORUS, a mean paired difference of 3.08 points ($t(29) = 3.16$, $p = .0036$; Wilcoxon $p = .0081$; Hedges' $g = 0.56$). Mean COMET increased from 0.837 to 0.852, a mean paired difference of 0.015 ($t(29) = 3.51$, $p = .0015$; Wilcoxon $p = .0019$; Hedges' $g = 0.62$). For both metrics, 73.3% of participants had higher scores under CHORUS than Baseline.

Quality trends over task order provided additional evidence of adaptation (Fig. 7). BLEU increased over rows under CHORUS (slope

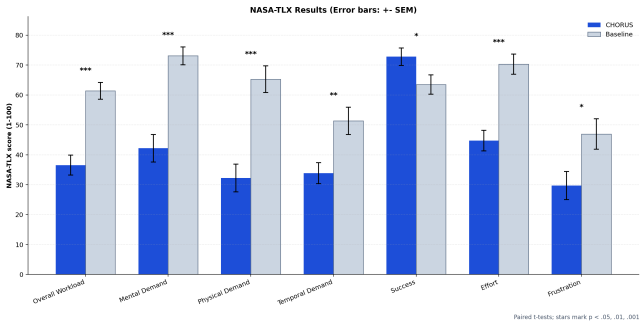

Figure 6: NASA-TLX scores for CHORUS and Baseline.

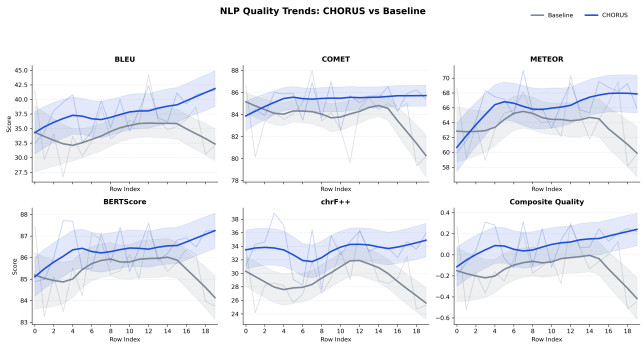

Figure 7: Task-order trends in translation-quality metrics under CHORUS and Baseline.

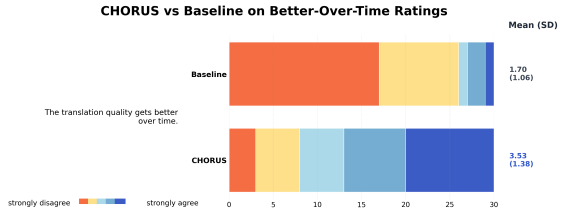

Figure 8: Human expert ratings of whether translation quality improved over time under CHORUS and Baseline.

= 0.297, $p = .014$), as did METEOR [3] (slope = 0.00236, $p = .024$), while Baseline remained largely flat on these metrics. A composite quality index was flat for Baseline (slope = 0.00062, $p = .923$) and showed an upward trend for CHORUS (slope = 0.0119, $p = .059$). BLEURT [64] also suggested that Baseline quality declined over rows (slope = -0.00830, $p < .001$), while CHORUS showed a smaller, non-significant decrease (slope = -0.00295, $p = .101$).

Human expert ratings showed the same direction. Three translation experts rated whether translation quality improved over time. CHORUS received a mean rating of 3.53 ($SD = 1.38$), whereas Baseline received 1.70 ($SD = 1.06$), with Baseline ratings concentrated toward disagreement and CHORUS ratings showing more agreement (Fig. 8).

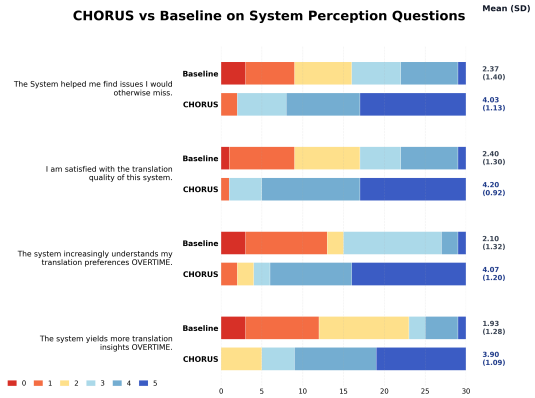

Figure 9: Participant responses to perception questions.

