HyoJung Han, Nishant Balepur, **Jordan Boyd-Graber**, and Marine Carpuat. **Measuring User's Mental Models of Speech Translation in Human-MT Collaboration**. *Association for Computational Linguistics*, 2026.


```
@inproceedings{Han:Balepur:Boyd-Graber:Carpuat-2026,
Title = {Measuring User's Mental Models of Speech Translation in Human-MT Collaboration},
Author = {HyoJung Han and Nishant Balepur and Jordan Lee Boyd-Graber and Marine Carpuat},
Booktitle = {Association for Computational Linguistics},
Year = {2026},
Url = {http://cs.umd.edu/~jbg//docs/2026_acl_mental_models_speech_translation.pdf},
}
```


Accessible Abstract: When people see a translation from an AI, what causes them to trust (or reject) it? We ask people to make translated question "good enough" that a QA system can answer it correctly. In other words, they point out where they think the AI made mistakes that are important enough to cause confusion. Humans think errors happen in translation when they see contradictions, can use their own language knowledge (e.g., cognates), and when there is a lack of fluency.

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Measuring Users' Mental Models of Speech Translation in Human-AI Collaboration

HyoJung Han **Nishant Balepur** **Jordan Boyd-Graber** **Marine Carpuat**
University of Maryland, College Park, USA
{hjhan, nbalepur, jbg, marine}@cs.umd.edu

Abstract

Millions of people use machine translation (MT) tools daily, yet little is known about their perception of what systems can and cannot do. This paper studies users' mental models of speech translation systems through a new framework based on cross-lingual question answering, where users either accept MT output or request professional re-translation to answer questions based on the information presented in a foreign language. By analyzing user behavior and accuracy trends across varying translation qualities, we examine to what extent they can predict where the system is likely to be wrong, and how this mental model evolves. Users develop stronger mental models with practice, especially when they have some knowledge of the source language, primarily by relying on surface-level error cues. Moreover, providing speech transcriptions can help users develop better mental models. Our results show the promise of cross-lingual question answering as a downstream task for studying MT mental models, and advancing our understanding of human–AI collaboration.

1 Introduction

Millions of people use machine translation (MT) tools daily, including in both casual (Calefato et al., 2016; Xiao et al., 2025a) and high-stakes contexts where errors can have serious consequences (Liebling et al., 2020; Nunes Vieira, 2024). To use MT effectively, users must understand inputs and scenarios where systems work reliably and where they fail, so they can make informed choices about e.g., when to trust outputs and when to seek human translation. This is particularly needed in *speech* translation, where audio inputs further amplify variability in output quality (Spechbach et al., 2019; Han et al., 2024b) and reflect more realistic user scenarios. This understanding of a MT system's strengths and limitations is a much needed form of MT literacy (Bowker and Ciro, 2019; O'Brien and Ehrensberger-Dow, 2020), and it forms the basis of users' **mental model**s (MMs)—their internal understanding of how an AI system works, when it succeeds or fails, and how to act on its outputs (Norman, 1983; Hoffman et al., 2023). Such models are critical for integrating AI recommendations into decision-making (De-Arteaga et al., 2020; Sieker et al., 2024).

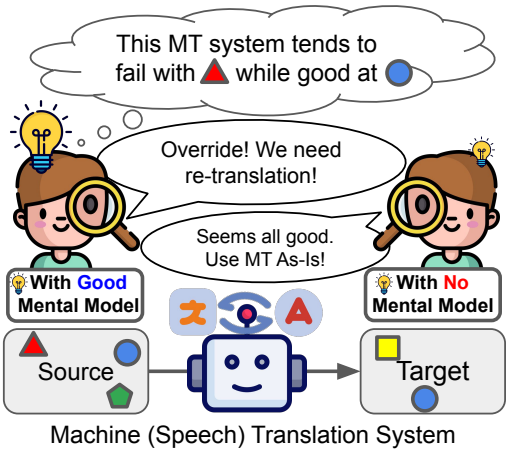


Figure 1: Users' mental model, or their understanding of an MT system's strengths and weaknesses, is essential for effectively integrating MT system's output or appropriately intervening on it. Geometric shapes in the figure represent the features of input/output that the users can use in updating their mental model of MT.

