Should I Share this Translation?
Evaluating Quality Feedback for User Reliance on Machine Translation

Dayeon Ki[✤] Kevin Duh[*] Marine Carpuat[✤]
[✤]University of Maryland [*]Johns Hopkins University
{dayeonki,marine}@umd.edu kevinduh@cs.jhu.edu

Abstract

As people increasingly use AI systems in work and daily life, mechanisms that help them use AI responsibly are urgently needed, especially when they are not equipped to verify AI predictions themselves. We study a realistic Machine Translation (MT) scenario where monolingual users decide whether to share an MT output, first without and then with quality feedback. We compare four types of quality feedback: explicit feedback that directly give users an assessment of translation quality using (1) error highlights and (2) LLM explanations, and implicit feedback that helps users compare MT inputs and outputs through (3) backtranslation and (4) question–answer (QA) tables. We find that all feedback types, except error highlights, significantly improve both decision accuracy and appropriate reliance. Notably, implicit feedback, especially QA tables, yields significantly greater gains than explicit feedback in terms of decision accuracy, appropriate reliance, and user perceptions – receiving the highest ratings for helpfulness and trust, and the lowest for mental burden.[1]

1 Introduction

Artificial Intelligence (AI) are increasingly deployed to support human decision-making across a wide range of domains (Buçinca et al., 2021; Dastin, 2022; Ma et al., 2024). As these systems are adopted by the general public, there is a growing need for feedback mechanisms that help users construct their own functional explanations – reasoning grounded in the goals and consequences of an AI output to determine how and when to rely on it for safe and effective use (Lombrozo and Wilkenfeld, 2019; Schoeffer et al., 2024). Prior work has evaluated various forms of feedback through human studies, such as error highlights (Khashabi et al., 2018; Carton et al., 2020) or free-text explanations (Bussone et al., 2015; Bansal et al., 2021; Buçinca et al., 2021). However, many human-centered studies evaluating the impact of feedback in real-world application settings are still needed. Although designing such studies is challenging, as it requires accounting for the knowledge and assumptions people bring to decision-making tasks (Lage et al., 2019; Wiegreffe and Marasovic, 2021), it remains essential to adopt a human-centered evaluation since these feedback mechanisms are ultimately intended to support human users (Boyd-Graber et al., 2022; Zhu et al., 2024; Carpuat et al., 2025).

To address this, and in line with prior work that empirically investigates real-world use cases in other domains (Hong et al., 2020; Bhatt et al., 2020; Liao et al., 2020), we evaluate AI feedback for Machine Translation (MT), where many users critically need support because they lack the language proficiency needed to evaluate MT outputs. Imagine a situation during the COVID-19 pandemic: You regularly read official guidelines in English, but your Spanish-speaking neighbor cannot access this information. You turn to MT to share this information – but as an English monolingual, how can you determine whether the Spanish MT output is accurate enough to safely share, or if it contains critical errors risking misinformation? This is a practical yet challenging scenario for monolingual users, who lack both source language proficiency (Bowker and Ciro, 2019; Liebling et al., 2020) and domain expertise (Nourani et al., 2020; Lee and Chew, 2023) to reliably evaluate MT quality, and who often lack effective strategies for deciding when to trust imperfect MT (Xiao et al., 2025b).

Prior studies have proposed quality feedback mechanisms to support MT decision-making, such as paraphrases, Quality Estimation (QE) scores, and backtranslation, but findings on their impact on user confidence and decision accuracy remain

[1] https://github.com/dayeonki/mt_quality_feedback

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12069–12092
November 4-9, 2025 ©2025 Association for Computational Linguistics

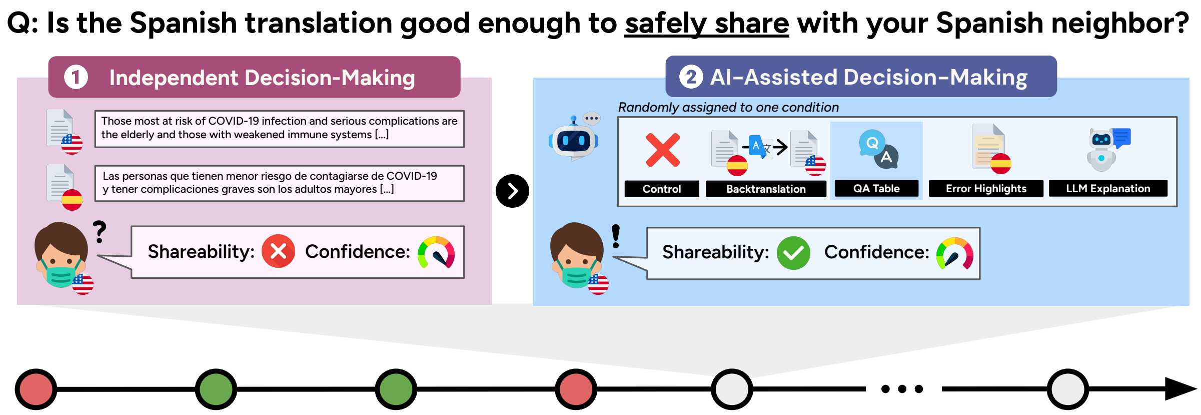

Q: Is the Spanish translation good enough to safely share with your Spanish neighbor?
1 Independent Decision-Making
Those most at risk of COVID-19 infection and serious complications are the elderly and those with weakened immune systems [...]
Las personas que tienen menor riesgo de contagiarse de COVID-19 y tener complicaciones graves son los adultos mayores [...]
?
Shareability:
Confidence:
2 AI-Assisted Decision-Making
Randomly assigned to one condition
Control
Backtranslation
QA Table
Error Highlights
LLM Explanation
!
Shareability:
Confidence:

Figure 1: Overview of our study setup. In our human study, each English-speaking monolingual participant reviews a sequence of 20 decision-making examples. Each example is shown in a two-step process: ❶ **Independent decision-making:** Participants first make judgments based solely on the English source and its Spanish MT output 🟪 and ❷ **AI-Assisted decision-making:** They then reassess the same example with one of five randomly assigned conditions (one control and four treatments) 🟦. For each step, they respond to two questions: **(1) Shareability:** To the best of your knowledge, is the Spanish translation good enough to safely share with your Spanish-speaking neighbor? and **(2) Confidence:** How confident are you in your assessment?

mixed (Zouhar et al., 2021; Mehandru et al., 2023). Building on this, we provide a more comprehensive assessment of which feedback types best support users in forming functional explanations to make reliable MT decisions. We evaluate four types of quality feedback, grouped by their mode of explicitness: **(1) Explicit** quality assessments of MT output (error highlights and LLM explanation); and **(2) Implicit** assessments that support input/output comparison (backtranslation and QA table).

We conduct a between-subjects human study with 91 English-speaking monolingual participants, where they are asked to decide whether Spanish MT outputs are safe to share with a hypothetical Spanish-speaking neighbor. For each of 20 examples, participants first make a decision (Independent step), then reassess the same example with a randomly assigned condition (AI-Assisted step), as illustrated in Figure 1.

Our findings reveal that all quality feedback interventions except error highlights significantly improve both decision accuracy and appropriate reliance (§4.1). Implicit feedback generally outperforms explicit types: backtranslation yields significantly higher appropriate reliance than error highlights, and QA table leads to significantly higher gains in both metrics than both explicit feedback types (§4.2). While explicit feedback prompts more decision changes, it also results in higher over-reliance, which are cases where participants change from correct to incorrect decisions after viewing feedback (§4.3). We also find that participants are better at recognizing good translations than identifying problematic ones (§5.1). In terms of self-reported perceptions on mental burden, helpfulness, and trust, QA table feedback consistently receives the best ratings (§5.2). Finally, we present participants' responses to identify which aspects of each quality feedback they found to be helpful (§5.3).

Together, these results highlight the value of quality feedback that supports users' implicit interpretation of MT outputs rather than explicitly telling them *what* to do, supporting a more *user-driven* process for making reliable decisions.

2 Background & Research Questions

2.1 Reliance on AI Systems

A growing body of work has examined the nature of human reliance on AI systems (Lai et al., 2023), particularly in scenarios involving risk and uncertainty (Jacovi et al., 2021). While the overarching goal is to design trustworthy AI, studying trust is complex and multifaceted. To operationalize this, prior works have developed methods to study human behavior when using AI systems with a focus on reliance (de Fine Licht and Brülde, 2021), defined as the "decision to follow someone's recommendation" (Vereschak et al., 2021). Various metrics have been proposed to assess the degree of user reliance, including agreement percentage (how often a user agrees with the AI prediction) (He et al., 2023a), confidence-weighted accuracy (decision accuracy weighted by user confidence) (Mehandru et al., 2023), and switch percentage (how often a user changes their decision after seeing AI feedback) (Schmitt et al., 2021).

One of the core challenge in human-AI collaboration is achieving appropriate reliance, which is accepting correct AI advice while rejecting incorrect advice (Eckhardt et al., 2024). In contrast, under-reliance (rejecting correct advice) and over-reliance (accepting incorrect advice) are both undesirable. We build our work on this line of measuring user reliance in AI systems, but more specifically in the context of Machine Translation (MT).

2.2 Impact of Feedback on Reliance

In AI-assisted decision-making, AI systems typically play a supportive role, offering explanations in various formats, such as recommendations, confidence scores (Yin et al., 2019), (un)certainty estimates, output rationales (Bussone et al., 2015; Bansal et al., 2021; Buçinca et al., 2021), or a combination thereof (Zhang et al., 2020). These AI feedback are intended to help human users decide whether and when to rely on AI outputs (Lai et al., 2021). However, empirical studies have shown that despite the intended benefits of AI explanations in fostering human–AI collaboration, they often lead to increased user confidence without corresponding improvements in decision accuracy, resulting in over-reliance on the AI system (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021; Kim et al., 2025).

2.3 Impact of Quality Feedback on Reliance on MT

In the context of MT, the role of the AI system (i.e., MT system) takes on a different character since monolingual users often lack the mechanisms to reliably assess MT quality, and the AI prediction (i.e., MT output) is not a direct prediction for the user's decision-making task. This unique property allows MT systems to be used for many *implicit* decision-making scenarios (*e.g.,* Is this translation good enough to share with a friend? To translate an official document?). This contrasts to traditional AI-assisted decision-making tasks, which typically focus on classification settings (*e.g.,* recidivism prediction (Wang and Yin, 2021)), where the AI prediction directly maps onto a single decision. Given this difference, our work focuses on quality feedback, which are more generic assessments of MT quality rather than direct recommendations, and ask whether users can rely on such feedback to make more informed decisions.

Various forms of quality feedback have been proposed, including quality estimation (QE) score, backtranslation (Agrawal et al., 2022), error highlights that flag problematic spans in the MT output (Eksi et al., 2021; Rubino et al., 2021; Briakou et al., 2023), textual explanations of metric outputs (Fomicheva et al., 2022; Xu et al., 2023; Jiang et al., 2024; Lu et al., 2024), and question–answer (QA) pairs designed to indicate potential errors in the translation (Sugiyama et al., 2015; Krubiński et al., 2021; Han et al., 2022; Ki et al., 2025; Fernandes et al., 2025).