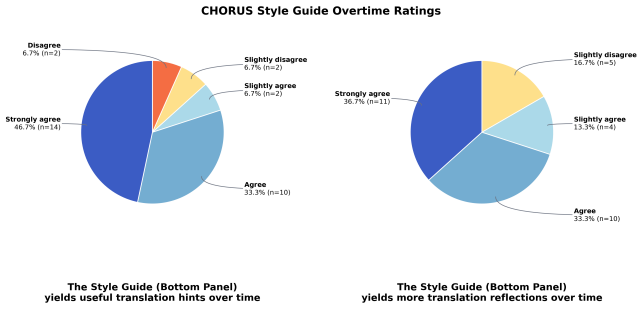

Figure 10: Style-guide dimension distributions over time.

6.3 Structured Revision and Reflection

Open-ended responses helped explain why CHORUS reduced effort and improved quality. Participants attributed the lower effort to targeted support, clearer problem visibility, and less need to manually formulate prompts. They described CHORUS as making issue-specific feedback easier to inspect than broad chatbot responses, and several participants noted that focusing on one quality dimension at a time reduced the burden of mentally tracking accuracy, terminology, fluency, and style together (Fig. 9).

Participants also described CHORUS as supporting more deliberate quality decisions. They reported that dimension-specific suggestions made trade-offs more explicit, helped them catch issues they might otherwise miss, and supported verification before confirming a sentence. In contrast to a static chatbot, participants described CHORUS as increasingly reflecting their ongoing edits, preferred correction patterns, and habitual areas of focus, reducing the need to repeatedly restate the same priorities.

The Style Guide and radar chart supported structured review by making revision history and MQM coverage visible. Participants used the radar chart to see which dimensions received more attention and which were underrepresented, and many reported using the visualization to rebalance their focus in later revisions. The Style Guide also helped participants understand MQM dimensions, identify possible misclassifications in their own edits, and justify decisions with more confidence (Fig. 10).

7 Discussion

The results suggest that CHORUS improves professional translation not by replacing translator judgment, but by reorganizing LLM assistance around how translators revise. MQM agents made quality concerns more visible and easier to inspect, reducing the need to repeatedly prompt, compare broad responses, and mentally track several revision goals at once. This shift may help explain the 33.8% reduction in completion time, the lower Live Effort, and the lower NASA-TLX workload (RQ1). Because CHORUS differs from the baseline in both its multi-agent backend and its purpose-built interface, we interpret these gains as the effect of the integrated workflow rather than of the agent architecture alone. It also aligns with prior work on mixed-initiative translation, where useful automation supports professional control rather than removing it [15, 23, 26].

CHORUS improved translation metrics (BLEU +3.08, COMET +0.015) and expert ratings (RQ2). One explanation is that participants could evaluate whether a suggestion improved accuracy, fluency, or another MQM dimension before committing to it. Making trade-offs inspectable reduced the risk of accepting suboptimal rewrites and supported more deliberate final decisions.

CHORUS also supported reflection beyond sentence-level editing (RQ3), complementing the inspectability and reduced re-prompting that participants highlighted in their qualitative feedback. The Style Guide and radar chart made revision history and quality coverage visible, helping participants notice where their attention was concentrated, identify possible blind spots, and explain decisions using shared quality categories.

More broadly, CHORUS illustrates a division of labor for human-centered multi-agent systems. Agents can surface dimension-specific issues, organize alternatives, and accumulate interaction history, while the professionals remain responsible for interpreting context, weighing trade-offs, and making the final decision. This pattern may generalize to other revision-heavy domains where people coordinate multiple specialized AI agents under shifting priorities.

8 Limitations and Future Work

Our evaluation assessed CHORUS as an integrated workflow because the goal of evaluation was to holistically understand the efficacy of the system for professional translation. Efficacy of individual components remains the future work.

In addition, our evaluation was limited to sentence-level English–Chinese tasks from WMT24. Other language pairs, and domain-specific effort signals remain to be tested, and we hope to expand the user study through multilingual and document-level studies in future work.

9 Conclusion

This paper presented CHORUS, an MQM-aligned multi-agent workspace for professional translators. CHORUS separates revision into seven inspectable dimensions and integrates users' interaction history to reduce repeated MQM-specific adjustments, and visualizes revision patterns through an MQM review interface. In a study with 30 professional translators, CHORUS reduced completion time, effort, and workload while improving translation quality evaluated via both

expert ratings and existing methods. Findings suggest that CHORUS can better support professional work by preserving translators' control and revealing quality concerns for inspections.

10 Safe and Responsible Innovation Statement

CHORUS explores a human-accountable paradigm for multi-agent interaction in professional translation, where AI support remains explainable, inspectable, and subject to human judgment. Responsible deployment should protect confidential materials, audit uneven performance across users and domains, and prevent over-reliance on agent suggestions in high-stakes settings.

