Toward Machine Translation Literacy: How Lay Users Perceive and Rely on Imperfect Translations

Yimin Xiao[✤] **Yongle Zhang**[✤] **Dayeon Ki**[*] **Calvin Bao**[*] **Marianna J. Martindale**[✤] **Charlotte Vaughn**[❖] **Ge Gao**[✤] **Marine Carpuat**[*]

[✤]College of Information [*]Department of Computer Science [❖]Language Science Center
University of Maryland, College Park
yxiao@umd.edu

Abstract

As Machine Translation (MT) becomes increasingly commonplace, understanding how the general public perceives and relies on imperfect MT is crucial for contextualizing MT research in real-world applications. We present a human study conducted in a public museum ($n = 452$), investigating how fluency and adequacy errors impact bilingual and non-bilingual users' reliance on MT during casual use. Our findings reveal that non-bilingual users often over-rely on MT due to a lack of evaluation strategies and alternatives, while experiencing the impact of errors can prompt users to reassess future reliance. This highlights the need for MT evaluation and NLP explanation techniques to promote not only MT quality, but also MT literacy among its users.

1 Introduction

As machine translation (MT) becomes more deeply embedded in daily life through apps and chatbots, people increasingly rely on it for casual, everyday tasks: understanding signs, browsing foreign-language content, and making quick decisions. While this wide adoption signals the success of NLP technologies, it also raises questions about public understanding and appropriate use (Carpuat et al., 2025). Are users equipped to detect errors or understand their consequences? Can they calibrate their trust in systems? Do they know what MT can and cannot do? In other words, as the reach of MT has increased, what do we know about the general public's MT literacy (Bowker and Ciro, 2019b)?

This paper responds to this year's EMNLP theme of "Advancing our Reach: Interdisciplinary Recontextualization of NLP," which calls for rigorous evaluation of how NLP technologies actually impact society and intersect with other fields. While benchmark scores for MT continue to improve, these evaluations alone do not tell us how the general public perceives and relies on MT. Work in Translation Studies emphasizes the need for MT literacy as translation tools gain a broad range of users, who may lack the language proficiency or background knowledge to critically evaluate outputs (O'Brien and Ehrensberger-Dow, 2020; Bowker, 2025). However, designing interventions that can support such a large and diverse population requires a better understanding of how people interact with MT in the wild: how they perceive its quality, how they rely on it, and what might influence those decisions.

In this work, we study how people's reliance on MT is impacted by fluency and adequacy errors during casual, low-stakes use. Our study builds on that of Martindale and Carpuat (2018), which measured user trust in presence of MT errors, but without controlling for their impact on decision-making. Here, we go further by drawing from HCI methods for studying trust and reliance in AI (Vereschak et al., 2021). We also conduct our experiment in a public museum setting (Vaughn et al., 2024), enabling us to recruit participants from many walks of life and ground MT use in a specific environment.

We found that bilingual and non-bilingual users rely on MT differently, as can be expected. More surprisingly, we found that non-bilingual users often rely on imperfect MT not because they assume the outputs to be correct, but because they lack strategies to approach assessing outputs and making decisions on their basis. Interestingly, experiencing the impact of MT errors in low-stakes settings still prompted users to reevaluate their future use of the tool. These findings motivate several directions for future MT and NLP research, including the development of MT systems that support users in assessing and recovering from errors, and the development of tools to support MT literacy training inspired by the task conducted here. In the process, we hope to illustrate the benefits of recontextualizing MT and NLP work in an interdisciplinary

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33997–34014
November 4-9, 2025 ©2025 Association for Computational Linguistics

fashion to address the societal implications of MT.

2 Research Questions & Background

The research questions (RQs) addressed in this paper are motivated by a body of work spanning the translation studies, HCI and MT literatures.

How is MT Used? This is a hard question to answer because MT is available to anyone with an internet connection (Savoldi et al., 2025). By 2021, the Google Translate app alone had over a billion installations (Pittman, 2021), with an estimated 99.97% of MT users being non-professionals (Nurminen, 2021). Surveys of UK residents show high satisfaction for low-stakes uses (Vieira et al., 2022), but public service professionals also frequently use MT in their work without formal training (Nunes Vieira, 2024). Another concern is the use of MT in high-stakes contexts like healthcare and law, where errors can cause significant harm (Khoong et al., 2019; Vieira et al., 2021; Lee et al., 2023). Furthermore, MT tools do not yet meet user needs equally across socioeconomic and geographic contexts (Santy et al., 2021), negatively impacting daily lives for groups such as migrant workers in India and immigrant populations in the U.S. (Liebling et al., 2020; Valdez et al., 2023).

In response, researchers emphasize the importance of raising awareness about the strengths and limitations of MT technology (Vieira et al., 2021). Users not only lack an understanding of how MT operates and might fail, they also do not grasp the risks and complexities inherent in the translation process itself (O'Brien and Ehrensberger-Dow, 2020; Bowker, 2025). Efforts to improve translation and MT literacy have emerged, particularly in academic settings (Bowker and Ciro, 2019a; Bowker, 2025). However, extending these efforts to the general public remains challenging due to the diversity of user needs and the difficulty in reaching all relevant audiences.

What Makes MT "Good"? Methods for evaluating MT quality have evolved alongside MT technology itself. White et al. (1993) identified two core evaluation dimensions: fluency, or "well-formedness" of the system outputs in the target language, and adequacy, "the extent to which the semantic content of [..] texts from each source language was present in the translations". Automatic metrics emerged to provide rapid quality assessments, comparing system outputs to professional reference translations using n-gram overlap (Papineni et al., 2002; Popović, 2015) or neural methods (Ma et al., 2019; Freitag et al., 2022). As MT systems advanced, human evaluation regained prominence, with protocols ranging from holistic quality ratings (Graham et al., 2017), error annotation by type and severity (Lommel et al., 2014; Freitag et al., 2021), to post-editing of MT outputs (Raunak et al., 2023; Xu et al., 2024). In this approach, third party annotators evaluate translation quality to establish a ground truth rating. While this is an effective guide for system development, we also need to measure users' first person perception of MT to address the MT literacy gap.

How is MT Perceived? User studies of MT and AI systems highlight that people's perception, reliance, and trust are shaped by the types of errors they encounter and their level of source language proficiency. Exposure to translation errors or uncertainties in AI outputs can affect users' trust and confidence in the system (Zhang et al., 2020; Kim et al., 2024). For MT, fluency errors play an important role, with evidence that they impact reported trust in MT more than adequacy errors (Martindale and Carpuat, 2018), and that they serve as a heuristic for judging overall translation quality (Robertson and Díaz, 2022). Further, when the outputs from MT and AI appeared to be fluent and natural sounding, people with limited proficiency experience significant challenge to understand and assess the nuanced meaning expressed in these outputs (Xiao et al., 2024). Findings on the role of domain expertise in AI evaluation (Lee and Chew, 2023; Nourani et al., 2020) also raise the question of how users' source language proficiency influences their ability to assess translation quality and calibrate their reliance on MT.

Research Questions. This contexts motivates the following Research Questions (RQs):

RQ1. How do people perceive the quality of translations containing different error types, and subsequently, how do their decision-making accuracy and confidence vary with these error types?

RQ2. How do people with varying proficiency in the source language perceive and evaluate translations containing different error types and make decisions based on these translations?

RQ3. How does people's trust in the MT system differ with exposure to different types of MT errors, and does their proficiency in the source language mediate this trust?