Only a few human studies have evaluated quality feedback on user decision-making and reliance on MT. For example, Zouhar et al. (2021) show that backtranslation significantly increased user confidence in translations, even when it did not improve decision accuracy. Mehandru et al. (2023) demonstrate that backtranslation can help users detect critical errors more effectively than QE scores in clinical settings. However, results across these studies remain mixed, and comparisons with more recent feedback mechanisms are still lacking. Our work aims to address this gap.

Research Questions. Given this context, we address the following RQs:

RQ1. How accurately and appropriately do monolingual users decide whether to share translations when provided with quality feedback?

RQ2. How does their decision-making performance vary across different quality feedback and the two modes of explicitness?

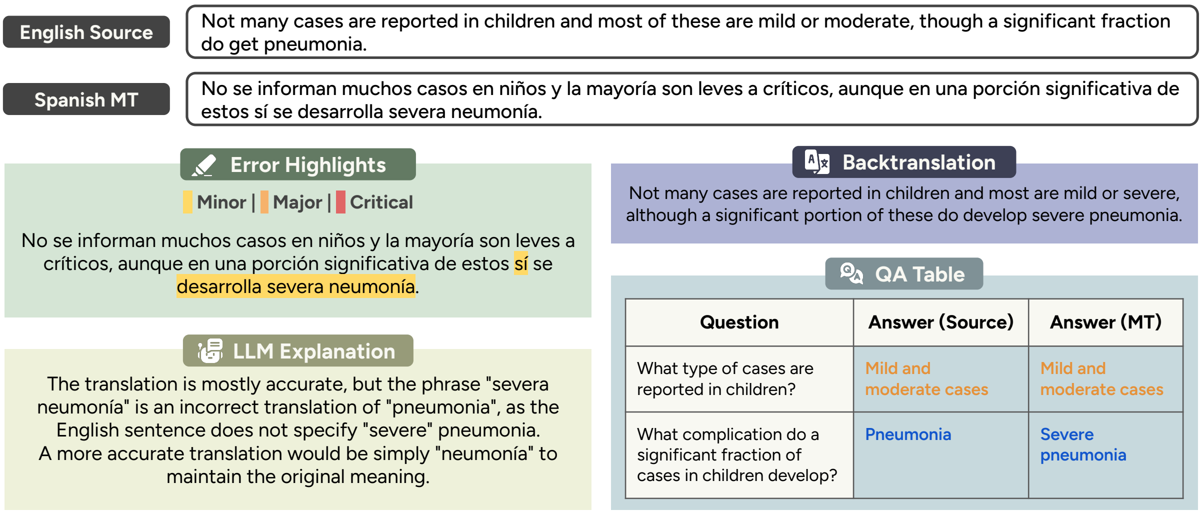
RQ3. How do users change their decisions in response to each type of quality feedback?

3 Methods

In this section, we describe the experimental study conducted to address our RQs. We outline the overall study design (§3.1), four types of quality feedback used as treatment conditions (§3.2), stimuli collection process (§3.3), participant details (§3.4), and our dependent variables (§3.5).

3.1 Study Design

We study how different types of quality feedback impact users' decision-making regarding MT shareability through a sequence of 20 examples in a between-subjects design, as illustrated in Figure 1. We use the notion of shareability to capture not only perceived MT quality but also the potential *risk* of miscommunication, highlighting the potential consequences in high-stakes contexts. This



English Source

Not many cases are reported in children and most of these are mild or moderate, though a significant fraction do get pneumonia.

Spanish MT

No se informan muchos casos en niños y la mayoría son leves a críticos, aunque en una porción significativa de estos sí se desarrolla severa neumonía.

Error Highlights

🟨 Minor | 🟧 Major | 🟥 Critical

No se informan muchos casos en niños y la mayoría son leves a críticos, aunque en una porción significativa de estos sí se desarrolla severa neumonía.

LLM Explanation

The translation is mostly accurate, but the phrase "severa neumonía" is an incorrect translation of "pneumonia", as the English sentence does not specify "severe" pneumonia. A more accurate translation would be simply "neumonía" to maintain the original meaning.

Backtranslation

Not many cases are reported in children and most are mild or severe, although a significant portion of these do develop severe pneumonia.

QA Table

Question	Answer (Source)	Answer (MT)
What type of cases are reported in children?	Mild and moderate cases	Mild and moderate cases
What complication do a significant fraction of cases in children develop?	Pneumonia	Severe pneumonia

Figure 2: During the AI-assisted decision-making step, each *treatment* group participant is presented with an English source, Spanish translation, and one of four randomly assigned quality feedback types. For error highlights, we also show a color-coded legend (🟨 Minor | 🟧 Major | 🟥 Critical) and for QA table, answer texts are displayed in orange when they are identical or highly similar, else, blue.

framing aligns with how people make such decisions in practice: while the choice to share or not is often made implicitly in real-world use cases, our study makes this decision more explicit, yet still allows participants to make their own judgment as they naturally would.

Specifically, we situate participants in a scenario where an English monolingual speaker reads official COVID-19 guidelines in English and decide whether the Spanish MT output is of sufficient quality to safely share with a Spanish-speaking neighbor. Each example is presented in two steps: ❶ **Independent** step, where participants first make judgments based solely on the English source and its Spanish translation, and a subsequent ❷ **AI-Assisted** step, where the same example is either shown again (*control* condition) or paired with a specific type of quality feedback to support decision-making (*treatment* condition). Examples are presented in randomized order, and two attention checks were included.

❶ Independent Decision-Making. For each example, participants are first asked to make judgments based solely on an English sentence and its corresponding Spanish translation. They are asked to answer two questions: **(1) Shareability:** To the best of your knowledge, is the Spanish translation good enough to safely share with your Spanish-speaking neighbor? (with binary options: ✅ Safe to share as-is, and ❌ Needs bilingual review before sharing); and **(2) Confidence:** How confident are you in your assessment? (on a five-point Likert scale from 1:Very Unconfident to 5:Very Confident). Since the recruited participants are monolingual English speakers, they are instructed to do their best in assessing the shareability of the MT outputs, despite not being fluent in Spanish.

❷ AI-Assisted Decision-Making. Subsequently, participants are randomly assigned to one of five conditions: a control condition or one of four treatment conditions, each involving a different type of quality feedback (§3.2). Those in the control condition view the same 20 examples twice in succession without receiving any quality feedback. In all conditions, participants answer the same two questions: shareability and confidence.

Pre-/Post-Task Survey. Before starting the main study, each participant is asked to answer four pre-task questions regarding their first language, proficiency in English and Spanish, and frequency of using AI translation tools in daily life or work. After completing the main study, each participant answers three post-task questions about their experience with the randomly assigned condition in terms of perceived mental burden, helpfulness, and trust for future use (Hoffman et al., 2019). Detailed descriptions are provided in Appendix C.1.

3.2 Types of Quality Feedback Intervention

Figure 2 illustrates an example of each of four quality feedback interventions. Error highlights and LLM explanation provide **explicit** quality assessments of MT output, whereas backtranslation and QA table offer **implicit** assessments that guide participants to compare MT input and output. Details on how each feedback is shown to participants are in Appendix C.1. To control for feedback quality, we balance the error rates of feedback predictions across all types (Appendix D).[2]

Error Highlights. We adopt an off-the-shelf QE system, xCOMET-XXL[3] (Guerreiro et al., 2024), to generate error annotations. Each English source and its corresponding Spanish MT is passed through the trained QE model, which produces error spans along with associated confidence scores and severity levels (minor, major, or critical). We display the highlighted error spans with a color-coded legend (🟨 Minor | 🟧 Major | 🟥 Critical). Confidence scores are not shown to participants. When the identified error span is a subword segment, we highlight at the word level to improve readability. Error annotations are presented only on the MT output, reflecting how the QE model naturally operates. If the QE model does not produce any error annotations, no highlights are shown, and the following message is displayed: "*AI did not detect any errors*". On average, each example contains 1.43 annotated error spans.

LLM Explanation. We generate natural language explanations using LLaMA-3.3 70B (Grattafiori et al., 2024).[4] Instead of instructing the model to make a shareability decision, we prompt it to assess the overall quality of the Spanish MT relative to the English source text. The exact prompt is provided in Appendix A.1. For digestibility, we constrain the model to generate responses of fewer than three sentences. The generated explanations have an average length of 46.35 words.

Backtranslation. We use the Google Translate API[5] to backtranslate the Spanish MT output since it represents one of the most widely used consumer-facing commercial MT systems (Pitman, 2021). The translation quality is reasonable, as indicated by QE scores between the Spanish MT and its backtranslation: 0.860 from COMET-QE (Rei et al., 2020) and 0.962 from xCOMET-QE XL (Guerreiro et al., 2024). Participants are not informed about the specific MT system used; instead, they are simply shown a brief explanation stating that the backtranslation represents "*how the AI system translates the Spanish MT back into English*".

QA Table. We use the AskQE framework (Ki et al., 2025) for question generation and answering, where questions are generated from the source text and answers are drawn from both the source and the backtranslated MT output. Specifically, we adopt an optimized version of AskQE that uses LLaMA-3.3 70B and entailed facts to guide question generation, and Google Translate for backtranslation of the Spanish MT output. All prompts are provided in Appendix A.2. On average, each example yields 2.65 questions, with an average question length of 10.04 words. As illustrated in Figure 2, we present the QA pairs in a table format with the following statements: "*The questions are about the original English content*" and "*The answers are based on two sources: the original English text and the Spanish translation, which has been translated back into English for display*". When the two answers are identical or highly similar, they are displayed in orange, else, blue. Similarity is computed using a soft variant of exact matching, with normalization for punctuation, case, whitespace, and articles.

3.3 Stimuli Collection

We sample 40 English-Spanish examples from the ContraTICO dataset (Ki et al., 2025),[6] which contains contrastive, synthetic MT errors in the COVID-19 domain. Each reference translation from the TICO-19 dataset (Anastasopoulos et al., 2020) is perturbed using eight linguistic perturbations, categorized as either minor or critical based on the potential real-world impact of the MT error. We recruit five bilingual annotators to independently annotate each MT output for gold shareability labels. Inter-annotator agreement, measured by Fleiss' Kappa[7], is moderate (0.449). We select 20 examples for the study based on high agreement scores determined by majority vote, with 10 examples per label. Further details are in Appendix B.1.[8]

[2] We compare the computational and time efficiency for generating each feedback in Appendix E.

[3] https://huggingface.co/Unbabel/XCOMET-XXL

[4] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[5] https://translate.google.com/

[6] Criteria for data selection are detailed in Appendix B.2.

[7] https://en.wikipedia.org/wiki/Fleiss_kappa

[8] English source sentences contain 471 words in total (average 23.6 words per sentence), and the Spanish translations contain 590 words in total (average 29.5 words per sentence).