Despite its importance, the study of users' mental models in MT remains limited. The HCI literature provides a starting point for studying the role of MM in human-AI collaboration, but primarily focuses on classification and regression tasks, where models have simpler error boundaries (Bansal et al., 2019b; Kelly et al., 2023), and AI predictions directly align with the users' decision needs (Liu et al., 2024b; Vats et al., 2025). As a result, existing methods do not easily port to MT, where natural language outputs can be imperfect in many ways, and users must determine how to use them to make informed downstream decisions.

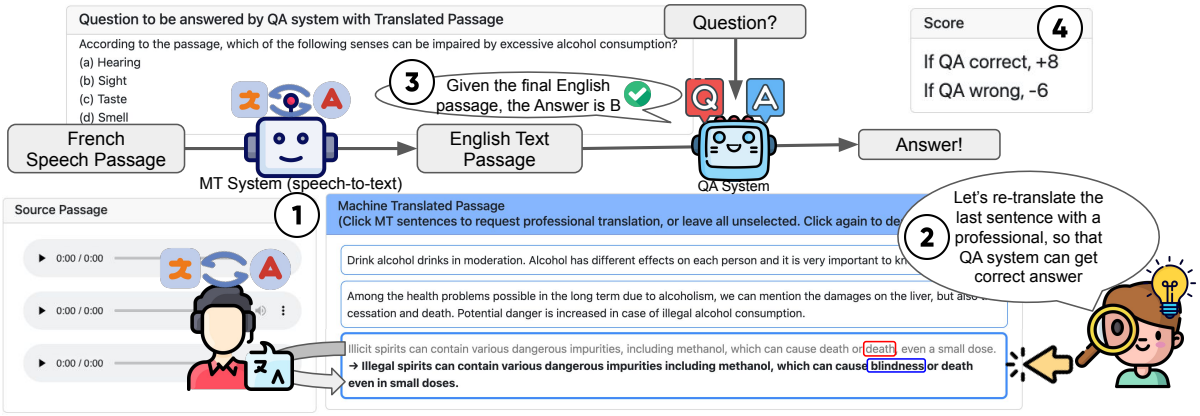


Figure 2: Our proposed framework for measuring and updating users' mental models of MT system. ①: MT system translates the source passage input. ②: users decide which one among each segment to re-translate by a professional translator, and which one to use as-is, forming the final passage. ③: the QA system does the QA consuming the passage. ④: the user's reward is by the QA correctness minus how much they re-translated.

This paper seeks to **understand and improve users' mental models of MT** (Figure 1) by introducing a framework based on cross-lingual question answering (QA), where users aim to maximize task performance by either accepting MT output as-is or intervening with professional re-translation. This process naturally teaches them to recognize MT's error patterns and refine their mental models (Figure 2). This downstream framing emphasizes the "fitness for purpose" of translations, and encourages users to do a cost-benefit analysis of the impact of potential errors on answers, rather than assessing MT quality in isolation (Mehandru et al., 2023; Xiao et al., 2025a). In this setup, users' efforts to maximize task performance naturally reveal how they form and refine their mental models over the course of the task.

This framework lets us study human MMs for speech translation, and investigate how different factors impact their formation: users' language proficiency, input/output features they rely on, and types of interventions most effective for updating MMs. As we will see, users refine their MMs over time (§5.1): fluent and intermediate users had consistent improvements, but basic users struggled to update their models (§5.2). When examining error features, users were most responsive to output surface-level cues such as incompleteness, while topic-related errors remained the hardest to detect (§5.3). Finally, transcription explanation proved most effective by offering additional clues about error sources, while error span explanation boosted accuracy but encouraged over-reliance, limiting MM development (§5.4).

Overall, this work introduces a concrete framework to measure users' mental models of MT systems, and provides a foundation for exploring how to promote MT literacy and appropriate MM development in the wild, and for studying MM for a broader range of generation tasks such as summarization, research, and dialogue.

2 Mental Model for Machine Translation

We first review how mental models have been defined and studied in prior work (§2.1), and then describe how we design our study to investigate mental models in the context of MT (§2.2).