RQ4. What primary strategies do people use in evaluating translations? And does the adopted strategy vary by people's proficiency in the source language or MT error types?

3 Methods

In this section, we detail the experimental study we conducted to explore our RQs.

3.1 Study Design

We designed a mixed 2 × 3 experiment with two factors. **(1) MT Correctness** is *within*-subject factor with two levels: Correct vs. Incorrect MT and **(2) MT Error Type** is *between*-subject with three levels: FLUENCY ERROR WITHOUT IMPACT, ADEQUACY ERROR WITHOUT IMPACT, and ADEQUACY ERROR WITH IMPACT. Participants were randomly assigned to one of three conditions.

The task is designed to mimic a low-stakes MT-mediated communication scenario in the museum setting where the study takes place – thus grounding the study in a shared real-world context. The task design is inspired by prior HCI work on MT based on controlled experiments in mock scenarios for collaboration and communication (Yamashita et al., 2009; Xu et al., 2014; Gao et al., 2013; Wang et al., 2013). Prior studies have explored MT's impact on tasks such as problem-solving (Yamashita et al., 2009; Zhang et al., 2022) and brainstorming (Gao et al., 2013; Wang et al., 2013) in team-based settings. Others have investigated MT use in informal, everyday contexts, such as exchanging greetings, engaging in casual conversations (Xu et al., 2014), or facilitating housing purchases between newly arrived migrants and local people (Xiao et al., 2025). For the museum setting, it was important to have a task that can be understood quickly, by participants of any background and any age (Vaughn et al., 2024), and thus we exploited the only context that we know all participants share: the museum itself. Another important consideration in our task design is our focus on low-stakes scenarios that are characteristic of everyday MT use (Vieira et al., 2022), as these experiences are formative for users' trust in the technology.

Participants were asked to complete a navigation task to assist a fictional Spanish-speaking character trace her steps in the museum and retrieve a lost item. The task consists of four trials (T1-T4). In each trial, participants received a stimulus composed of one Spanish message and its translated version in English. The Spanish message describes a location the fictional character was at. Upon viewing the stimulus, participants were asked to 1) rate their perception of the translation quality, 2) select one out of two images that best match the location described in the Spanish message, 3) rate their confidence in their selection. After completing each trial, participants were told whether their selection was correct or not. Figure 1 shows an example stimulus with the two candidate images for participants to select from. Screenshots of the full task for an example stimulus are provided in Appendix Figure 4. We manipulated the MT translation in each stimulus according to our experimental conditions. For the within-subject factor, participants viewed correct English translations in T1 and T4 and incorrect English translations in T2 and T3. For the between-subject factors, participants were randomly assigned into one of the three conditions (FLUENCY ERROR WITHOUT IMPACT, ADEQUACY ERROR WITHOUT IMPACT, ADEQUACY ERROR WITH IMPACT) and viewed stimuli containing corresponding errors in T2 and T3. Presenting correct translations at the beginning (T1) helps participants calibrate their understanding of the task and establish initial trust in the translation system. Introducing errors in the middle trials (T2 and T3) enables us to observe how users respond to disruptions after forming these expectations. Within this structured ordering, we exhaustively counterbalanced the order of stimuli and randomly assigned participants to one of the pre-generated stimulus sequence.

The experiment was implemented in Qualtrics. Prior to the task, participants viewed written instructions and two example stimuli. They were also to complete pre-task and post-task surveys to share demographic information, English and Spanish proficiency, and overall task experience. To ensure the clarity and engagement of our task materials, we conducted several pilots before the formal study, including think-aloud sessions across age groups and crowd-sourced surveys. Trained research assistants were available onsite to assist participants as needed, and optionally to debrief after the study for participants who wanted to know more about the study, as well as MT and generative AI.

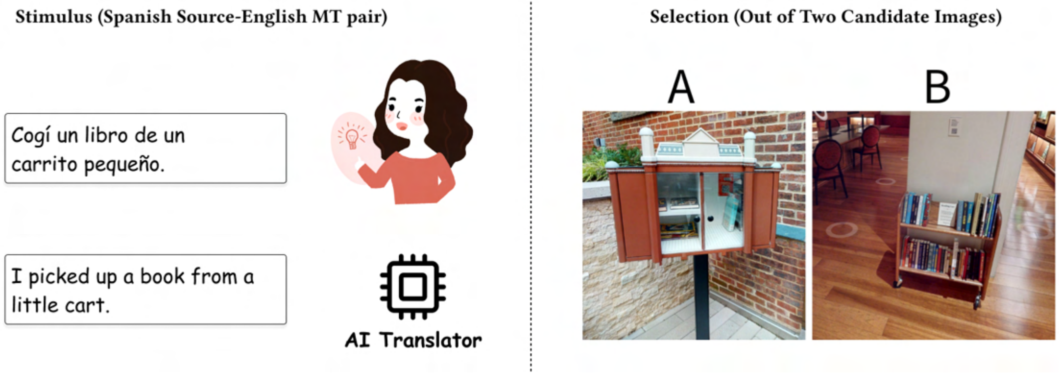

Stimulus (Spanish Source-English MT pair)
Cogí un libro de un carrito pequeño.
I picked up a book from a little cart.
AI Translator
Selection (Out of Two Candidate Images)
A
B

Figure 1: Example stimulus (**Stimulus**) with two candidate images for participants to select from (**Selection**). Image A shows a small outdoor library box and image B shows a small, low wooden shelf holding several books.

3.2 Stimuli Collection

Stimuli were designed to balance study control with museum setting needs. We aimed for realistic Spanish-to-English MT examples illustrating Correct vs. Incorrect translations, featuring three error types: Fluency Error without Impact, Adequacy Error without Impact, and Adequacy Error with Impact. Texts were kept concise and readable for broad accessibility, including for young audiences and non-native English speakers. Content was linked to the museum context to boost engagement, leveraging the only shared real-world context among visitors. Due to these constraints, using naturally occurring MT errors was not feasible; thus, we constructed the stimuli as described below.

Construction. Each stimulus includes a Spanish source sentence, its English MT, and two candidate images – only one of which matches the Spanish description. After manually creating concise and museum-relevant Spanish sentences, we generate English translations using two methods. First, we sampled LLM translations. We prompt models like LLaMA-3 8B (Grattafiori et al., 2024), ChatGPT, and GPT-4 (OpenAI et al., 2024) with diverse decoding strategies (e.g., top-p, top-k, sampling) to produce both accurate translations and natural MT errors. Second, we introduce controlled changes to the Spanish input (Xu et al., 2023) -– such as misspellings, deletions, round-trip MT, paraphrasing, and style transfer – to elicit subtle MT errors like tense shifts or word omissions.

Validation. We crowdsource three independent validation checks to ensure that text and images are interpreted as intended. These checks helped filter out ambiguous messages and unintuitive text-image pairings. Native English speakers assessed translation fluency and selected matching images based on English translations, while proficient Spanish speakers did so based on the Spanish source. Through this process, we selected 16 final stimuli across four categories (Correct, Fluency Error without Impact, Adequacy Error without Impact, Adequacy Error with Impact, as illustrated in Table 1), organized into pairs for trials T1–T4. Participants were randomly assigned one T1+T4 pair and one T2+T3 pair.