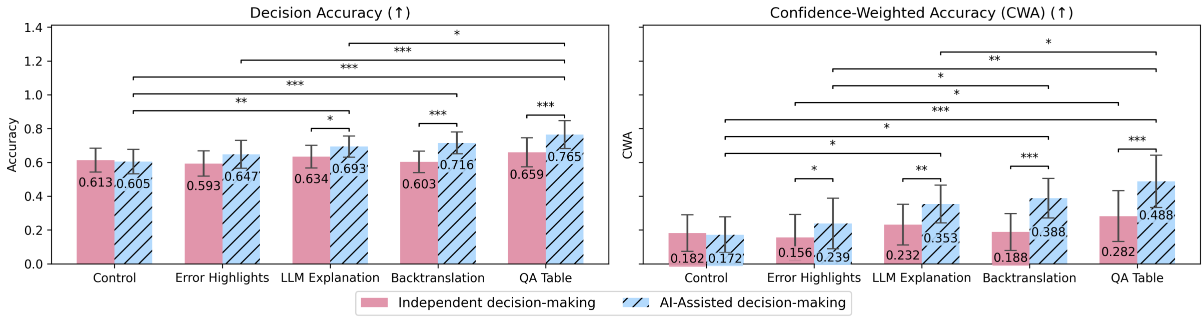


Figure 3: Average decision accuracy (*left*) and CWA (*right*) for each condition. Paired-sample t-tests are performed to compare independent and AI-assisted performance and linear mixed-effects ANOVA with Bonferroni corrections to compare different treatment conditions. *: significant with p-value < 0.05; **: $p < 0.01$; ***: $p < 0.001$; Non-marked ones are not statistically significant. Detailed results are provided in Appendix F.1.

3.4 Participants

We recruited 91 participants residing in the United States who self-identified English as their first, primary, and fluent language. Recruitment was conducted in two stages to exclude participants proficient in Spanish: **(1)** A pre-screening survey, where participants reported their English and Spanish proficiency on a five-point scale; and **(2)** The main task, limited to those who reported high English and low Spanish proficiency in the pre-screening phase. Each main task participant received 5 USD for completing the task (equivalent to 20 USD/hour),[9] and 30 participants who achieved over 70% overall decision accuracy received an additional 2 USD bonus. Our institution's IRB approved to conduct the study. Participants provided informed consent prior to the study. Further details are in Appendix C.2.

Of the 91 participants, 90 reported English as their first language, and one reported both Filipino and English. The average self-reported English proficiency was 5/5 and Spanish proficiency was 1.83/5.[10] Reported monthly MT usage varied: 5 participants (5.49%) never used MT, 24 (26.4%) rarely used it, 32 (35.2%) used it sometimes, 19 (20.9%) often, and 11 (12.1%) used MT almost every day. Data from one participant who failed both shareability checks was excluded from analysis. Participants were randomly assigned to one of five conditions, with 18 in each group.

3.5 Dependent Variables

Decision Accuracy. For each example e, we ask participants to decide whether the translation is of sufficient quality to safely share using a binary scale. We compute decision accuracy by comparing each participant's shareability judgment $\hat{s}$ against the gold label s^* for each example $e \in E$:

$$\textbf{DecisionAcc.}(E) = \frac{1}{E} \sum_{e \in E} \mathbb{1}(\hat{s} = s^*) \quad (1)$$

Confidence-Weighted Accuracy (CWA). Following Mehandru et al. (2023), we combine decision accuracy and confidence scores using confidence weighting (Ebel, 1965; Marshall et al., 2017) to evaluate whether participants made the correct decision weighted by their confidence in that decision. This metric serves as a measure of (in)appropriate reliance, where higher scores indicate accurate decisions made with well-calibrated confidence. Formally, for each example $e \in E$, we combine shareability $\hat{s}$ and confidence c as follows:

$$\begin{aligned} \textbf{CWA}(E) &= \frac{1}{E} \sum_{e \in E} \operatorname{sign}(\hat{s}) \cdot \frac{c}{5} \\ \operatorname{sign}(\hat{s}) &= \begin{cases} 1, & \text{if } \hat{s} = s^* \\ -1, & \text{otherwise} \end{cases} \end{aligned} \quad (2)$$

Switch Percentage. Switch percentage is a widely used behavioral measure of reliance, capturing how often participants change their decisions after viewing AI feedback (Srivastava et al., 2022; He et al., 2023b). In our context, it reflects how quality feedback influences final shareability judgments (Eckhardt et al., 2024). We compute three metrics following the framework of Schemmer et al. (2023): **(1) Over-reliance**: the proportion of cases where a participant changes from a correct to an incorrect decision after feedback; **(2) Under-reliance**: the proportion of cases where a participant does not change from an incorrect decision to a correct one after the quality feedback; **(3) Appropriate reliance**: the proportion of cases where a participant

[9] The task took a median of 14 minutes to complete.

[10] We detail language proficiency scale in Appendix C.

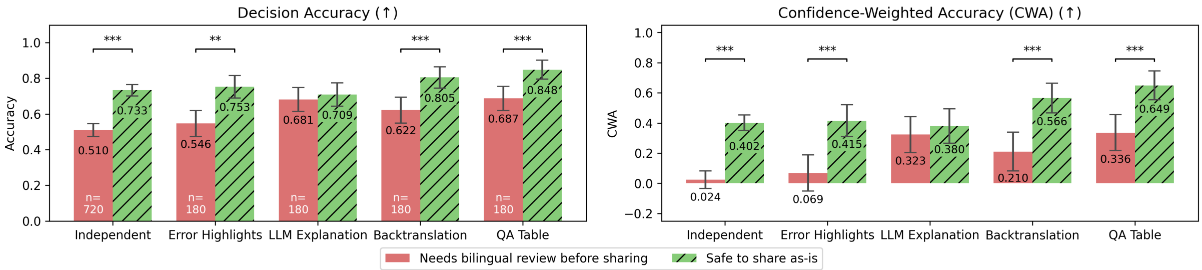


Figure 4: Average decision accuracy (*left*) and CWA (*right*) for each type of quality feedback and shareability label. **n** indicates the number of examples aggregated for each condition and label. **Independent** aggregates responses made without quality feedback across all conditions. **: statistically significant with p-value < 0.01; ***: p < 0.001; Non-marked ones are not statistically significant. Detailed results are provided in Appendix F.3.

either corrects an incorrect decision after receiving feedback (switch) or maintains a correct decision (no switch). To account for the differing consequences of reliance depending on shareability, we further break down each reliance metric by shareability label, as detailed in Appendix F.2.

4 Results

We begin by comparing independent and AI-assisted decision-making performance (§4.1). We then evaluate the four quality feedback types in detail, in terms of decision accuracy, CWA (§4.2), and switch percentage (§4.3).

4.1 RQ1: Does Quality Feedback Improve MT Decision-Making?

We perform paired-sample t-tests to compare independent and AI-assisted performance. As shown in Figure 3, participants in all four treatment conditions generally exhibit higher decision accuracy (*left*) and appropriate reliance, measured by CWA (*right*) in the AI-assisted decision-making step compared to the independent step. We observe statistically significant gains in average decision accuracy for LLM explanation (9.31%; $p < 0.05$), backtranslation (18.7%; $p < 0.001$), and QA table feedback (16.1%; $p < 0.001$), resulting in an overall average improvement of 8.32% across all conditions. CWA improves across all conditions with greater extent, averaging 15.3%, indicating that providing any quality feedback is more effective at helping participants make accurate decisions with well-calibrated confidence than at improving decision accuracy alone.

We further find that both decision accuracy (M = 0.605, S.E. = 0.036) and CWA (M = 0.172, S.E. = 0.054) in the AI-assisted step are significantly lower in the control condition than in all treatment conditions except for the error highlights group. Moreover, the within-group difference between the independent and AI-assisted steps in the control condition is not statistically significant. This indicates that our two-step setup does not induce learning effects but the observed gains in decision accuracy and appropriate reliance stem from the quality feedback, not from repetition.

4.2 RQ2: Which Feedback is Most Effective?

We perform linear mixed-effects ANOVA, followed by Bonferroni correction for multiple comparisons across treatment conditions. Implicit feedback types generally outperform explicit ones. Participants who received QA table feedback have significantly higher AI-assisted decision accuracy (M = 0.765, S.E. = 0.022) than those in the error highlights (M = 0.647, S.E. = 0.043; $p < 0.01$) and the LLM explanation group (M = 0.693, S.E. = 0.042; $p < 0.05$), as shown in Figure 3. No significant difference was found between QA table and backtranslation (M = 0.716, S.E. = 0.025).

A similar pattern emerges for appropriate reliance (CWA). QA table group achieved significantly higher CWA (M = 0.488, S.E. = 0.040) than error highlights (M = 0.239, S.E. = 0.071; $p < 0.01$) and LLM explanations (M = 0.353, S.E. = 0.070; $p < 0.05$). Backtranslation (M = 0.388, S.E. = 0.072; $p < 0.05$) also yielded significantly higher CWA than error highlights. These results suggest that QA table feedback is the most effective overall, outperforming both explicit feedback types in supporting accurate and well-calibrated MT decisions.

Participants' independent decision-making performance did not significantly differ across conditions, except for CWA, where QA table feedback (M = 0.282, S.E. = 0.040) significantly outperform error highlights (M = 0.156, S.E. = 0.041; $p < 0.05$).

Overall, our findings show that all three quality feedback types except error highlights significantly

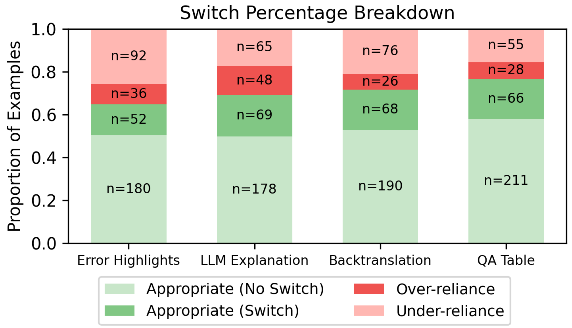


Figure 5: Breakdown of switch percentages by quality feedback type, showing appropriate, over-, and under-reliance.

improve both decision accuracy and appropriate reliance compared to no feedback, with implicit feedback types (backtranslation and QA table) showing stronger and more consistent statistical effects (§4.1). Among them, QA table consistently yields the greatest gains in both metrics (§4.2).

4.3 RQ3: Which Feedback do Users Rely on Most Appropriately?

For each quality feedback, we compute switch percentage to capture participants' behavioral patterns of reliance in Figure 5. Under-reliance is highest in the error highlights group (25.6%), followed by backtranslation (21.1%), LLM explanation (18.1%), and QA table (15.3%). Interestingly, implicit feedback types (QA table (7.78%) and backtranslation (7.22%)) yield lower over-reliance than explicit ones (LLM explanation (13.3%) and error highlights (10.0%)). Across all conditions, participants are more likely to maintain their initial decisions (regardless of correctness) than to change them, as under-reliance consistently exceeds over-reliance, and appropriate reliance (no switch) exceeds (switch).

We further examine switch percentages by shareability label in Appendix F.2. Participants show higher over-reliance and lower under-reliance for shareable examples than non-shareable ones.[11] This suggests that participants are more likely to change their decisions when initially judging translations as shareable, but tend to maintain their decisions when judging them as non-shareable.

Condition	Mental burden	Helpfulness	Trust
Control	5.83	2.89	-
Error Highlights	4.94	3.83	3.89
Explanation	4.06	**4.39**	4.11
Backtranslation	4.06	**4.39**	4.06
QA Table	**4.00**	**4.39**	**4.22**

Table 1: Average mental burden (1-7, ↓), helpfulness (1-5, ↑), and trust for future use (1-5, ↑) for each condition group. Best scores for each metric are **bold**.