2.1 Background and Prior Work

Conceptualizations of mental models have been introduced in contexts ranging from cognitive psychology (Norman, 1983; Klein and Hoffman, 2008; Iida et al., 2024) to explainable AI (Miller, 2018; Mueller et al., 2019; Brachman et al., 2025). We adopt the definition of users' mental models of an AI system as their understanding of its capabilities and limitations: its strengths, weaknesses, and the boundaries of its errors (Bansal et al., 2019a; Kelly et al., 2023; Vats et al., 2025).

Previous studies place users in human–AI collaborative tasks, where participants are rewarded for making decisions that reflect appropriate expectations of the system's successes and failures. Kelly et al. (2023) asked participants to predict an AI agent's accuracy across rounds of 12 trivia questions, with rewards tied to how closely their estimates matched actual performance—thus capturing their mental model of the system. In Bansal

et al. (2019a,b), participants were tasked with a binary decision of determining whether an object was defective, guided by recommendations from an AI system. They could accept or override the recommendation, with monetary rewards reflecting correct choices and steep penalties for errors, simulating high-stakes decision-making.

However, they focus on a classification task in which the system directly provides a decision to the user, a setup that does not align with MT. We address this gap by proposing a new approach to measuring and developing users' mental models in MT (Section 2.2).

2.2 Measuring Mental Models for MT Systems in Human+MT Context

A key question in our study is how to meaningfully assess users' mental models of MT. There are several ways to do that. One straightforward suggestion is to present both the source and its translated version, then ask the user to evaluate how well the source has been rendered, similar to a quality estimation (QE) task but with a human (Specia et al., 2010; Callison-Burch et al., 2012). The user's assessment can be compared to actual metric scores and rewarded if their predictions are close, as in Kelly et al. (2023).

However, this setup may not fully align with real-world usage, as what makes a translation "good" by standard metrics does not necessarily reflect what users find "useful". In other words, standard metrics can overlook whether the MT quality is "good enough". For instance, generic measures may flag case disagreement or other minor stylistic issues but overlook critical errors from a user's perspective (Tomita et al., 1993; Krubiński et al., 2021; Han et al., 2022). If we want to measure usefulness, we need to assess "fitness for purpose"—how well a translation supports a downstream task the user cares about—rather than simply how accurately it predicts a quality score (Hovy et al., 2002; Bowker, 2010; Liu et al., 2024a). Thus, focusing on utility—while also prompting users to weigh the potential impact of translation errors on their decisions—may be more valuable than intrinsic definitions of translation quality, especially in MT as an intermediate tool scenario (Mehandru et al., 2023; Xiao et al., 2025a).

As an alternative to a QE-based setup, **we propose to use a cross-lingual downstream task to measure the mental model of machine translation**, where users can naturally learn to recognize MT's error patterns and refine their mental models. We choose reading comprehension question answering as a downstream task because it can measure translation quality by testing whether key information is preserved well enough to answer the question, thereby emphasizing quality as fitness for purpose (Jones et al., 2005; Moghe et al., 2023; Agrawal et al., 2024; Ki et al., 2025). The user sees the non-English source information and its translated English version. The goal is to complete a QA task using the translated English passage as effectively as possible. But the translation is not static: for each translation segment (here, individual sentences), users can either accept the MT output as is or request a re-translation by a "professional" (i.e., get a gold reference translation).

To control for the possibility that users might answer questions using world knowledge rather than the translation, we rely on a fixed QA system, which isolates the influence of their MMs of MT. The reward is determined by whether the QA *system* answers correctly given the final translated input. To prevent users from getting professional translations for everything, the reward decreases every time they request a re-translation from a base reward.

3 Key Research Questions

We explore three primary research questions: how language proficiency affects the evaluation and development of mental models (§3.1), which input/output features users rely on to develop and update their mental models (§3.2), and what types of explanation are most effective for updating a mental model (§3.3).

3.1 RQ1: How Does Language Proficiency Affect Development of Mental Models?

Prior work shows that users who are not bilingual struggle to assess MT outputs (Xiao et al., 2025b) and that interventions based on quality estimation and backtranslation affect reliance inconsistently, depending on fluency (Mehandru et al., 2023; Zouhar et al., 2021). In this work, we study how mental models develop as a function of source-language proficiency. This can help us understand user decisions at a finer-grained level and inform training strategies to promote more appropriate reliance in the future.