3.3 Participants

We collected data at the Language Science Station (LSS; Vaughn et al. 2024), a research and public engagement laboratory located in the Planet Word museum in Washington, D.C. The LSS invites museum visitors to contribute to research studies within the galleries and to engage in conversations about language with students and educators from local universities. This setting was particularly well suited to our study: it enabled us to recruit a large and diverse sample of participants with varying Spanish proficiency, experience with MT, and demographic backgrounds. Moreover, the LSS's mission of science communication and education aligns closely with our research goal of promoting MT and AI literacy.

Participants were required to have sufficient English proficiency to comprehend the task instruction to participate. Participants were consented prior to the start of the study and were not compensated for their participation. The study protocol was approved by the University of Maryland's Institutional Review Board.

In total, 517 people participated in our study. 65 participants did not complete the task or par-

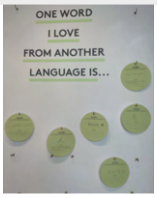
Spanish Source	English MT	Correct Image	Incorrect Image
Fluency Error without Impact			
Miré un instrumento de cuerda en una vitrina.	Look at a bristle instrument in a showcase.	 (*Instruments behind the glass windows of a music shop with a blue storefront labeled "Zithers Music Shop".*)	 (*Karaoke setup on a wall, with the title "Unlock the Music" and visuals of lyrics and a man playing guitar.*)
Adequacy Error with Impact			
Pisé letras y señales de muchos idiomas.	Write letters and signs of many languages.	 (*Floor near an elevator with characters and signs in many languages spread around.*)	 (*Green sticky notes on a wall with the text "One word I love from another language is ...".*)

Table 1: Stimuli examples for two error types that impact the correct vs. incorrect image decision. **Spanish Source:** Spanish source sentence; **English MT:** MT of the corresponding Spanish source; **Correct Image:** Image aligned with the Spanish content; **Incorrect Image:** Image that is plausible but does not align with the Spanish. The full set of stimuli can be found in Appendix A.2.

ticipated in the task with other people such as a family member or friend. After filtering out these responses, our sample included valid responses from 452 participants, including 269 females, 159 males, 15 people identified as non-binary or genderfluid, and 9 people who preferred not to disclose their gender. Their average age was 35.12 years old (S.E. = 1.65). For Spanish proficiency, 63 participants (13.94%) reported no proficiency at all, 353 participants (78.10%) reported some proficiency, and 36 participants (7.96%) reported high proficiency. For English proficiency, 406 participants (89.82%) reported high English proficiency, with 46 participants (10.18%) reporting some English proficiency. For self-reported usage of MT tools, 159 (35.18%) participants never used MT, 162 participants (35.84%) rarely used MT, 74 participants (16.37%) sometimes used MT, 40 participants (8.85%) often used MT, and 17 participants (3.76%) used MT almost everyday.

3.4 Dependent Variables

We collected several dependent measures to address our research questions. For each stimulus, we assessed perceived translation quality, decision accuracy, and confidence. At the end of the survey, each participant reported their willingness to reuse the MT system and any evaluation strategies used.

- **Translation Quality Perception**: Participants rated each English translation after seeing the Spanish source, choosing whether it seemed correct (1), unclear (0.5), or problematic (0).
- **Decision Accuracy**: Participants selected which of two images best matched the Spanish message. A correct choice scored 1; incorrect, 0.
- **Decision Confidence**: Participants rated confidence in their decision on a 5-point Likert scale.
- **Willingness to Reuse MT**: Participants rated willingness to reuse the system on a 5-point Likert scale.
- **Evaluation Strategy**: Participants selected which of three strategy types best matched what they did: no strategy/intuitive, comparative analysis of Spanish and English texts, or Spanish proficiency-based judgment.

4 Results

We address each RQ in turn, by presenting the results of the statistical analysis and its implications (Section 4.1-4.3), before summarizing qualitative feedback (Section 4.4).

4.1 RQs 1 & 2: Perception of MT and Decision Making Based on Translations

The RQs 1 and 2 ask about participants' perception of MT quality and reliance on MT outputs in decision-making tasks, and how they are influenced by MT error types and users Spanish proficiency.

Perception of Translation Quality. We fitted a Cumulative Link Mixed Model (CLMM) with a logit link to investigate how participants' perceptions of translation quality across error types and effects of Spanish proficiency. In the model setup, we treated participants' Perception of Translation Quality for each stimulus as dependent variable. Our two fixed-effect independent variables were MT error types (Correct, FLUENCY ERROR WITHOUT IMPACT, ADEQUACY ERROR WITHOUT IMPACT, and ADEQUACY ERROR WITH IMPACT) and Spanish Proficiency (No Spanish Proficiency, Some Spanish Proficiency, and High Spanish Proficiency). We treated Participant ID and Stimulus ID as random effects. Our co-variates included individual participants' age, gender, English proficiency, and prior experience with MT. We applied Bonferroni corrections to adjust multiple comparisons. Figure 2 illustrates the core results.

We find that participants' perception of MT quality was influenced by both their language proficiency and MT error types. FLUENCY ERROR WITHOUT IMPACT were perceived significantly less believable than Correct (Coefficient = -2.97, $p < .001$). Participants with higher Spanish proficiency reported higher ratings of Perception of Translation Quality (Coefficient = 1.08, $p < .001$).

Further, we observed significant interaction effects between MT error types and Spanish proficiency. Pair-wise comparisons revealed that participants with High Spanish Proficiency were able to detect MT errors, regardless of the error types. Participants with Some Spanish Proficiency were able to perceive FLUENCY ERROR WITHOUT IMPACT but not ADEQUACY ERROR WITH IMPACT or ADEQUACY ERROR WITHOUT IMPACT. Participants with No Spanish Proficiency were not able to perceive any MT errors, even the fluency errors which were detectable by monolingual English speakers based on our validation studies.

These results shed new light on prior work suggesting that fluency and adequacy errors impact people's perceptions of MT quality differently. Here, participants' assessment of quality varies depending on their proficiency in the source language. Surprisingly, this is the case even when error cues are visible in the target language.

Decision-Making Accuracy. We used a Generalized Linear Mixed Model (GLMM) with a logit link. Similarly, we treated MT Error Type and Spanish Proficiency as fixed-effect independent variables, and Participant ID and Stimulus ID as random effects; We applied the same set of control variables and correction method as before.

The analysis shows a significant main effect for ADEQUACY ERROR WITH IMPACT, where participants showed significantly lower Decision-Making Accuracy when faced with ADEQUACY ERROR WITH IMPACT compared to Correct (Coefficient = -2.18; $p < .001$). There was also a significant interaction between ADEQUACY ERROR WITH IMPACT and Spanish Proficiency, motivating pair-wise comparisons. Figure 3 illustrates the core results.

These results show that participants with No or Some Spanish Proficiency were less accurate in their Decision-Making Accuracy based on MT with misleading errors (ADEQUACY ERROR WITH IMPACT), while those with High Spanish Proficiency showed no difference in Decision-Making Accuracy across different MT errors. This further illustrates the disparate impact of MT errors on users depending on different proficiency levels.

Decision-Making Confidence. We used a CLMM model with a logit link, with the same set of fixed-effect, random-effect and control variables and correction method as above. we found that Spanish Proficiency has a significant main effect on Decision-Making Confidence. Specifically, participants with higher Spanish proficiency demonstrated significantly higher Decision-Making Confidence compared to those with lower proficiency (Coefficient = 0.225, z = 2.47, $p < 0.05$). However, MT error types did not show significant main effects, and there were no significant interaction effects between MT Errors types and Spanish Proficiency.