5 Analysis

5.1 Shareable vs. Non-shareable MT

As shown in Figure 4, treatment condition participants consistently achieve significantly higher decision accuracy ($p < 0.01$ for error highlights; $p < 0.001$ for others) and CWA scores ($p < 0.001$) on shareable examples than on non-shareable ones. One exception is the LLM explanation group, where the differences are statistically not significant. This suggests that participants generally make more accurate and appropriate decisions when evaluating good translations than problematic ones, indicating that helping users reliably identify critical MT errors remains a challenge.

5.2 Self-reported Perception vs. Actual Performance

In the post-task survey, participants rated the quality feedback they received in terms of perceived mental burden, helpfulness, and trust for future use. As shown in Table 1, the control condition group reported highest mental burden (5.83) and lowest helpfulness (2.89), suggested that repeated exposure without any quality feedback increased cognitive load without enhancing perceived utility.

Among the four treatment conditions, QA table group reported the highest level of trust (4.22), aligning with findings from Section 4.2 that this feedback is most effective at improving appropriate reliance. In contrast, participants who received error highlights gave the lowest ratings for helpfulness (3.83) and trust (3.89), which is consistent with their relatively poor performance in both decision accuracy and appropriate reliance (CWA). Interestingly, while LLM explanations do not yield large gains in decision accuracy or CWA, they have relatively high ratings for helpfulness (4.39) and trust (4.11), which may reflect the over-reliance discussed in Section 4.3. In terms of mental burden, the error highlights group reported higher score (4.94) than other groups. We attribute this to the

[11] We refer to *shareable* as examples labeled “Safe to share as-is,” and *non-shareable* as those labeled “Needs bilingual review before sharing”.

nature of error highlights as a *target*-side feedback mechanism (Leiter et al., 2024), which displays highlights on the Spanish MT output, making it difficult to interpret for monolingual source speakers.

5.3 What Makes Quality Feedback *Helpful*?

We present treatment condition participants' responses on how they used quality feedback in their decision-making and which aspects they found helpful or unhelpful. For error highlights, two participants found them largely unhelpful, noting that the highlights were only shown on the Spanish MT, illustrating a key limitation of a target-side feedback. However, some appreciated the explicitness of the highlights, stating that they pointed to "*areas that not have been accurately translated*" or "*key translation mistakes*". Similarly, four participants valued LLM explanations for showing "*exactly what is correct or incorrect*". Some further appreciated for offering insights into alternative translations or contextual relevance.

In contrast, participants who received implicit feedback described a more self-directed decision-making process. Backtranslation was considered helpful for verifying "*some tiny unsure details*" or checking whether "*the core meaning of the original English text was preserved in the Spanish MT*" by comparing the two English texts. Similarly, QA table encouraged participants to revisit the MT output when a mismatch was detected ("*if the statement is blue, I double check the phrase again*"), "*compare the words and make determinations*" themselves. Detailed comments are provided in Appendix F.4.

6 Conclusion

We explore the utility of quality feedback in helping monolingual source speakers make reliable MT decisions. We conduct a between-subjects human study where participants decide whether Spanish MT outputs are safe to share, first independently and then with one of five conditions. The four treatment conditions include different types of MT quality feedback: two explicit (error highlights and LLM explanation) and two implicit (backtranslation and QA table).

We find that all feedback types except error highlights significantly improve decision accuracy and appropriate reliance (§4.1). Implicit feedback, especially QA table, outperforms explicit feedback in both objective performance and self-reported ratings (§4.2, §5.2) and while explicit feedback prompts more decision changes, it also increases over-reliance (§4.3). We further show that participants are better at confirming good translations than detecting problematic ones (§5.1).

Overall, our findings underscore the value of feedback that guides users' implicit interpretation rather than prescribing decisions. Implicit methods may be especially effective, as they preserve users' agency in the decision-making process (Savoldi et al., 2025; Xiao et al., 2025a). Along with insights into what participants found helpful across feedback types (§5.3), our work calls for further research exploring feedback to help users reliably identify critical MT errors in realistic use cases.

7 Limitation

Presentation differences. We acknowledge that presentation differences, such as formatting or visual salience, can influence user behavior independently of the underlying feedback type. To mitigate this, we iteratively designed and piloted each feedback interface to ensure clarity and minimize usability discrepancies across conditions. Specifically, we conducted seven rounds of internal pilot testing and usability checks to identify sources of confusion, refine language, or layout. While some format-specific differences were necessary (*e.g.,* error highlights inherently rely on color to convey span-level quality signals, while QA tables require side-by-side comparisons), we carefully calibrated to ensure that each condition represented a "best of its breed" version of that feedback type. Therefore, although we acknowledge the possibility of minor usability-related effects, we believe these differences are unlikely to fully account for the performance gaps observed between feedback types.

Study with monolingual target speakers. Our human study focuses on monolingual source speakers who self-identified as proficient in English but not in Spanish, simulating a scenario in which non-English speakers encounter MT outputs of COVID-19 articles originally written in English. An important complementary study remains – evaluating with the monolingual target speakers. This would require modifying our current quality feedback setup: **(1)** Error highlights would continue to be displayed on the MT output; **(2)** LLM explanations would need to be presented in the target language; **(3)** Backtranslation would no longer be a suitable feedback; and **(4)** For QA table, questions would need to be generated from the backtranslated source, with answers derived from both the backtranslation and the (possibly perturbed) MT output. Future work can explore this variant through a human-centered study to assess how different forms of quality feedback influence MT decision-making for monolingual target speakers.

Limited scope. The scope of our study is limited to the current experimental setup. While we focus on the English-Spanish language pair, motivated by Spanish being the most widely spoken non-English language in the United States (U.S. Census Bureau, 2022), our findings may not generalize to other language pairs. For consistency and fair comparison, we use the same model (LLAMA-3.3 70B) to generate both LLM explanations and QA table feedback, and the same MT system (Google Translate) to produce backtranslations and the backtranslated MT outputs used in QA generation. Our evaluation is also limited to four types of quality feedback: error highlights, LLM explanations, backtranslation, and QA table.

Moreover, we intentionally focused on a single domain (communicating COVID-19 protocols) and a specific user population with specific language background in order to ensure that our study reflects realistic decision-making scenarios. This controlled setting helped participants better operationalize the notion of shareability and limited potential confounding factors in the study. We believe the core findings ought to generalize to other language pairs since MT quality was not a major factor, and possibly other domains with similar notions of shareability.

Understanding the effect of quality feedback on MT reliance across a wider range of contexts is an important direction of future work. For instance, the doctor-patient setting represents a valuable use case where the doctor as a user of MT has more consideration and technical expertise in deciding whether MT outputs are shareable. Another example could be dialogues, where additional rounds of human and automatic feedback is possible. We view our study as a foundational step toward more extensive future research in this area.

Use of synthetic dataset. While our study aimed to simulate a realistic decision-making scenario, the MT outputs themselves are drawn from a synthetically constructed dataset CONTRATICO. This design choice allowed us to systematically control for specific error types and severity levels. However, we acknowledge that these errors may not fully capture the complexity and variability of naturally occurring MT errors.

Ethics Statement

This study was approved by our university's Institutional Review Board (IRB). All participants provided informed consent prior to participation and were compensated according to the rates specified in the consent form. Additionally, participants who achieved an overall accuracy above 70% received a performance-based incentive.

Acknowledgements

We would like to thank the members of the CLIP lab at University of Maryland for their constructive feedback and support. Specifically, we are grateful to Hiba El Oirghi, Yu Hou, HyoJung Han, Kartik Ravisankar, Nishant Balepur, Navita Goyal, and Calvin Bao for their help in testing the user interface. We especially appreciate Yimin Xiao for extensive discussions on the statistical analysis in the paper. This work was supported in part by the Human Language Technology Center of Excellence at Johns Hopkins University, by NSF Fairness in AI Grant 2147292, 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

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via backtranslation at the WMT 2022 quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 593–596, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part*

2) at EMNLP 2020, Online. Association for Computational Linguistics.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA. Association for Computing Machinery.

Lynne Bowker and Jairo Buitrago Ciro. 2019. 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, Leeds.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

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. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE.

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.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications.

Karl de Fine Licht and Bengt Brülde. 2021. On defining "reliance" and "trust": Purposes, conditions of adequacy, and new definitions. *Philosophia*, 49:1981–2001.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Robert L Ebel. 1965. Confidence weighting and test reliability. *Journal of Educational Measurement*, 2(1):49–57.

Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. 2024. A survey of AI reliance. *Preprint*, arXiv:2408.03948.

Melda Eksi, Erik Gelbing, Jonathan Stieber, and Chi Viet Vu. 2021. Explaining errors in machine translation with absolute gradient ensembles. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 238–249, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi,

Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind ThattAI, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris CAI, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, ItAI Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, KAI Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin

Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, SAI Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, VijAI Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 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. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

HyoJung Han, Marine Carpuat, and Jordan Boyd-Graber. 2022. SimQA: Detecting simultaneous MT errors through word-by-word question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023a. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023b. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for explainable AI: Challenges and prospects. *Preprint*, arXiv:1812.04608.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025. 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.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *Preprint*, arXiv:1902.00006.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-AI decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.

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.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.

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, page 1–13, New York, NY, USA. Association for Computing Machinery.

Tania Lombrozo and Daniel A. Wilkenfeld. 2019. *Mechanistic versus functional understanding*, chapter 11. New York, NY: Oxford University Press.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

ShuAI Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. 2024. Beyond recommender: An exploratory study of the effects of different AI roles in AI-assisted decision making. *Preprint*, arXiv:2403.01791.

James A.R. Marshall, Gavin Brown, and Andrew N. Radford. 2017. Individual confidence-weighting and group decision-making. *Trends in Ecology & Evolution*, 32(9):636–645.

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.

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.

Jeff Pitman. 2021. Google translate: One billion installs, one billion stories. Engineering Manager, Google Translate.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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)*, pages 2685–2702, Online. Association for Computational Linguistics.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422.

Anuschka Schmitt, Thiemo Wambsganss, Matthias Soellner, and Andreas Janson. 2021. Towards a trust reliance paradox? exploring the gap between perceived trust in and reliance on algorithmic advice. In *Proceedings of the International Conference on Information Systems (ICIS)*, number 14 in ICIS 2021 Proceedings.

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–18. ACM.

Divya K. Srivastava, J. Mason Lilly, and Karen M. Feigh. 2022. Improving human situation awareness in AI-advised decision making. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, pages 1–6.

Kyoshiro Sugiyama, Masahiro Mizukami, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. An investigation of machine translation evaluation metrics in cross-lingual question answering. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 442–449, Lisbon, Portugal. Association for Computational Linguistics.

U.S. Census Bureau. 2022. What languages do we speak in the united states? Accessed: 2025-05-14.

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).

Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Yimin Xiao, Cartor Hancock, Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, Marine Carpuat, and Ge Gao. 2025a. 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.

Yimin Xiao, Yongle Zhang, Dayeon Ki, Calvin Bao, Marianna J. Martindale, Charlotte Vaughn, Ge Gao, and Marine Carpuat. 2025b. Beyond Benchmarks: Exploring Machine Translation Error Perception and Reliance Among the General Public. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

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.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. 2024. Explanation in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico. Association for Computational Linguistics.