References

[1] Vicent Alabau, Ragnar Bonkb, Christian Buck, Michael Carlb, Francisco Casacuberta Nolla, Mercedes García-Martínez, Jesus Gonzalez Rubio, Philipp Koehn, Luis Alberto Leiva Torres, Bartolomé Mesa-Lao, et al. 2013. CASMACAT: An open source workbench for advanced computer aided translation. (2013).

[2] Abeer Alabbas and Khalid Alomar. 2025. A weighted composite metric for evaluating user experience in educational chatbots: balancing usability, engagement, and effectiveness. *Future Internet* 17, 2 (2025), 64.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[4] Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation*. 62–69.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Vicent Briva-Iglesias. 2025. Are AI agents the new machine translation frontier? Challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891* (2025).

[7] Vicent Briva-Iglesias, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. Large language models" ad referendum": How good are they at machine translation in the legal domain? *arXiv preprint arXiv:2402.07681* (2024).

[8] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*. 249–256.

[9] Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering* 26, 2 (2020), 137–161.

[10] Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. 181–190.

[11] Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. arXiv:2502.12404 [cs.CL] https://arxiv.org/abs/2502.12404

[12] Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3393–3402.

[13] Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of llm domain adaptation for business it machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. 610–622.

[14] Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving llm-based machine translation with systematic self-correction. *CoRR* (2024).

[15] George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation* 12, 1 (1997), 175–194.

[16] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics* 9 (2021), 1460–1474.

[17] Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, et al. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*. 47–81.

[18] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*. 733–774.

[19] António Góis and André FT Martins. 2019. Translator2vec: Understanding and representing human post-editors. In *Proceedings of Machine Translation Summit XVII: Research Track*. 43–54.

[20] Attila Görög. 2014. Quality evaluation today: the dynamic quality framework. In *Proceedings of Translating and the Computer 36*.

[21] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*. 33–41.

[22] Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833* (2019).

[23] Spence Green, Jeffrey Heer, and Christopher D Manning. 2015. Natural language translation at the intersection of AI and HCI. *Commun. ACM* 58, 9 (2015), 46–53.

[24] Spence Green, Sida I Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D Manning. 2014. Human effort and machine learnability in computer aided translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1225–1236.

[25] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[26] Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: a practical interactive machine translation system. arXiv. *arXiv preprint arXiv:2105.13072* (2021).

[27] James W Hunt and Thomas G Szymanski. 1977. A fast algorithm for computing longest common subsequences. *Commun. ACM* 20, 5 (1977), 350–353.

[28] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745* 1, 10 (2023).

[29] Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 4253–4273.

[30] Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*. 107–120.

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

[32] Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Vol. 5. Kent State University Press.

[33] Joseph B Kruskal. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review* 25, 2 (1983), 201–237.

[34] Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2019. Interactive-predictive neural machine translation through reinforcement and imitation. In *Proceedings of Machine Translation Summit XVII: Research Track*. 96–106.

[35] Philippe Langlais, George Foster, and Guy Lapalme. 2000. TransType: a computer-aided translation typing system. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*.

[36] Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of artificial intelligence research* 67 (2020), 653–672.

[37] Arle Lommel. 2018. Metrics for translation quality assessment: A case for standardising error typologies. In *Translation quality assessment: From principles to practice*. Springer, 109–127.

[38] Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*. 165–172.

[39] Arle Lommel, Serge Gladkoff, Alan K Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, et al. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*. 75–94.

[40] Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica* 12 (2014), 0455–463.

[41] Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*. 5570–5587.

[42] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural*

Information Processing Systems 36 (2023), 46534–46594.

[43] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7297–7306.

[44] Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*. 1–11.

[45] Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. 227–237.

[46] MQM Council. [n. d.]. MQM (Multidimensional Quality Metrics). https://themqm.org/. Accessed: 2026-03-20.

[47] OpenAI. 2026. GPT-5.3 Instant: Smoother, More Useful Everyday Conversations. https://openai.com/index/gpt-5-3-instant/. Accessed: 2026-07-20.

[48] Daniel Ortiz-Martinez, Luis A Leiva, Vicent Alabau, Ismael Garcia-Varea, and Francisco Casacuberta. 2011. An interactive machine translation system with online learning. In *Proceedings of the ACL-HLT 2011 System Demonstrations*. 68–73.

[49] Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics* 13 (2025), 73–95.

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[51] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[52] Alvaro Peris and Francisco Casacuberta. 2019. Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language* 58 (2019), 98–126.