Recap & Implications. These results highlight the importance of accounting for language proficiency differences in human studies of MT and in MT interaction design. We find that bilingual and non-bilingual users rely on MT in predictably different ways, highlighting a fundamental fairness issue in the use of MT. More surprisingly, lack of Spanish proficiency impacted participants' ability to perceive fluency errors, which were in principle detectable based on the English alone, and even some knowledge of the language was not sufficient to compensate for impactful adequacy errors.

This emphasizes the need for developing and testing MT and NLP techniques to guide error assessment. Future work could draw from methods for highlighting differences between MT input and outputs (Briakou et al., 2023), and a wealth of existing techniques to automatically estimate the quality of a translation or detect potential errors (Specia et al., 2010; Fomicheva et al., 2021; Kocmi and Fe-

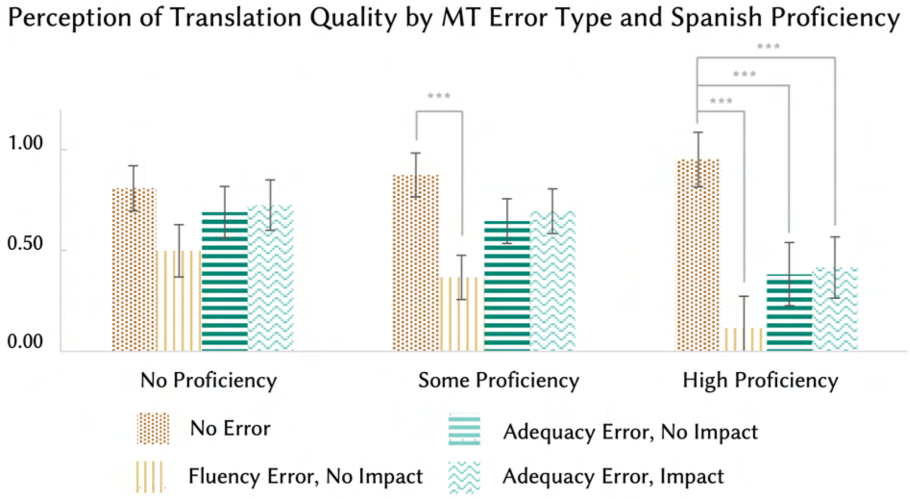


Figure 2: Perception of Translation Quality (i.e., Believability) by MT Error Type (legend) and Spanish Proficiency (x-axis). Note: *$p < .05$; **$p < .01$; ***$p < .001$.

dermann, 2023; Guerreiro et al., 2023; Jung et al., 2024). However, how to present such information to effectively support users in practice remains an open question (Tsai and Wang, 2015; Zouhar et al., 2021; Mehandru et al., 2023).

4.2 RQ3: Individual Participants' Willingness to Reuse MT

For our RQ3, we analyzed the relationship between MT Error type (three between-subject levels: FLUENCY ERROR WITHOUT IMPACT, ADEQUACY ERROR WITHOUT IMPACT), and ADEQUACY ERROR WITH IMPACT), Spanish proficiency (three levels: No Spanish Proficiency, Some Spanish Proficiency, High Spanish Proficiency), and participants' Willingness to Reuse the AI translator by evaluating their willingness to reuse the tool in the future. We analyzed results per participant, using an ANOVA to examine the main effects and interactions, and controlling for covariates such as age, gender, English proficiency, and MT experience.

The analysis revealed a significant main effect of MT Error type on Willingness to Reuse in the AI translator (F(2, 374) = 4.798, $p < .01$). Specifically, participants' willingness to reuse the tool varied across the error types, suggesting that the type of MT error influences users' trust in the system. To further explore these differences, we conducted a Tukey's HSD post-hoc test. Participants reported significantly lower Willingness to Reuse in the AI when FLUENCY ERROR WITHOUT IMPACT were present (Mean = 2.81, S.E. = 0.20) compared to conditions where there were ADEQUACY ERROR WITHOUT IMPACT (Mean = 3.00, S.E. = 0.21; $p < .05$). Willingness to Reuse was also significantly lower when participants encountered ADEQUACY ERROR WITH IMPACT (Mean = 2.90, S.E. = 0.20) compared to situations where there was ADEQUACY ERROR WITHOUT IMPACT (Mean = 3.00, S.E. = 0.21, $p < .05$). There was no significant difference between ADEQUACY ERROR WITH IMPACT and FLUENCY ERROR WITHOUT IMPACT. There were no significant main effects for Spanish Proficiency or interactions between MT Error Type and Spanish Proficiency.

Recap & Implications. These results indicate that participants expressed lower trust in MT systems, as measured by their willingness to reuse the tool, after exposure to certain error types (FLUENCY ERROR WITHOUT IMPACT or ADEQUACY ERROR WITH IMPACT) but not others (ADEQUACY ERROR WITHOUT IMPACT).

This finding has practical implications. It suggests that controlled, fictional interactions with MT might influence future tool use, even through a brief session. Such settings make it possible to control for the impact of errors, facilitating risk awareness without waiting for natural errors. This can form the basis for future MT literacy interventions, motivating MT and NLP technical work to support their development. For example, semi-

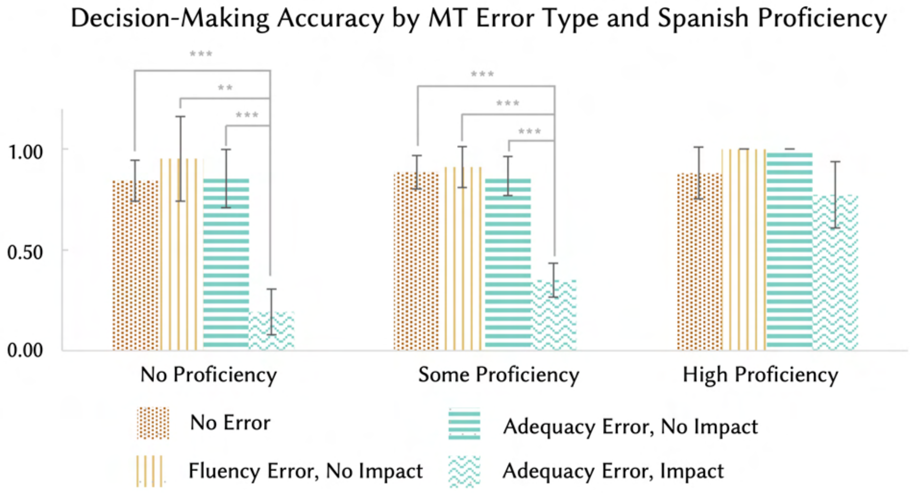


Figure 3: Decision Accuracy by MT Error Type and Spanish Proficiency. Note: *$p < .05$; **$p < .01$; ***$p < .001$.

automatically producing stimuli, such as scoring or generating appropriate MT input/output pairs for a given context, decision, and error type, would be beneficial. More broadly, this highlights the need for system developers to understand technology's failure modes, whether for AI systems generally (Ehsan et al., 2022) or MT specifically (Robertson et al., 2023), as a complement to improving benchmark performance.

4.3 RQ4: Individual Participants' Strategy Used to Evaluate MT Quality

Last, we analyzed the main strategy reported by participants segmented by their Spanish Proficiency. We built a multinomial logistic regression model with the strategy categories as the dependent variable and MT Error Type and Spanish proficiency as predictors, while controlling for covariates as in previous models. In this model, No Strategy was set as the reference category for comparison.