Vilém Zouhar and Ondřej Bojar. 2020. Outbound translation user interface ptakopět: A pilot study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6967–6975, Marseille, France. European Language Resources Association.

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 Prompts

A.1 LLM Explanation

We show prompt used for generating explanations with LLAMA-3.3 70B (Grattafiori et al., 2024).

Prompt A.1: LLM Explanation

Task: Your task is to evaluate the quality of the Spanish translation of the English sentence. Give your explanation in less than 3 sentences.

English sentence: `{source}`
Spanish translation: `{target}`
Explanation:

A.2 Question-Answer (QA) Table

We use the same prompts for both question generation (A.2.1) and answering (A.2.2) as those used in ASKQE (Ki et al., 2025).

Prompt A.2.1: Question Generation (QG)

Task: You will be given an English sentence and a list of atomic facts, which are short sentences conveying one piece of information. Your goal is to generate a list of relevant questions based on the sentence. Output the list of questions in Python list format without giving any additional explanation.

*** Example Starts ***
Sentence: It was declared a pandemic by the World Health Organization (WHO) on 11 March 2020.
Atomic facts: [“It was declared a pandemic on 11 March 2020.”, “The World Health Organization (WHO) declared it a pandemic.’]
Questions: [“What was declared on 11 March 2020?", "Who declared it a pandemic?”]

Sentence: The number of accessory proteins and their function is unique depending on the specific coronavirus.
Atomic facts: [“The number of accessory proteins is unique depending on the specific coronavirus.”, “The function of accessory proteins is unique depending on the specific coronavirus.”]
Questions: [“What is unique depending on the specific coronavirus?”, “What is unique about the function of accessory proteins?”]
*** Example Ends ***

Sentence: `{sentence}`
Atomic facts: `{atomic facts}`
Questions:

Prompt A.2.2: Question Answering (QA)

Task: You will be given an English sentence and a list of relevant questions. Your goal is to generate a list of answers to the questions based on the sentence. Output only the list of answers in Python list format without giving any additional explanation.

*** Example Starts ***
Sentence: Some patients have very mild symptoms, similar to a cold.
Questions: [“What kind of symptoms do some patients have?”, “What are the symptoms similar to?”]
Answers: [“Very mild symptoms”, “A cold”]

Sentence: Diabetes mellitus (784, 10.9%), chronic lung disease (656, 9.2%), and cardiovascular disease (647, 9.0%) were the most frequently reported conditions among all cases.
Questions: [“What were the most frequently reported conditions among all cases?”, “What percentage of cases reported diabetes mellitus?”, “What percentage of cases reported chronic lung disease?”, “What percentage of cases reported cardiovascular disease?”]
Answers: [“Diabetes mellitus, chronic lung disease, and cardiovascular disease”, “10.9%”, “9.2%”, “9.0%”]
*** Example Ends ***

Sentence: `{sentence}`
Questions: `{questions}`
Answers:

B Dataset Details

B.1 Gold Annotation

We provide details of the gold annotation process used to collect gold shareability labels. We use Qualtrics[12] to design the survey and Prolific[13] to recruit annotators fluent in both English and Spanish. A total of 40 examples, each consisting of an English sentence and its Spanish translation, were presented in randomized order. As shown in Figure 6, annotators were asked to judge whether the Spanish translation was **(1)** Safe to share as-is or **(2)** Needs bilingual review before sharing. The survey took a median completion time of 30 minutes. We recruited 5 annotators and compensated each with 8 USD (equivalent to 16 USD/hour).

We select 20 examples with high agreement scores (based on majority vote) for use in the main task. The final set includes 10 examples per label. The average agreement score is 1.0 for examples labeled as “Safe to share as-is” (shareable) and 0.8 for those labeled as “Needs bilingual review before sharing” (non-shareable). The shareable set includes 5 non-error examples and 5 minor error examples, comprising 2 synonym, 2 word order, and 1 spelling error, based on the error taxonomy

[12] `https://www.qualtrics.com`
[13] `https://www.prolific.com`

in Ki et al. (2025). The non-shareable set includes 1 minor error (intensifier) and 9 critical errors: 6 alteration, 2 omission, and 1 expansion with impact error examples.

B.2 Data Selection Criteria

We ensure the selected examples from the ContraTICO dataset (Ki et al., 2025) meet the following criteria before running the gold annotation: **(1)** Examples have a balanced distribution across three error severity levels: no error, minor errors, and critical errors in MT; **(2)** Examples have comparable lengths across error severity levels to minimize the influence of sentence length on participants' confidence (Zouhar and Bojar, 2020); **(3)** Examples are relevant to the scenario context. We focus on the Wikivoyage[14] subset of the dataset, which contains announcements and protocols related to COVID-19 (e.g., "It was declared a pandemic by the World Health Organization (WHO) on 11 March 2020.") instead of informal conversations (e.g., "and does this pain move from your chest?") or sentences with highly technical terms (e.g., "Like all coronaviruses, virions consist of single-stranded positive-sense RNA enclosed within an envelope.").

C Study Design Details

We present details about our human study design.

C.1 Annotation Interface

We built a custom annotation interface, with screenshots shown in Figure 7 following the task flow: **(1)** Consent to Participate, **(2)** Pre-task survey, **(3)** Task instructions and compensation details, **(4)** Tutorial, **(5)** Independent decision-making task, **(6)** AI-assisted decision-making task, and **(7)** Post-task survey. Participants were required to answer all pre- and post-task survey questions. We present an interactive tutorial before the main task to ensure that participants understand the questions they will be asked for each example. We also illustrate how each type of quality feedback was presented during the AI-assisted decision-making step.

Pre-Task Survey. The pre-task survey includes the following four questions:

- **First language:** What is your first language (or languages)?
- **Proficiency in English:** What is your level of proficiency in English?
    - ○ I cannot understand any English words or sentences at all.
    - ○ I can read some English words and very simple sentences.
    - ○ I can read short, simple texts in English, such as messages from friends.
    - ○ I can read English texts about everyday life, such as short novels or news articles.
    - ○ I can read long and difficult texts in English, such as opinion essays or scientific papers, without help.
- **Proficiency in Spanish:** What is your level of proficiency in Spanish?
    - ○ I cannot understand any English words or sentences at all.
    - ○ I can read some English words and very simple sentences.
    - ○ I can read short, simple texts in English, such as messages from friends.
    - ○ I can read English texts about everyday life, such as short novels or news articles.
    - ○ I can read long and difficult texts in English, such as opinion essays or scientific papers, without help.
- **AI Translation Tool Usage:** In the past month, how often did you use AI translation tools (e.g., Google Translate, ChatGPT)?
    - ○ Never: Never in the past month
    - ○ Rarely: Fewer than once a week
    - ○ Sometimes: Two or three times a week
    - ○ Often: More than three times a week, but not every day
    - ○ Always: Almost every day

Post-Task Survey (Control). In the post-task survey for the control condition, we include the following two questions:

- **Mental Workload:** How much mental burden (e.g., thinking, deciding) did you experience while doing the task?
    - ○ 1 (Low)
    - ○ ...
    - ○ 7 (High)
- **Helpfulness:** How helpful was seeing the same example twice in your Spanish translation quality assessment?
    - ○ 1 (Very unhelpful)
    - ○ 2 (Unhelpful)
    - ○ 3 (Neutral)
    - ○ 4 (Helpful)
    - ○ 5 (Very helpful)

Post-Task Survey (Treatment). In the post-task survey for treatment conditions, we dynamically replace FEEDBACK with the participant's assigned quality feedback type. The survey includes the following three questions:

- **Mental Workload:** How much mental burden

[14] https://www.wikivoyage.org/

(e.g., thinking, deciding) did you experience while using the FEEDBACK?
  - 1 (Low)
  - ...
  - 7 (High)
- **Helpfulness:** How helpful was the FEEDBACK in assisting the Spanish translation quality assessment?
  - 1 (Very unhelpful)
  - 2 (Unhelpful)
  - 3 (Neutral)
  - 4 (Helpful)
  - 5 (Very helpful)
- **Trust for Future Use:** Would you use the FEEDBACK again in the future?
  - 1 (Very unlikely)
  - 2 (Unlikely)
  - 3 (Neutral)
  - 4 (Likely)
  - 5 (Very likely)

Depending on the participant's response to the **Helpfulness** question, we present an optional follow-up question:

- **If the response is 1 or 2:** In what ways was the information provided by the FEEDBACK confusing or unhelpful?
- **If the response is 3:** How did you use the information provided by the FEEDBACK in your assessment?
- **If the response is 4 or 5:** In what ways was the information provided by the FEEDBACK helpful?

C.2 Recruitment Process

Pre-screening Survey. We conducted our user studies on the Prolific platform. For the pre-screening survey, we used the same language proficiency questions as in the pre-task survey (Appendix C) to reliably recruit monolingual English speakers. To ensure high-quality responses, we limited participation to users with a Prolific approval rate above 90% and at least 10 prior submissions. We recruited 205 participants and invited 123 who selected "I can read long and difficult texts in English without help" for English proficiency, and either "I cannot understand any Spanish words or sentences at all" or "I can read some Spanish words and very simple sentences, such as greetings and common expressions" for Spanish proficiency. Each pre-screening participant received 0.20 USD, totaling 55.20 USD including Prolific platform fees.

Main task. Of the 123 invited participants, 91 completed the main task. Each received a base payment of 5 USD for 20 minutes of participation, with an additional 2 USD performance-based bonus for those achieving over 70% overall decision accuracy. A total of 30 participants qualified for this bonus. One participant who failed both attention check questions was compensated but excluded from the final analysis, resulting in 90 valid responses. The median task completion time was 14 minutes, corresponding to an effective pay rate of 20 USD per hour. The total cost for the main task, including Prolific fees, was 588.01 USD.

D Inferring AI Feedback Decisions

During dataset selection, we ensure that each feedback condition includes a relatively balanced number of examples reflecting both correct and incorrect AI predictions. Since none of the feedback types provide a direct prediction of shareability judgment, we apply tailored proxy measures to approximate the intended shareability label for each feedback type, as detailed below.

Error Highlights. For each error annotation, we take the highest error severity level h (No error, Minor, Major, or Critical) as the QE model's decision. The decision is mapped to the two-point shareability scale as follows:

$$\begin{cases} \text{✅ Safe to share as-is,} & h \in \{\text{No error}, \text{Minor}\} \\ \text{❌ Needs bilingual review,} & h \in \{\text{Major}, \text{Critical}\} \end{cases} \quad (3)$$

LLM Explanation. We prompt the same model used to generate textual explanations (LLaMA-3.3 70B) to make shareability judgments using the same information and option format provided to participants in the task instructions. Exact prompt used is outlined in Appendix A.1.