3.2 RQ2: Which Input/Output Features Do Users Rely On to Update Their MM?

To understand which characteristics they usually rely on to detect errors and update their mental model, we categorize input and output features of naturally occurring speech translation errors. We focus on four main categories: translation errors containing *rare words or named entities*, inputs with *phonetic ambiguities or noise*, *domain-specific errors*, and outputs that exhibit *incompleteness or unnaturalness*.[1]

Source speech containing ***rare words or named entities*** often leads to errors in translation systems, particularly when contextual cues are limited (Gaido et al., 2021). ***Phonetic ambiguities or noisy inputs*** are also a key weakness of speech translation systems (Anwar et al., 2023; Han et al., 2024a; Mirzayev, 2025). As noted by Irvine et al. (2013), ***domain-specific errors*** occur when the system struggles more in certain topics (e.g., sports) while translating more reliably in others (e.g., science). Finally, ***incompleteness or unnaturalness*** in translation is often the easiest indicator of error for users to detect (Martindale and Carpuat, 2018).

While we do not explicitly prime participants with these features during the task, we also gather their reflections in a post-survey on "what characteristics they have learned to look for when identifying incorrect translations" to cross-check whether the categories we observe in error analysis are also salient to participants in forming or updating their mental models of the system.

3.3 RQ3: What Type of Explanation Is Most Effective for Updating Mental Models?

Providing appropriate explanations can improve users' ability to understand and predict system behavior, as clear feedback can help them make sense of complex models (Rutjes et al., 2019; Xie et al., 2022; Boyd-Graber et al., 2022). Building on this, we investigate which types of translation explanations most effectively help users build and refine their MMs of an MT system.

Among several types of explanations that a machine translation model can generate by itself, we choose **transcriptions** as internal explanations that can be directly produced by the speech translation model. The transcription provides insights into how the model processes and represents the source inputs. Even when the transcription is not explicitly used to generate translation, it offers clues for errors stemming from wrong audio representation and its possible improvements in translation (Dou et al., 2025).

Beyond explanations generated by the speech translation model itself, external models such as quality estimation (QE) can also provide insights. These models output segment-level scores and highlight error spans, helping users spot unreliable translations. In our setup, we use **error spans**—highlighted segments of the translation that contain errors—as external explanations (Zouhar et al., 2021; Briakou et al., 2023), since an overall score alone is often not reliable or useful to support users' understanding and MM development (Mehandru et al., 2023; Ki et al., 2025). We compare a baseline with no explanations against conditions that provide transcriptions and error spans as local explanations.

4 User Study Design

This section outlines the user study setup to address our research questions. We describe how we curated the question set to present to participants for effectively evaluating and developing their mental models, considering translation error features (§4.1). We then detail the interface in our experiments (§4.2). Finally, we describe the participants and their proficiency groups (§4.3).

4.1 Curation of QA Set

We curated 16 reading comprehension sets from 2M-BELEBELE (Costa-jussà et al., 2025), where each passage consists of French audio sentences and is paired with an English multiple-choice question. We begin with question sets that Mistral-7B[2](Jiang et al., 2023, as QA model) answers correctly when provided with a gold English passage. Each French speech sentence is translated into English text using Whisper (Radford et al., 2022), a widely adopted speech translation system. We then provide the translated passage along with the corresponding multiple-choice question to the QA model to do the reading comprehension task and record the system's multiple-choice response. We further apply category classification[3] (Wettig et al., 2025) to each passage and collect XCOMET scores

[1] We identified these four major categories based on our preliminary error analysis and prior literature about MT errors. More details in Section 4.1

[2] mistralai/Mistral-7B-Instruct-v0.3

[3] WebOrganizer/TopicClassifier-NoURL

(Guerreiro et al., 2024) based on gold English transcription and translation output.