Results reveal several significant predictors of strategy choice. Participants in FLUENCY ERROR WITHOUT IMPACT condition were significantly less likely to report using the Comparison Strategy (Coefficient = -2.30, SE = 0.26, $p < .001$) and Proficiency Strategy (Coefficient = -0.95, SE = 0.34, $p < .01$). Participants in ADEQUACY ERROR WITHOUT IMPACT condition were also less likely to report the use of Comparison Strategy (Coefficient = -1.95, SE = 0.23, $p < .001$) and Proficiency Strategy (Coefficient = -1.39, SE = 0.41, $p = .001$). Participants in the ADEQUACY ERROR WITH IMPACT condition were more likely to report Comparison Strategy (Coefficient = 1.78, SE = 0.31, $p < .001$) but less likely to report Proficiency Strategy (Coefficient = -1.80, SE = 0.33, $p < .001$). Spanish proficiency was found to have a significant effect on strategy choice. Participants with higher Spanish proficiency were less likely to report using Comparison Strategy (Coefficient = -5.34, SE = 0.44, $p < .001$), but more likely to report using Proficiency Strategy (Coefficient = 9.97, SE = 0.25, $p < .001$).

Recap & Implications. These results indicate that different error conditions may prompt different strategies for evaluating translation quality. As expected, participants with higher Spanish proficiency are able to directly assess MT quality. More interestingly, participants attempt to compare the MT and the original more when they experience the negative impact of adequacy errors by losing points. However, the decision and confidence measures (Section 4.1) show that they are not successful when they are not proficient in Spanish. This calls for more work to design techniques that helps people make such comparisons. Providing cognitive forcing functions to encourage users to engage in a strategic evaluation of outputs has shown promise for other AI-decision making tasks (Buçinca et al., 2021), and it remains to be seen how to design MT specific solutions. A promising strategy is to leverage question-answering to surface inconsistencies between the source and its translation (Ki et al., 2025a; Fernandes et al., 2025), which has been shown to help users decide whether a translation is reliable enough to share safely (Ki et al., 2025b). However, it remains unclear whether such feedback also supports users in making more accurate inferences when answering content-specific questions, as examined in this study.

4.4 Post-Task Qualitative Feedback

Post-task debrief sessions with participants indicated that the study prompted them to want to know more about MT. Notes taken by research assistants frequently mentioned the following debrief topics: participants' personal experience using MT for work and travel, understanding of the technical mechanisms of MT and why MT makes errors, participants' evaluation strategy used in the task, explanation of the translation errors in our stimuli, and general questions about our research design. Although we did not directly measure the educational benefits of this research participation experience, these discussion topics indicated a general interest of the public in acquiring information on MT systems following participation in the study.

This feedback underscores a key research gap: supporting users in developing appropriate mental models of MT tools. While prior work has largely focused on feedback about the quality of individual outputs, less attention has been given to helping users understand the broader capabilities and limitations of these systems. Our findings point to the potential of simulations to promote MT literacy outside classroom settings. Although museum visitors who opt into such studies may be more motivated to learn than randomly selected MT users, this approach remains promising for professionals who rely on MT for communication with little or no formal training (Nunes Vieira, 2024; Mehandru et al., 2022). A more ambitious long-term goal is to design generic translation apps that explicitly foster MT literacy for all users.

5 Conclusion

In summary, we presented a human-study investigating how fluency and adequacy errors impact bilingual and non-bilingual users' reliance on MT during casual use. Our findings confirm that bilingual and non-bilingual users perceive and rely on potentially imperfect MT outputs differently. More surprisingly, they suggest that non-bilingual users over-rely on MT not because they have high confidence in their correctness, but because they do not know what else to do: they lack strategies to evaluate outputs and reason about how to use them. However, experiencing the impact of errors in the study settings was sufficient to prompt users to reassess future reliance.

These findings motivate several directions for future MT and NLP research, including the development of MT systems that support users in assessing and recovering from errors, the development of tools to support MT literacy training inspired by the task conducted here, as well as understanding how trust formation in casual settings common to everyday users (Vieira et al., 2022) impact their behaviors in high-stakes use cases.

More broadly, we aim to underscore the value of complementing intrinsic evaluations of MT quality with studies of how people experience and respond to MT, thereby motivating future work at the intersection of NLP, HCI, and Translation Studies (Carpuat et al., 2025). We also highlight the value of conducting human studies beyond the lab and crowdsourcing platforms to engage diverse segments of the public, using data collection as an opportunity for science communication (Vaughn et al., 2024) and promoting AI literacy.

Limitations

Our research has several limitations.

Conducting the study in a museum setting introduced several constraints. The low-stakes nature of the task, featuring a fictional interlocutor and minimal consequences for incorrect decisions, may not fully capture real-world decision-making dynamics. Additionally, the short interaction duration limits the ability to examine trust development over time (Holliday et al., 2016).

Our participant pool also reflects a selection bias; while it likely includes a broader age range than typical university lab studies, museum visitors who choose to engage with a scientific study are not necessarily representative of the general population and may scrutinize translations more carefully than typical users.

Furthermore, our study focuses on a single decision-making task closely aligned with reading comprehension, which may not capture the complexity of real-world MT usage scenarios.

Therefore, we caution against overgeneralizing our findings and highlight the need for further research to explore the relationships between

MT quality perception, decision-making, and trust across users with varying language proficiency levels and in diverse real-world scenarios.

Acknowledgements

We would like to thank the Language Science Station team and all the research assistants for hosting our research and facilitating data collection with museum visitors, including Ge (Stella) Huang, Sarah Nam, Kelly Brennan, Ashley Chau, Jamie Flynn, Miriam Franklin-Grinkorn, Lyosha Genzel, Tessa Goldlust, Tzipporah Harker, Isabel Harris, Maliel Henry, Mikoalis Bedrosian Kajen, Melanie Kurtz, Eamon Maloney, Kennedy Nwosu, Jessica Orozco Contreras, Cristelle Torbeso, Stacey Torbeso, and Nichole Tramel. We are also grateful to the Planet Word leadership and staff for opening their doors to our study, and to the museum visitors who generously shared their time and insights with us. This work was supported in part by NSF Language Science Station Grant 2116959 and by the Institute for Trustworthy AI in Law and Society (TRAILS), which is supported by the National Science Foundation under Award No. 2229885. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Lynne Bowker. 2025. Machine Translation Literacy. In Lee McCallum and Dara Tafazoli, editors, *The Palgrave Encyclopedia of Computer-Assisted Language Learning*, pages 1–4. Springer Nature Switzerland, Cham.

Lynne Bowker and Jairo Buitrago Ciro. 2019a. Expanding the Reach of Knowledge Through Translation-Friendly Writing. In *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*, pages 55–78. Emerald Publishing Limited.

Lynne Bowker and Jairo Buitrago Ciro. 2019b. Towards a Framework for Machine Translation Literacy. In *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*, pages 87–95. Emerald Publishing Limited.

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. 2023. Explaining with Contrastive Phrasal Highlighting: A Case Study in Assisting Humans to Detect Translation Differences. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

Marine Carpuat, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. An Interdisciplinary Approach to Human-Centered Machine Translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Upol Ehsan, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daume III. 2022. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Preprint*, arXiv:2211.06753.

Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, Andre Martins, and Graham Neubig. 2025. Do LLMs understand your translations? evaluating paragraph-level MT with question answering. In *Second Conference on Language Modeling*.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results. *Preprint*, arXiv:2110.04392.

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*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 449–458.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Preprint*, arXiv:2310.10482.

Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, page 164–168, New York, NY, USA. Association for Computing Machinery.