Backtranslation. Following Ki et al. (2025), we compute XCOMET-QE scores between the Spanish MT and its backtranslation. We fit a three-component Gaussian Mixture Model (GMM).[15] GMM is a probabilistic clustering model that assumes data points are generated from a mixture of Gaussian distributions, assigning a probability to each point for belonging to a specific cluster. We hypothesize two clusters corresponding to different ranges of QE scores: **(1)** ✅ "Safe to share as-is" cluster with higher QE scores and **(2)** ❌ "Needs

[15] https://scikit-learn.org/stable/modules/mixture.html

Feedback	Model (*size*)	Computation	Avg. Time
Error Highlights	xCOMET-XXL (10.7B)	× 3	03:52
LLM Explanation	LLAMA-3.3 (70B)	× 8	01:56
Backtranslation	Google Translate	-	01:24
QA Table	LLAMA-3.3 (70B)	× 8	03:48

Table 2: Average computational and time efficiency for four types of quality feedback. **Computation** is reported in GPU units (RTX A5000). **Avg. Time** is shown in MM:SS.

bilingual review before sharing" cluster with lower QE scores. Each of the 20 examples is assigned a probability to each cluster and is categorized into the cluster with the highest probability.

QA Table. We follow a similar procedure to that used for backtranslation. However, instead of using QE scores, we compute ASKQE scores through a two-step process: (1) measuring overlap between answers from the source A_{src} and from the backtranslated MT output A_{bt}, and (2) aggregating question-answer similarities into a segment-level metric. We use F1, a standard similarity measure in QA research (Rajpurkar et al., 2016; Deutsch et al., 2021). Formally, for each example $e \in E$, we compute:

$$\text{AskQE}(e) = \sum_{i=1}^{E} \frac{\text{F1}(A_{\text{src}}, A_{\text{bt}})}{E} \qquad (4)$$

We fit a two-component Gaussian Mixture Model (GMM) to the ASKQE scores and assign each example to the cluster with the highest probability.

The selected examples were then manually reviewed by the authors to ensure that the proxy measures provided a reasonable approximation of the feedback's implied shareability judgments. This process resulted in a relatively balanced distribution of correct and incorrect examples across feedback types: we included 9, 10, 11, and 9 correct examples for the Error Highlights, LLM Explanation, Backtranslation, and QA Table conditions, respectively (with corresponding incorrect counts of 11, 10, 9, and 10). This helped ensure that each feedback condition reflected a comparable underlying error rate.

E Computational & Time Efficiency

We compare the average computational and time efficiency of generating four types of quality feedback across 20 examples chosen from the gold annotation process. As shown in Table 2, all feedback demonstrate sufficient efficiency for deployment in user-facing applications.

F Detailed Results

F.1 Independent vs. AI-Assisted

We present detailed results for both the independent and AI-assisted decision-making steps in terms of decision accuracy and CWA in Table 5. We show that the QA table feedback yields the highest overall and AI-assisted performance across both metrics, while error highlights have the lowest scores.

F.2 Switch Percentage

To account for the differing consequences of reliance relative to shareability, we first define four outcome types based on the participant's shareability judgment $\hat{s}$ and the gold label s^*:

- **True Positive (TP)**: s^* is shareable, $\hat{s} = s^*$
- **False Positive (FP)**: s^* is shareable, $\hat{s} \neq s^*$
- **True Negative (TN)**: s^* is non-shareable, $\hat{s} = s^*$
- **False Negative (FN)**: s^* is non-shareable, $\hat{s} \neq s^*$

Using this, we categorize decision transitions as shown in Table 3: **(1)** Over-reliance: changing from a correct to an incorrect decision after feedback (TP → FP and TN → FN); **(2)** Under-reliance: maintaining an incorrect decision after feedback (FP → FP and FN → FN); **(3)** Appropriate reliance: either maintaining a correct decision (TP → TP and TN → TN) or correcting an incorrect one after feedback (FP → TP and FN → TN).

	TP	FP	TN	FN
TP	Appropriate	Over	-	-
FP	Appropriate	Under	-	-
TN	-	-	Appropriate	Over
FN	-	-	Appropriate	Under

Table 3: Categorization of decision transitions from Independent (*rows*) to AI-Assisted (*columns*). **Appropriate:** Appropriate reliance; **Over:** Over-reliance; **Under:** Under-reliance.

In Table 4, we present detailed switch percentage results for each quality feedback type, broken down into each constituent of under-, over-, and appropriate reliance. Across all conditions, participants show higher over-reliance and lower under-reliance for shareable examples than non-shareable ones.

F.3 Per Shareability Label

We show detailed decision accuracy and CWA scores by shareability label (✅ Safe to share as-is and ❌ Needs bilingual review before sharing) in Table 6. Across all conditions, both decision accuracy and CWA are consistently higher for examples

Feedback	Reliance	Transition	Value (%)
Error Highlights	Appropriate reliance	TP → TP	113 (31.4)
		FP → TP	23 (6.39)
		TN → TN	67 (18.6)
		FN → TN	29 (8.06)
	Over-reliance	TP → FP	20 (5.56)
		TN → FN	16 (4.44)
	Under-reliance	FP → FP	26 (7.22)
		FN → FN	66 (18.3)
LLM Explanation	Appropriate reliance	TP → TP	100 (27.8)
		FP → TP	24 (6.67)
		TN → TN	78 (21.7)
		FN → TN	45 (12.5)
	Over-reliance	TP → FP	31 (8.61)
		TN → FN	17 (4.72)
	Under-reliance	FP → FP	25 (6.94)
		FN → FN	40 (11.1)
Backtranslation	Appropriate reliance	TP → TP	115 (31.9)
		FP → TP	31 (8.61)
		TN → TN	75 (20.8)
		FN → TN	37 (10.3)
	Over-reliance	TP → FP	10 (2.78)
		TN → FN	16 (4.44)
	Under-reliance	FP → FP	24 (6.67)
		FN → FN	52 (14.4)
QA Table	Appropriate reliance	TP → TP	123 (34.2)
		FP → TP	29 (8.06)
		TN → TN	86 (23.9)
		FN → TN	39 (10.8)
	Over-reliance	TP → FP	16 (4.44)
		TN → FN	12 (3.33)
	Under-reliance	FP → FP	11 (3.06)
		FN → FN	44 (12.2)

Table 4: Detailed switch percentage results by quality feedback type and reliance (appropriate, over-, under-reliance).

labeled as shareable, indicating that participants make more accurate and better-calibrated decisions for good translations than for those requiring bilingual review.

F.4 Participant Responses

We present detailed participants' comments of what aspects of the quality feedback they received to be helpful or unhelpful. This question varied depending on their rating of helpfulness, as shown in Appendix C.

Feedback	Decision Acc. (Total)	Decision Acc. (Ind.)	Decision Acc. (AI)	CWA (Total)	CWA (Ind.)	CWA (AI)
Control	0.609 [0.558, 0.660]	0.613 [0.542, 0.684]	0.605 [0.533, 0.678]	0.177 [0.101, 0.253]	0.182 [0.074, 0.290]	0.172 [0.066, 0.278]
Error Highlights	0.620 [0.547, 0.693]	0.593 [0.518, 0.668]	0.647 [0.565, 0.730]	0.197 [0.063, 0.331]	0.156 [0.020, 0.292]	0.239 [0.089, 0.388]
LLM Explanation	0.664 [0.610, 0.717]	0.634 [0.567, 0.701]	0.693 [0.631, 0.756]	0.292 [0.193, 0.392]	0.232 [0.111, 0.352]	0.353 [0.241, 0.465]
Backtranslation	0.659 [0.603, 0.715]	0.603 [0.540, 0.666]	0.716 [0.683, 0.847]	0.288 [0.188, 0.389]	0.188 [0.079, 0.298]	0.388 [0.272, 0.505]
QA Table	**0.712** [0.639, 0.786]	**0.659** [0.573, 0.746]	**0.765** [0.683, 0.847]	**0.385** [0.252, 0.519]	**0.282** [0.131, 0.433]	**0.488** [0.333, 0.643]

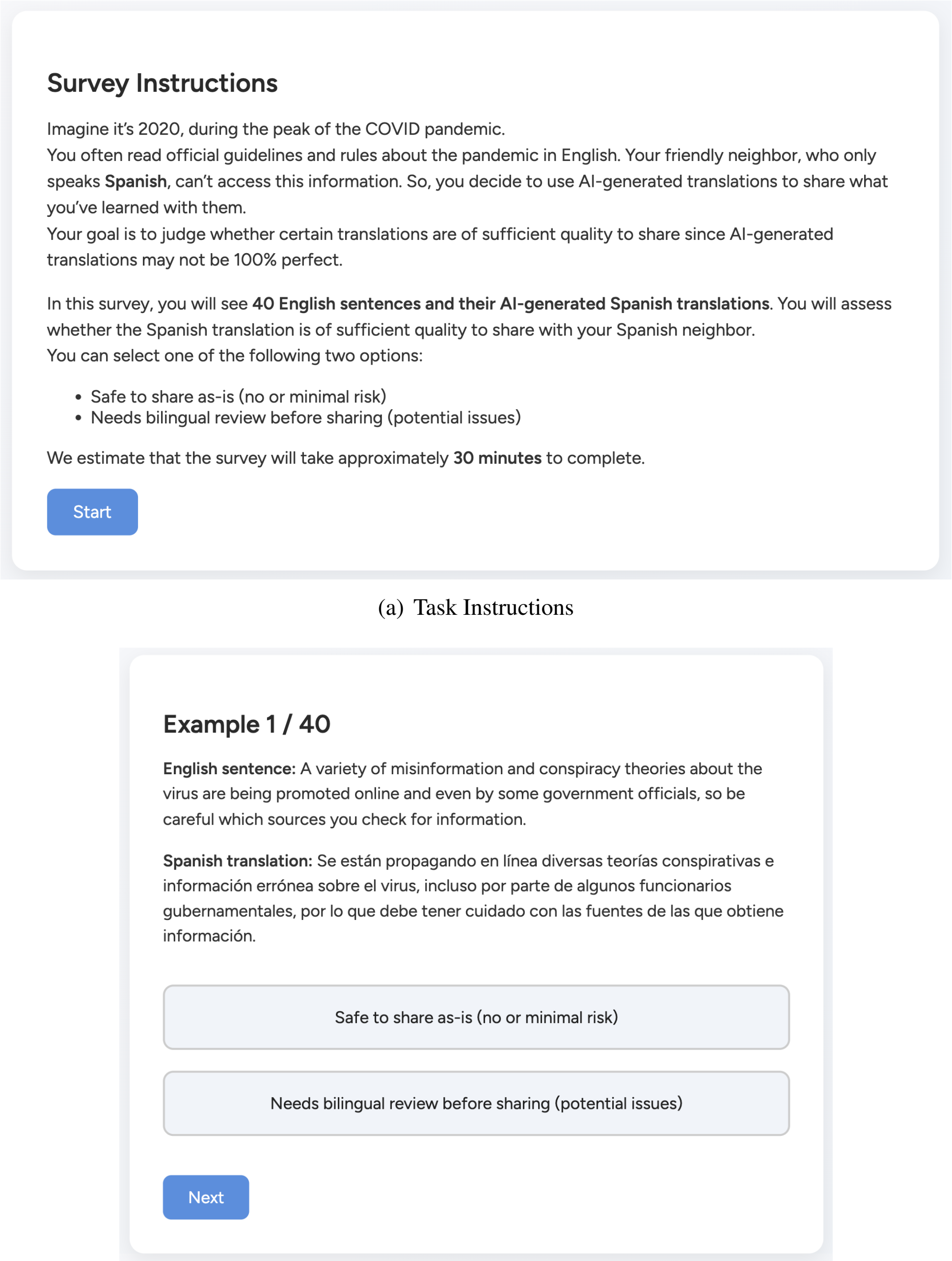
Table 5: Detailed results for decision accuracy and CWA by condition. **Ind.:** Independent decision-making step; **AI:** AI-assisted decision-making step. Values represent mean scores with corresponding 95% confidence intervals. Best scores for each column is **bold**.