Based on the set where the QA model is correct with the gold English passage but fails with the translated English passage, we run a preliminary analysis of speech translation errors that cause the QA system to fail and manually categorize them into groups. From the topic-wise analysis, translations in the Sports domain tend to score lower, while Science translations are higher. We thus narrow the scope to these two domains, treating Sports as more translation-challenging and Science as comparatively easier. Finally, we identify four major categories, as described in Section 3.2. Among the questions where translation errors lead to incorrect QA answers, we select 11 examples in which the feature categories are evenly represented, except for the topic, which is dominated by Sports. In contrast, we select five correctly translated questions, which are largely drawn from the Science domain.

4.2 Experiment Design

Interface. We implement an interface to train and test users' mental models of MT. The interface presents segment-level pairs of French audio and machine-translated English text along with reading comprehension multiple-choice questions (Figure 2). Users are instructed to select only translations that they believe require professional re-translation to ensure the QA system will produce correct answers based on the translated input.

Rewards. If the QA system answers correctly with the final passage, the user's score increases by the calculated reward. However, if the user fails to identify the problematic sentence containing a critical error causing a QA error (recall that the QA system by construction is correct on gold question translations), they *lose* six points. To discourage excessive re-translation requests, each request deducts points proportionally from the base reward of 12, depending on the number of sentences selected (e.g., selecting one of four sentences lowers the reward to nine, while selecting two lowers it to six); if all sentences are selected, the reward collapses to zero. The *maximum possible score* for each question is therefore achieved by re-translating only the minimal necessary segment (usually one).

Explanation. We also experiment with additional forms of explanation—transcriptions and error spans (Section 3.3)—to examine what types of explanation are most effective for updating a mental model. XCOMET extracts the error spans based on the gold English transcription and the speech-translated text, and transcriptions are generated by the same speech translation model. Users are automatically assigned to three different experimental conditions: Default (no explanation), Transcription, and Error span. In the transcription condition, the text is displayed beneath the audio, while in the error-span condition, users see highlighted spans indicating the detected errors in the translated text.

Assist (↓)	Basic	Intermediate	Fluent	All
Default	5	5	5	15
Transcription	4	6	5	15
Error Span	4	7	4	15
All	13	18	14	45

Table 1: Distribution of participants by French (source language) proficiency and the assistance experimental group with kinds of explanation. Most participants are proficient English (target language) speakers.

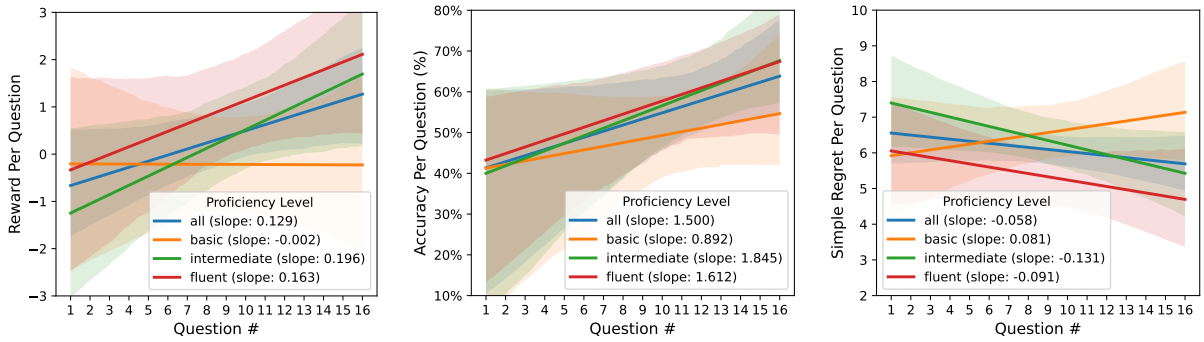
4.3 Defining Source Language Proficiency

We ran user studies with crowd workers on Prolific,[4] representing a diverse range of source-language fluency. The post-survey asks participants about their French proficiency. We also include four quality estimation (QE) tasks, presented before every four QA tasks, where participants rate the adequacy of the translation on a 1–6 scale (Figure 6). We compare their QE ratings with the converted XCOMET scores to assess user performance. Based on their self-reported proficiency and QE performance, we divide participants into three groups: basic, intermediate, and fluent (Table 1; QE setup and proficiency grouping detailed in Appendix A).