Dahyun Jung, Sugyeong Eo, and Heuiseok Lim. 2024. Towards Precise Localization of Critical Errors in Machine Translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3000–3012, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Elaine C. Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine*, 179(4):580–582.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025a. AskQE: Question Answering as Automatic Evaluation for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17478–17515, Vienna, Austria. Association for Computational Linguistics.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025b. Should I Share this Translation? Evaluating Quality Feedback for User Reliance on Machine Translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 822–835, New York, NY, USA. Association for Computing Machinery.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Min Hun Lee and Chong Jun Chew. 2023. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22.

Won Lee, Elaine C. Khoong, Billy Zeng, Francine Rios-Fetchko, YingYing Ma, Kirsten Liu, and Alicia Fernandez. 2023. Evaluation of Commercially Available Machine Interpretation Applications for Simple Clinical Communication. *Journal of General Internal Medicine*.

Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet Needs and Opportunities for Mobile Translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA. Association for Computing Machinery.

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

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA. Association for Machine Translation in the Americas.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2016–2025, New York, NY, USA. Association for Computing Machinery.

Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human*

Computation and Crowdsourcing, volume 8, pages 112–121.

Lucas Nunes Vieira. 2024. Uses of AI Translation in UK Public Service Contexts | CIOL (Chartered Institute of Linguists). https://www.ciol.org.uk/ai-translation-uk-public-services.

Mary Nurminen. 2021. *Investigating the Influence of Context in the Use and Reception of Raw Machine Translation*. Tampere University.

Sharon O'Brien and Maureen Ehrensberger-Dow. 2020. MT Literacy—A cognitive view. *Translation, Cognition & Behavior*, 3(2):145–164.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeff Pittman. 2021. Google Translate: One billion installs, one billion stories. https://blog.google/products/translate/one-billion-installs/.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Samantha Robertson and Mark Díaz. 2022. Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2223–2238, New York, NY, USA. Association for Computing Machinery.

Samantha Robertson, Zijie J. Wang, Dominik Moritz, Mary Beth Kery, and Fred Hohman. 2023. Angler: Helping Machine Translation Practitioners Prioritize Model Improvements. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–20, New York, NY, USA. Association for Computing Machinery.

Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, and Vivek Seshadri. 2021. Language Translation as a Socio-Technical System:Case-Studies of Mixed-Initiative Interactions. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '21, pages 156–172, New York, NY, USA. Association for Computing Machinery.

Beatrice Savoldi, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the Hands of Many:Centering Lay Users in Machine Translation Interactions. *Preprint*, arXiv:2502.13780.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Hsing-Lin Tsai and Hao-Chuan Wang. 2015. Evaluating the Effects of Interface Feedback in MT-embedded Interactive Translation. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 2247–2252, New York, NY, USA. Association for Computing Machinery.

Susana Valdez, Ana Guerberof Arenas, and Kars Ligtenberg. 2023. Migrant communities living in the Netherlands and their use of MT in healthcare settings. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 325–334, Tampere, Finland. European Association for Machine Translation.

Charlotte Vaughn, Hannah Mechtenberg, and Jessica Orozco Contreras. 2024. The Language Science Station at Planet Word: A language research and engagement laboratory at a language museum. *Linguistics Vanguard*, 10(s3):245–255.

Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):327:1–327:39.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Lucas Nunes Vieira, Carol O'Sullivan, Xiaochun Zhang, and Minako O'Hagan. 2022. Machine translation in society: Insights from UK users. *Language Resources and Evaluation*.

Hao-Chuan Wang, Susan Fussell, and Dan Cosley. 2013. Machine translation vs. common language: Effects on idea exchange in cross-lingual groups. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 935–944.

John S. White, Theresa A. O'Connell, and Lynn M. Carlson. 1993. Evaluation of machine translation. In

Proceedings of the Workshop on Human Language Technology - HLT '93, page 206, Princeton, New Jersey. Association for Computational Linguistics.

Yimin Xiao, Yuewen Chen, Naomi Yamashita, Yuexi Chen, Zhicheng Liu, and Ge Gao. 2024. (Dis)placed Contributions: Uncovering Hidden Hurdles to Collaborative Writing Involving Non-Native Speakers, Native Speakers, and AI-Powered Editing Tools. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2):403:1–403:31.

Yimin Xiao, Cartor Hancock, Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, Marine Carpuat, and Ge Gao. 2025. Sustaining human agency, attending to its cost: An investigation into generative ai design for non-native speakers' language use. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Bin Xu, Ge Gao, Susan R. Fussell, and Dan Cosley. 2014. Improving machine translation by showing two outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3743–3746, New York, NY, USA. Association for Computing Machinery.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback. *Preprint*, arXiv:2311.09336.

Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 679–688.

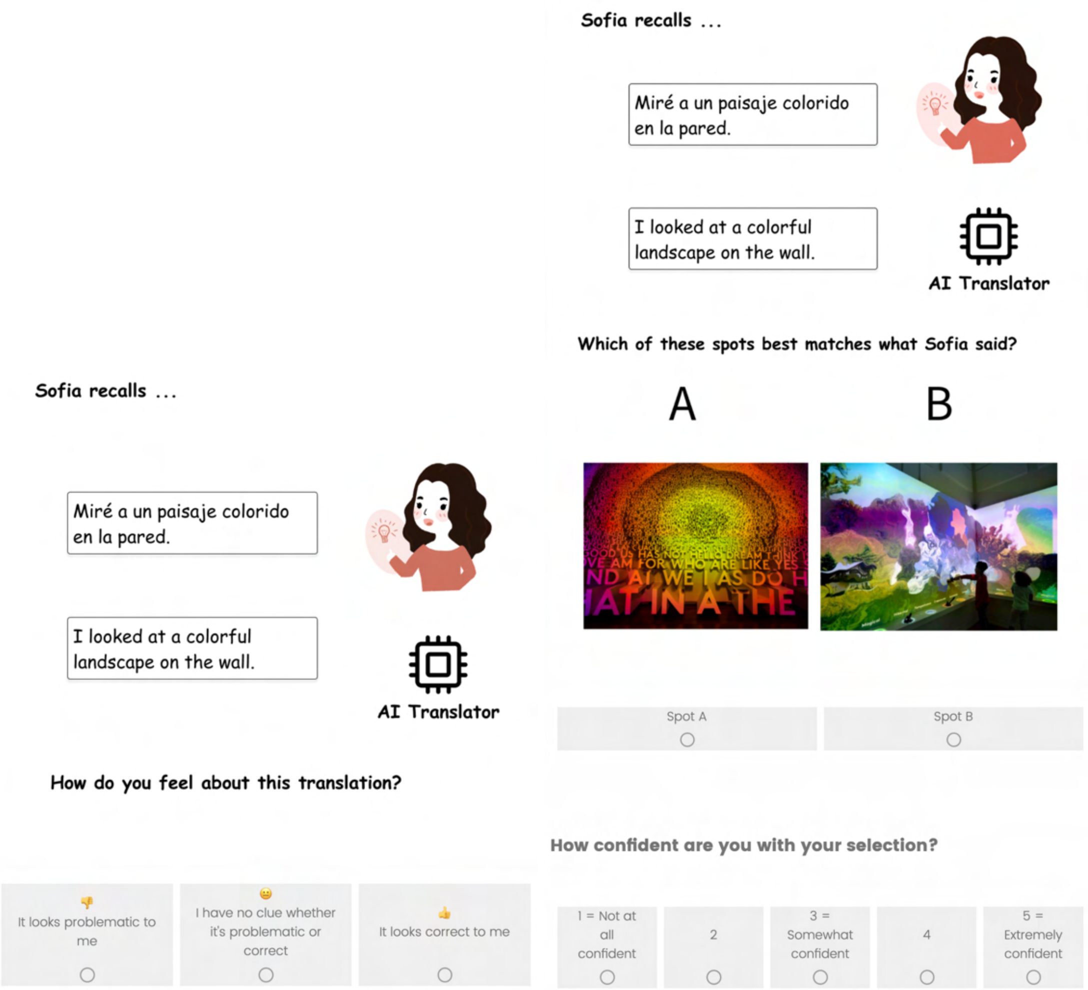
Yongle Zhang, Dennis Asamoah Owusu, Marine Carpuat, and Ge Gao. 2022. Facilitating global team meetings between language-based subgroups: When and how can machine translation help? *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–26.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305.