Step	Feedback	Decsion Acc. (✅)	Decision Acc. (❌)	CWA (✅)	CWA (❌)
Independent	-	0.733 [0.701, 0.765]	0.509 [0.473, 0.546]	0.402 [0.351, 0.453]	0.024 [-0.034, 0.081]
AI-Assisted	**Error Highlights**	0.753 [0.690, 0.816]	0.546 [0.474, 0.619]	0.415 [0.310, 0.520]	0.069 [-0.051, 0.189]
	LLM Explanation	0.709 [0.644, 0.774]	0.681 [0.614, 0.748]	0.380 [0.266, 0.494]	0.323 [0.204, 0.443]
	Backtranslation	0.805 [0.746, 0.864]	0.622 [0.549, 0.695]	0.566 [0.467, 0.664]	0.210 [0.082, 0.339]
	QA Table	**0.848** [0.795, 0.901]	**0.687** [0.619, 0.755]	**0.649** [0.554, 0.745]	**0.335** [0.217, 0.455]

Table 6: Detailed results for decision accuracy and CWA by quality feedback type and shareability label. ✅: Safe to share as-is; ❌: Needs bilingual review before sharing. Values represent mean scores with corresponding 95% confidence intervals. Best scores for each column is **bold**.

Feedback	Participant Responses
🖍	• The fact that it doesn't tell me what the highlighted translated words mean makes it useless. When it tells me certain things are translated incorrectly but doesn't tell me what the meaning of the translation is, that's so pointless. (👎) • Since I **don't understand Spanish** I can't even judge how wrong/right the error detection is. I can't see this being helpful to anyone except people who already know Spanish. (👎) • It gave me the suggestions about the information **I could not understand**. (👎) • It helped highlight **areas that not have been accurately translated**. (👍) • It let me know where to look, and if it thought there were errors. (👍) • It highlighted **key translation mistakes**, helping to ensure the translation was accurate before sharing. (👍)
🤖	• Provided **exactly** why it was incorrect or correct. (👍) • **Gave me the answers.** I don't know Spanish. (👍) • It has enabled to know whether the translation was accurate or not. (👍) • The information provided by explanation helped me **understand the context of what was being said** more clearly. (👍) • By explaining what was wrong with the translation and putting what would be the **correct translation**. (👍) • It clarified some phrasing that was clearly a **direct translation versus a natural translation**. (👍)
A文	• Since I'm not very fluent in Spanish, the AI-translation was helpful in **verifying some tiny unsure details**.(👍) • I made a variety of mistakes when attempting to translate myself, the AI was helpful in translating particular words which made a **major difference in the meaning** of the sentence. (👍) • It provides insight into whether the **core meaning and intent of the original text are preserved** in the translation. (👍) • The information **simplified** what was said and made it very easy to understand. (👍) • Since I'm not very fluent in Spanish, the AI translation was helpful in verifying some tiny unsure details. (👍)
QA	• If the statement is blue, I **double check** the phrase again. (👍) • I could compare the words and **make determinations**. (👍) • It was helpful in terms of giving a **breakdown** of the Spanish translation. (👍) • It provided some clarity in showing that there was a **difference in the translation**. (👍) • I liked seeing the orange and blue so I knew if similar or different. These helped when **slight differences** such as when things were mild or the other said severe. (👍) • There were **small but seriously important words** I missed that when giving advice on an issue that could mean life or death matter. (👍) • It helped me know what information was being displayed on the Spanish side. (👍) • I would like this to be included for future surveys. (👍)

Table 7: Participants' responses on how they used quality feedback in their decision-making process, specifically why the feedback was perceived as confusing or unhelpful (👎) or why it was considered helpful (👍).



Survey Instructions

Imagine it's 2020, during the peak of the COVID pandemic.
You often read official guidelines and rules about the pandemic in English. Your friendly neighbor, who only speaks **Spanish**, can't access this information. So, you decide to use AI-generated translations to share what you've learned with them.
Your goal is to judge whether certain translations are of sufficient quality to share since AI-generated translations may not be 100% perfect.

In this survey, you will see **40 English sentences and their AI-generated Spanish translations**. You will assess whether the Spanish translation is of sufficient quality to share with your Spanish neighbor.
You can select one of the following two options:

- Safe to share as-is (no or minimal risk)
- Needs bilingual review before sharing (potential issues)

We estimate that the survey will take approximately **30 minutes** to complete.

(a) Task Instructions

Example 1 / 40

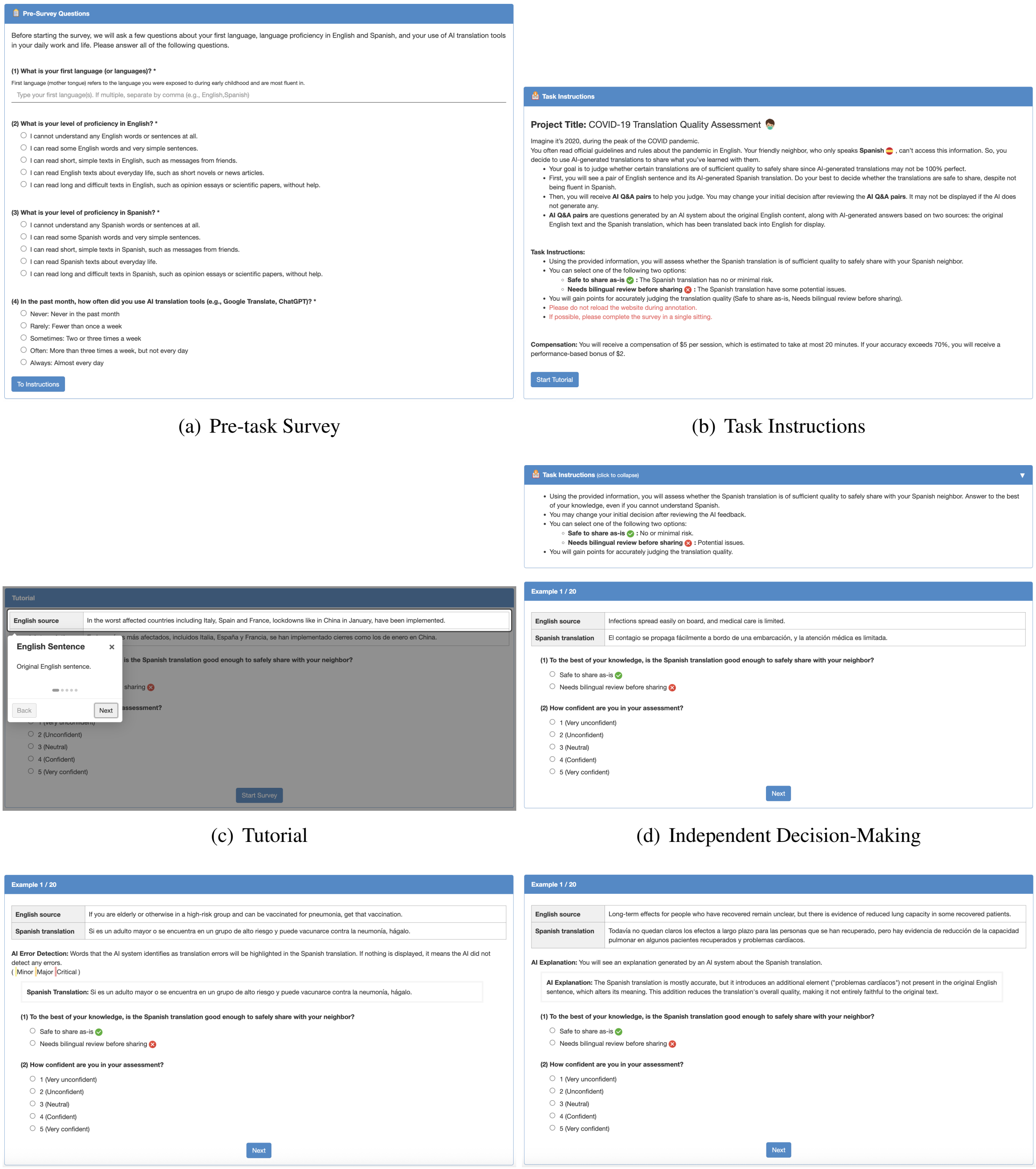
English sentence: A variety of misinformation and conspiracy theories about the virus are being promoted online and even by some government officials, so be careful which sources you check for information.

Spanish translation: Se están propagando en línea diversas teorías conspirativas e información errónea sobre el virus, incluso por parte de algunos funcionarios gubernamentales, por lo que debe tener cuidado con las fuentes de las que obtiene información.

Safe to share as-is (no or minimal risk)
Needs bilingual review before sharing (potential issues)
Next

(b) Example Question

Figure 6: Screenshots of the instructions provided to bilingual annotators, along with an example question.



Pre-Survey Questions

Before starting the survey, we will ask a few questions about your first language, language proficiency in English and Spanish, and your use of AI translation tools in your daily work and life. Please answer all of the following questions.

(1) What is your first language (or languages)? *

First language (mother tongue) refers to the language you were exposed to during early childhood and are most fluent in.

Type your first language(s). If multiple, separate by comma (e.g., English,Spanish)

(2) What is your level of proficiency in English? *

○ I cannot understand any English words or sentences at all.
○ I can read some English words and very simple sentences.
○ I can read short, simple texts in English, such as messages from friends.
○ I can read English texts about everyday life, such as short novels or news articles.
○ I can read long and difficult texts in English, such as opinion essays or scientific papers, without help.

(3) What is your level of proficiency in Spanish? *

○ I cannot understand any Spanish words or sentences at all.
○ I can read some Spanish words and very simple sentences.
○ I can read short, simple texts in Spanish, such as messages from friends.
○ I can read Spanish texts about everyday life.
○ I can read long and difficult texts in Spanish, such as opinion essays or scientific papers, without help.

(4) In the past month, how often did you use AI translation tools (e.g., Google Translate, ChatGPT)? *

○ Never: Never in the past month
○ Rarely: Fewer than once a week
○ Sometimes: Two or three times a week
○ Often: More than three times a week, but not every day
○ Always: Almost every day

To Instructions

(a) Pre-task Survey

Task Instructions

Project Title: COVID-19 Translation Quality Assessment

Imagine it's 2020, during the peak of the COVID pandemic.
You often read official guidelines and rules about the pandemic in English. Your friendly neighbor, who only speaks **Spanish** , can't access this information. So, you decide to use AI-generated translations to share what you've learned with them.

- Your goal is to judge whether certain translations are of sufficient quality to safely share since AI-generated translations may not be 100% perfect.
- First, you will see a pair of English sentence and its AI-generated Spanish translation. Do your best to decide whether the translations are safe to share, despite not being fluent in Spanish.
- Then, you will receive **AI Q&A pairs** to help you judge. You may change your initial decision after reviewing the **AI Q&A pairs**. It may not be displayed if the AI does not generate any.
- **AI Q&A pairs** are questions generated by an AI system about the original English content, along with AI-generated answers based on two sources: the original English text and the Spanish translation, which has been translated back into English for display.