[53] Meinrad Perrez and Michael Reicherts. 1996. A computer-assisted self-monitoring procedure for assessing stress-related behavior under real life conditions. *Ambulatory assessment. Computer-assisted psychological and psychophysiological Ambulatory assessment. Computer-assisted psychological and psychophysiological methods in monitoring and field studies* (1996), 51–67.

[54] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems* 34 (2021), 4816–4828.

[55] Mark A Przybocki, Gregory A Sanders, and Audrey N Le. 2006. Edit Distance: A Metric for Machine Translation Evaluation.. In *LREC*. 2038–2043.

[56] Vikas Raunak, Arul Menezes, and Hany Awadalla. 2023. Dissecting in-context learning of translations in GPT-3. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 866–872.

[57] Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 12009–12024.

[58] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*. 2685–2702.

[59] Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44, 3 (2018), 393–401.

[60] María Isabel Rivas Ginel and Joss Moorkens. 2025. Translators' trust and distrust in the times of GenAI. *Translation Studies* 18, 2 (2025), 283–299.

[61] RWS. [n. d.]. The LISA QA Model. https://docs.rws.com/en-US/sdl-multitrans-785465/the-lisa-qa-model-788069. Accessed: 2026-03-26.

[62] Klaus R Scherer. 2014. On the nature and function of emotion: A component process approach. In *Approaches to emotion*. Psychology Press, 293–317.

[63] Jörg Schütz. 1999. Deploying the SAE J2450 translation quality metric in language technology evaluation projects. In *Proceedings of Translating and the Computer 21*.

[64] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 7881–7892.

[65] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. 223–231.

[66] Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation* 23, 2 (2009), 117–127.

[67] Yixiao Song, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing Human Evaluation in Machine Translation with Comparative Judgement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 20536–20551.

[68] Maria Stasimioti and Vilelmini Sosoni. 2020. Translation vs post-editing of NMT output: Insights from the English-Greek language pair. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*. 109–124.

[69] Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. 240–251.

[70] David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15406–15427.

[71] Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)* 21, 1 (1974), 168–173.

[72] Dongqi Wang, Haoran Wei, Zhirui Zhang, Shujian Huang, Jun Xie, and Jiajun Chen. 2022. Non-parametric online learning from human feedback for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11431–11439.

[73] George Wang, Jiaqian Hu, and Safinah Ali. 2025. MAATS: A Multi-Agent Automated Translation System Based on MQM Evaluation. *arXiv preprint arXiv:2505.14848* (2025).

[74] Jiayi Wang, Ke Wang, Fengming Zhou, Chengyu Wang, Zhiyong Fu, Zeyu Feng, Yu Zhao, and Yuqi Zhang. 2024. Synslator: An interactive machine translation tool with online learning. In *Companion Proceedings of the ACM Web Conference 2024*. 1023–1026.

[75] Xi Wang, Shiyang Zhang, Fanfei Meng, and Lan Li. 2025. The Hidden Pitfalls of E-Dictionaries: How Inaccuracies Affect Chinese Language Users. *Lexicography* 12, 2 (2025), 107–130.

[76] Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. Transagents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 131–141.

[77] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15474–15492.

[78] Daichi Yamaguchi, Rei Miyata, Atsushi Fujita, Tomoyuki Kajiwara, and Satoshi Sato. 2024. Automatic Decomposition of Text Editing Examples into Primitive Edit Operations: Toward Analytic Evaluation of Editing Systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 1899–1914.

[79] Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. 2025. How Well Do Large Reasoning Models Translate? A Comprehensive Evaluation for Multi-Domain Machine Translation. *arXiv preprint arXiv:2505.19987* (2025).

[80] Kamer Ali Yuksel, Ahmet Gunduz, Abdul Baseet Anees, and Hassan Sawaf. 2025. Efficient Machine Translation Corpus Generation: Integrating Human-in-the-Loop Post-Editing with Large Language Models. *arXiv preprint arXiv:2502.12755* (2025).

[81] Enze Zhang, Jiaying Wang, Mengxi Xiao, Jifei Liu, Ziyan Kuang, Rui Dong, Eric Dong, Sophia Ananiadou, Min Peng, and Qianqian Xie. 2025. DITING: A Multi-Agent Evaluation Framework for Benchmarking Web Novel Translation. *arXiv preprint arXiv:2510.09116* (2025).

[82] Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 10961–10988.

[83] Shijie Zhang, Renhao Li, Songsheng Wang, Philipp Koehn, Min Yang, and Derek F Wong. 2025. HiMATE: A Hierarchical Multi-Agent Framework for Machine Translation Evaluation. *arXiv preprint arXiv:2505.16281* (2025).

[84] Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*. 468–481.

[85] Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1656–1671.

[86] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 19724–19731.

[87] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*. 2765–2781.