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021. Backtranslation Feedback Improves User Confidence in MT, Not Quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 151–161, Online. Association for Computational Linguistics.

A Appendix

A.1 Full Task of an Example Stimulus



Sofia recalls ...

Miré a un paisaje colorido en la pared.

I looked at a colorful landscape on the wall.

AI Translator

How do you feel about this translation?

👎 It looks problematic to me ○

😐 I have no clue whether it's problematic or correct ○

👍 It looks correct to me ○

Sofia recalls ...

Miré a un paisaje colorido en la pared.

I looked at a colorful landscape on the wall.

AI Translator

Which of these spots best matches what Sofia said?

A B

Spot A ○

Spot B ○

How confident are you with your selection?

1 = Not at all confident ○

2 ○

3 = Somewhat confident ○

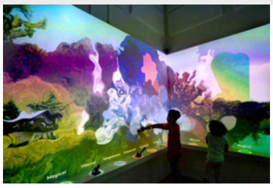
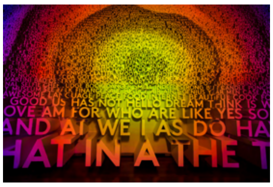
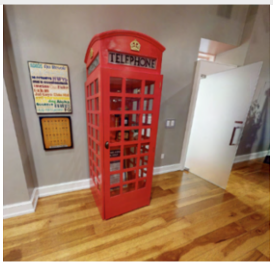
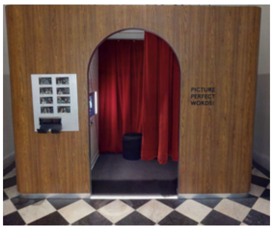
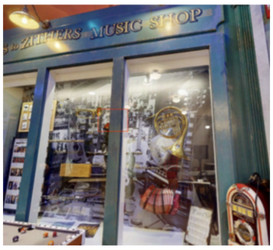
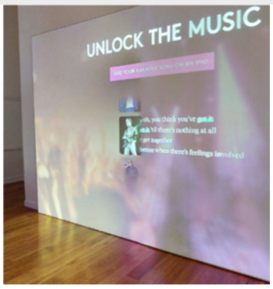
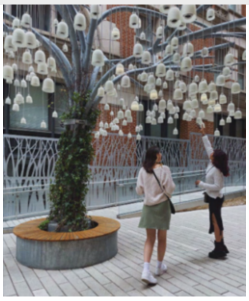
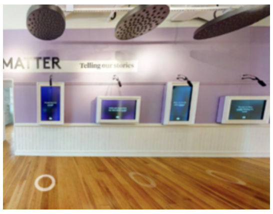
4 ○

5 = Extremely confident ○

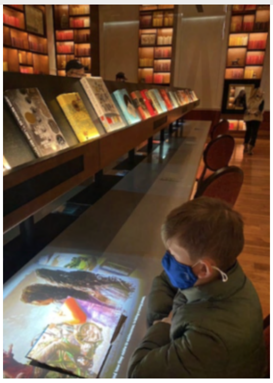
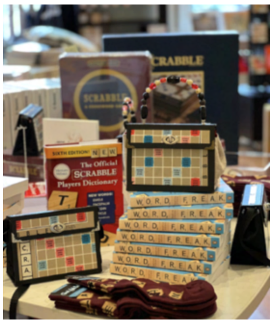
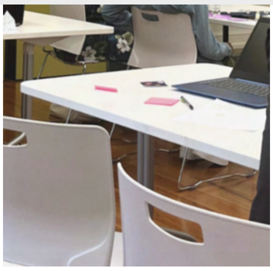
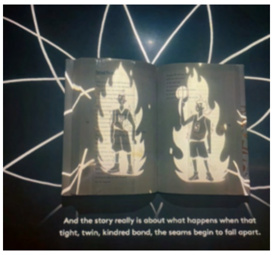
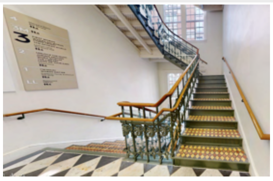
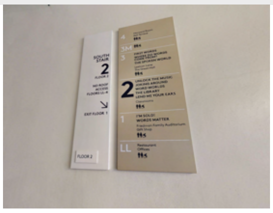
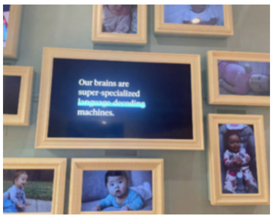
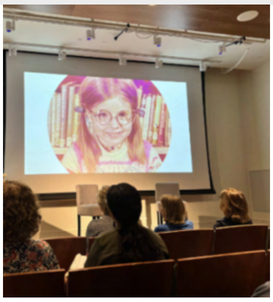
Figure 4: Screenshot of the full task for an example stimulus. **Left:** Participants rate their perception of the translation quality. **Right:** Next, participants select one out of two images that best match the location described in the Spanish message and rate their confidence in their selection. Image A shows colorful text radiating in a circular shape. Image B shows two children looking at a colorful, abstract digital display on the wall.

A.2 Final Stimuli

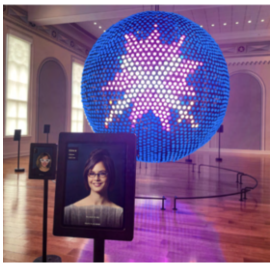
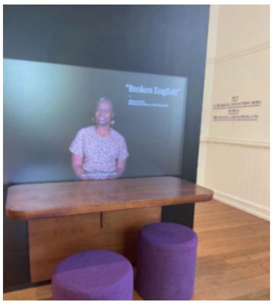
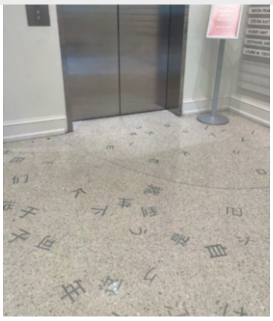
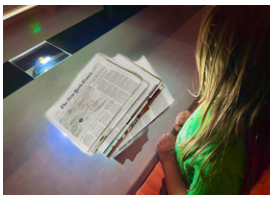
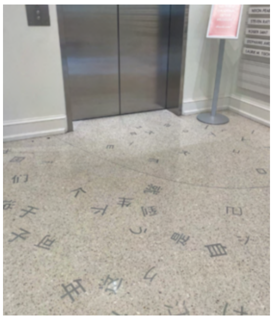
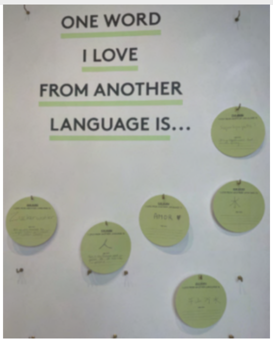
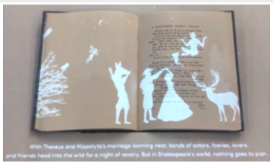
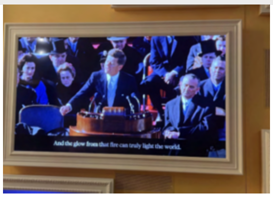
Table 2: All 16 stimuli used in our study. **Error Type:** Type of MT error present in the English MT; **Spanish Source:** Spanish source sentence; **English MT:** MT of the corresponding Spanish source; **Correct Image:** image describing the Spanish source; **Incorrect Image:** image that is plausible but does not describe the Spanish source.