Task Instructions:

- Using the provided information, you will assess whether the Spanish translation is of sufficient quality to safely share with your Spanish neighbor.
- You can select one of the following two options:
  - **Safe to share as-is ✓ :** The Spanish translation has no or minimal risk.
  - **Needs bilingual review before sharing ✗ :** The Spanish translation have some potential issues.
- You will gain points for accurately judging the translation quality (Safe to share as-is, Needs bilingual review before sharing).
- Please do not reload the website during annotation.
- If possible, please complete the survey in a single sitting.

Compensation: You will receive a compensation of $5 per session, which is estimated to take at most 20 minutes. If your accuracy exceeds 70%, you will receive a performance-based bonus of $2.

Start Tutorial

(b) Task Instructions

Tutorial

English source	In the worst affected countries including Italy, Spain and France, lockdowns like in China in January, have been implemented.
	s más afectados, incluidos Italia, España y Francia, se han implementado cierres como los de enero en China.

English Sentence ×

Original English sentence.

Back Next

is the Spanish translation good enough to safely share with your neighbor?

sharing ✗

assessment?

○ 2 (Unconfident)
○ 3 (Neutral)
○ 4 (Confident)
○ 5 (Very confident)

Start Survey

(c) Tutorial

Task Instructions (click to collapse) ▼

- Using the provided information, you will assess whether the Spanish translation is of sufficient quality to safely share with your Spanish neighbor. Answer to the best of your knowledge, even if you cannot understand Spanish.
- You may change your initial decision after reviewing the AI feedback.
- You can select one of the following two options:
  - **Safe to share as-is ✓ :** No or minimal risk.
  - **Needs bilingual review before sharing ✗ :** Potential issues.
- You will gain points for accurately judging the translation quality.

Example 1 / 20

English source	Infections spread easily on board, and medical care is limited.
Spanish translation	El contagio se propaga fácilmente a bordo de una embarcación, y la atención médica es limitada.

(1) To the best of your knowledge, is the Spanish translation good enough to safely share with your neighbor?

○ Safe to share as-is ✓
○ Needs bilingual review before sharing ✗

(2) How confident are you in your assessment?

○ 1 (Very unconfident)
○ 2 (Unconfident)
○ 3 (Neutral)
○ 4 (Confident)
○ 5 (Very confident)

Next

(d) Independent Decision-Making

Example 1 / 20

English source	If you are elderly or otherwise in a high-risk group and can be vaccinated for pneumonia, get that vaccination.
Spanish translation	Si es un adulto mayor o se encuentra en un grupo de alto riesgo y puede vacunarce contra la neumonía, hágalo.

AI Error Detection: Words that the AI system identifies as translation errors will be highlighted in the Spanish translation. If nothing is displayed, it means the AI did not detect any errors.
(Minor Major Critical)

Spanish Translation: Si es un adulto mayor o se encuentra en un grupo de alto riesgo y puede vacunarce contra la neumonía, hágalo.

(1) To the best of your knowledge, is the Spanish translation good enough to safely share with your neighbor?

○ Safe to share as-is ✓
○ Needs bilingual review before sharing ✗

(2) How confident are you in your assessment?

○ 1 (Very unconfident)
○ 2 (Unconfident)
○ 3 (Neutral)
○ 4 (Confident)
○ 5 (Very confident)

Next

(e) AI-Assisted: Error Highlights

Example 1 / 20

English source	Long-term effects for people who have recovered remain unclear, but there is evidence of reduced lung capacity in some recovered patients.
Spanish translation	Todavía no quedan claros los efectos a largo plazo para las personas que se han recuperado, pero hay evidencia de reducción de la capacidad pulmonar en algunos pacientes recuperados y problemas cardíacos.

AI Explanation: You will see an explanation generated by an AI system about the Spanish translation.

AI Explanation: The Spanish translation is mostly accurate, but it introduces an additional element ("problemas cardíacos") not present in the original English sentence, which alters its meaning. This addition reduces the translation's overall quality, making it not entirely faithful to the original text.

(1) To the best of your knowledge, is the Spanish translation good enough to safely share with your neighbor?

○ Safe to share as-is ✓
○ Needs bilingual review before sharing ✗

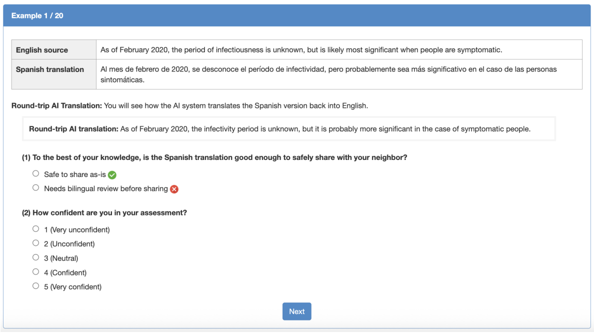
(2) How confident are you in your assessment?

○ 1 (Very unconfident)
○ 2 (Unconfident)
○ 3 (Neutral)
○ 4 (Confident)
○ 5 (Very confident)

Next

(f) AI-Assisted: LLM Explanation

Figure 7: Screenshots of our annotation interface, organized according to the task flow. Each example has a brief summary of the task instructions at the top, which participants can click to expand or collapse.



Example 1 / 20

English source	As of February 2020, the period of infectiousness is unknown, but is likely most significant when people are symptomatic.
Spanish translation	Al mes de febrero de 2020, se desconoce el período de infectividad, pero probablemente sea más significativo en el caso de las personas sintomáticas.

Round-trip AI Translation: You will see how the AI system translates the Spanish version back into English.

Round-trip AI translation: As of February 2020, the infectivity period is unknown, but it is probably more significant in the case of symptomatic people.

(1) To the best of your knowledge, is the Spanish translation good enough to safely share with your neighbor?

○ Safe to share as-is ✓

○ Needs bilingual review before sharing ✗

(2) How confident are you in your assessment?

○ 1 (Very unconfident)

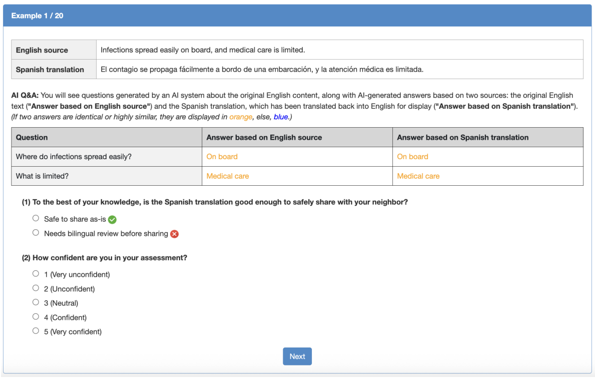
○ 2 (Unconfident)

○ 3 (Neutral)

○ 4 (Confident)

○ 5 (Very confident)

(h) AI-Assisted: Backtranslation



Example 1 / 20

English source	Infections spread easily on board, and medical care is limited.
Spanish translation	El contagio se propaga fácilmente a bordo de una embarcación, y la atención médica es limitada.

AI Q&A: You will see questions generated by an AI system about the original English content, along with AI-generated answers based on two sources: the original English text (**"Answer based on English source"**) and the Spanish translation, which has been translated back into English for display (**"Answer based on Spanish translation"**). *(If two answers are identical or highly similar, they are displayed in orange, else, blue.)*

Question	Answer based on English source	Answer based on Spanish translation
Where do infections spread easily?	On board	On board
What is limited?	Medical care	Medical care

(1) To the best of your knowledge, is the Spanish translation good enough to safely share with your neighbor?

○ Safe to share as-is ✓

○ Needs bilingual review before sharing ✗

(2) How confident are you in your assessment?

○ 1 (Very unconfident)

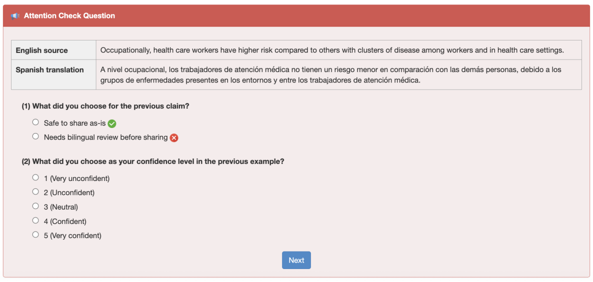
○ 2 (Unconfident)

○ 3 (Neutral)

○ 4 (Confident)

○ 5 (Very confident)

(i) AI-Assisted: QA Table



Attention Check Question

English source	Occupationally, health care workers have higher risk compared to others with clusters of disease among workers and in health care settings.
Spanish translation	A nivel ocupacional, los trabajadores de atención médica no tienen un riesgo menor en comparación con las demás personas, debido a los grupos de enfermedades presentes en los entornos y entre los trabajadores de atención médica.

(1) What did you choose for the previous claim?

○ Safe to share as-is ✓

○ Needs bilingual review before sharing ✗

(2) What did you choose as your confidence level in the previous example?

○ 1 (Very unconfident)

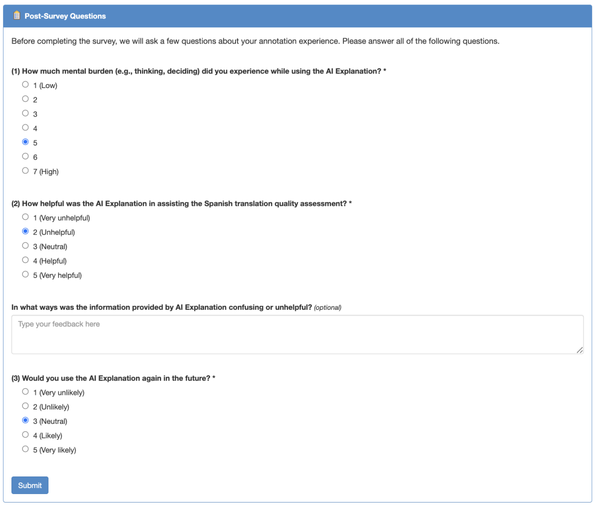
○ 2 (Unconfident)

○ 3 (Neutral)

○ 4 (Confident)

○ 5 (Very confident)

(j) Attention Check Question



Post-Survey Questions

Before completing the survey, we will ask a few questions about your annotation experience. Please answer all of the following questions.

(1) How much mental burden (e.g., thinking, deciding) did you experience while using the AI Explanation? *

○ 1 (Low)

○ 2

○ 3

○ 4

● 5

○ 6

○ 7 (High)

(2) How helpful was the AI Explanation in assisting the Spanish translation quality assessment? *

○ 1 (Very unhelpful)

● 2 (Unhelpful)

○ 3 (Neutral)

○ 4 (Helpful)

○ 5 (Very helpful)

In what ways was the information provided by AI Explanation confusing or unhelpful? *(optional)*

Type your feedback here

(3) Would you use the AI Explanation again in the future? *

○ 1 (Very unlikely)

○ 2 (Unlikely)

● 3 (Neutral)

○ 4 (Likely)

○ 5 (Very likely)

(k) Post-task Survey