Spanish Source	English MT	Correct Image	Incorrect Image
Correct			
Miré a paisaje colorido en la pared.	I looked at a colorful landscape on the wall.	 (*Two children looking at a colorful, abstract digital display on the wall.*)	 (*Colorful text radiating in a circular shape.*)
Correct			
Me paré al lado de una cabina roja brillante.	I stood next to a bright red booth.	 (*Red British-style telephone booth placed indoors on a polished wooden floor.*)	 (*Photo booth with an arched entrance, red curtains, and a small stool inside.*)
FLUENCY ERROR WITHOUT IMPACT			
Miré un instrumento de cuerda en una vitrina.	Look at a bristle instrument in a showcase.	 (*Instruments behind the glass windows of a music shop with a blue storefront labeled "Zithers Music Shop".*)	 (*Karaoke setup on a wall, with the title "Unlock the Music" and visuals of lyrics and a man playing guitar.*)
FLUENCY ERROR WITHOUT IMPACT			
Toqué los parlantes blancos que colgaban del árbol.	I touched the white speakers hanging from the aol.	 (*Two women touching the white speakers hanging from a tree.*)	 (*A purple wall displaying interactive screens with microphones handing on the top.*)

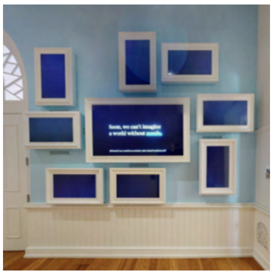
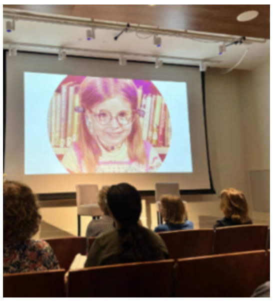
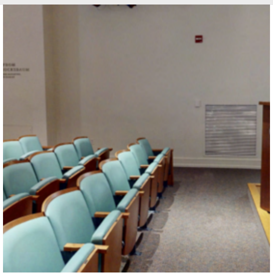
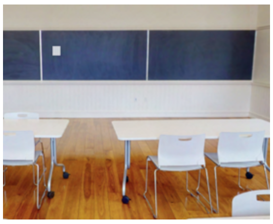
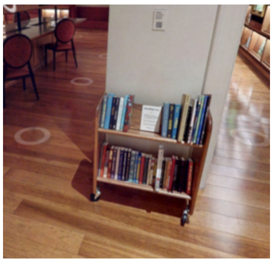
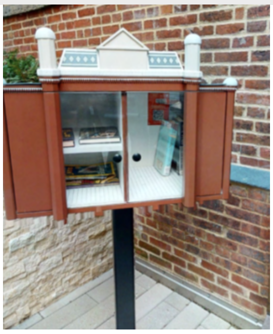
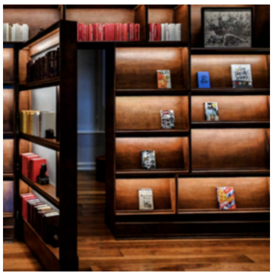
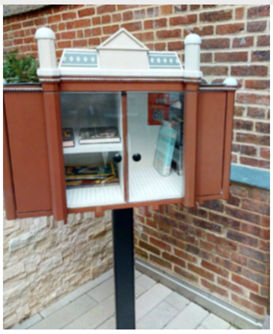
Continued on next page

Spanish Source	English MT	Correct Image	Incorrect Image
Fluency Error without Impact			
Miré a los libros para niños en la sala de la biblioteca.	I looked at the children's bros in the library room.	 (*A child wearing a blue mask engaging with an interactive display in a library.*)	 (*Scrabble-themed merchandise, including books and decorative items in a gift shop.*)
Fluency Error without Impact			
Escribí mis pensamientos sobre el libro en un nota adhesiva.	I wrote my thoughts on the sticky book.	 (*Pink sticky notes on a white table along with a pen and a laptop.*)	 And the story really is about what happens when that tight, twin, kindred bond, the seams begin to fall apart. (*Illuminated open book with depictions of two figures surrounded by flames.*)
Adequacy Error without Impact			
Subí al tercer piso.	Take it to the third floor.	 (*Stairs at the third floor with patterned tiles and decorative iron railings.*)	 (*Directory sign at the second floor.*)
Adequacy Error without Impact			
Escuché los balbuceos de una pequeña bebé.	I listened to the babbles of a little baby.	 Our brains are super-specialized language-decoding machines. (*Digital audio screen surrounded by framed photos of babies on a wall.*)	 (*Audience seated in a room watching a large screen with a picture of a smiling girl wearing glasses.*)
Adequacy Error without Impact			

Continued on next page

Spanish Source	English MT	Correct Image	Incorrect Image
Habían varias pantallas chicas.	There were several girls in glasses.	 (*Illuminated spherical globe with digital screens featuring women in glasses.*)	 (*Exhibit with a projected video featuring a person discussing a topic "Broken English".*)
Adequacy Error without Impact			
Pisé letras y señales de muchos idiomas.	Write letters and signs of many languages.	 (*Floor near an elevator with characters and signs in many languages spread around.*)	 (*Girl looking at a newspaper on a flat digital screen.*)
Adequacy Error with Impact			
Pisé letras y señales de muchos idiomas.	Write letters and signs of many languages.	 (*Floor near an elevator with characters and signs in many languages spread around.*)	 (*Green sticky notes on a wall with the text "One word I love from another language is ...".*)
Adequacy Error with Impact			
Vi una historia cobrar vida.	I saw history come to life.	 (*Illuminated open book with silhouettes of animals, trees, and people.*)	 (*Framed screen displaying a historical speech.*)
Adequacy Error with Impact			

Continued on next page

Spanish Source	English MT	Correct Image	Incorrect Image
Había varias pantallas chicas.	There were several girls in glasses.	 (*Display with multiple framed screens against a light blue wall.*)	 (*Audience seated in a room watching a large screen with a picture of a smiling girl wearing glasses.*)
Adequacy Error with Impact			
Me senté en el asiento al frente del auditorio.	I sat in the seat at the front of the classroom.	 (*Front view of the classroom-style seats with teal-cushioned chairs.*)	 (*Second row view of the classroom with white tables facing a wall with blackboards.*)
Correct			
Cogí un libro de un carrito pequeño.	I picked up a book from a little cart.	 (*Small, low wooden shelf holding several books.*)	 (*Small outdoor library box with doors that open to reveal books inside.*)
Correct			
Casi me paso la puerta oculta.	I almost missed the hidden door.	 (*Library with wooden shelves with a hidden door open.*)	 (*Doorway leading to another room with a wooden bookshelf.*)