5 Results: Participants Build MMs of MT

This section presents our main findings on how users form and refine their mental models of the MT system. Users build better mental models as they gain experience with the system and the task (§5.1). Next, higher language proficiency leads to better mental models (§5.2). We then investigate which input/output features users rely on when developing and updating their mental models (§5.3). Finally, we assess that transcription explanation is more effective in supporting mental model updates

[4] https://www.prolific.com/



(a) Reward Trend (↑) (b) QA Accuracy Trend (↑) (c) Simple Regret Trend (↓)

Figure 3: As participants progress, reward and accuracy increase (left, middle). Simple regret (maximum possible reward – actual reward) decreases as users become more precise in selecting only the problematic sentences (right). The slope values reflect the average change with each additional question. Participants progressively learn about the MT system over the course of the task, developing better mental models.

for speech translation than error span explanation (§5.4). Additionally, we present key phrases extracted from users' reflection notes (Appendix 6).

5.1 Participants Gradually Update their MMs

Figure 3 shows overall trends in three key metrics across all participants: reward, question accuracy, and simple regret (defined as the difference between the maximum possible reward and the actual reward; lower is better). Figure 3a and 3b present regression plots of reward and accuracy per question (higher is better), while Figure 3c shows a regression plot of **simple regret**. The slope value in the legend of each plot reflects the average change with each additional question. Especially, a negative slope in simple regret indicates that the user is approaching the optimal score by requesting retranslations only when necessary and avoiding redundant or incorrect selections, signaling that they are refining their mental models. In contrast, a positive slope suggests limited or no learning.

Overall positive slopes in reward and accuracy and negative slopes in simple regret suggest that ***users gradually refine their mental models***, enabling them to better identify MT errors that cause the QA system to fail.

5.2 A Little Bilingualism Goes a Long Way

While the overall trends suggest that users refine their mental models over time, Figure 3 also highlights differences across proficiency groups.

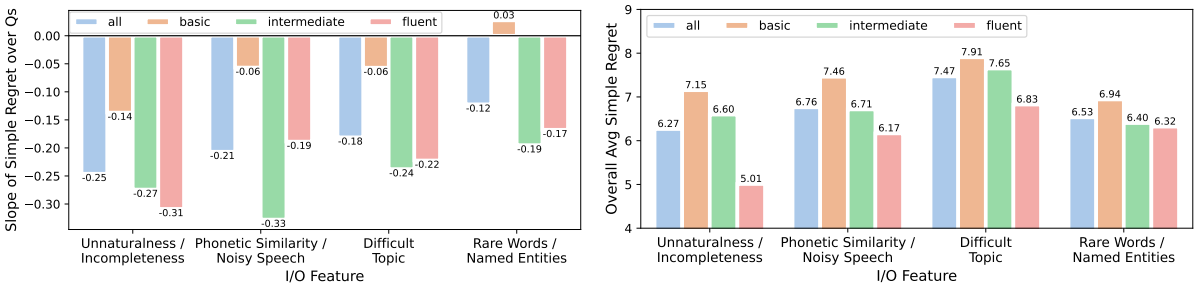
For reward and QA accuracy, fluent and intermediate users both show stronger positive slopes, while the basic group shows a flatter (Figure 3b) or even negative (Figure 3a) slope. In simple regret (Figure 3c), the basic group shows an upward trend, which reflects that they struggled to properly develop a mental model of the MT system; instead, they tend to select more sentences to avoid penalties from the incorrect QA outcome. Overall, the intermediate fluency group shows even steeper slopes compared to fluent users. One possible explanation is that fluent participants do well from the outset, even without fully developing their mental models, whereas intermediate users initially struggle but quickly learn to recognize the traits where the MT system fails and improve more rapidly. In summary, ***higher source-language proficiency leads to better mental models*** of the translation system.

5.3 Feature-Specific Trends in MM Updates

We group the question set according to the four input/output (I/O) features of MT errors categorized in Section 3.2, and analyze the trends of simple regret within those groups to identify which features are most salient for users when updating their mental models (Figure 4).

We analyze both the slopes of simple regret across corresponding questions' order (Figure 4a) and the overall average simple regret (Figure 4b).[5] Across all users, *incompleteness/unnaturalness* and *phonetic similarity or noisy speech* show the steepest negative slopes, meaning that these features are easy for users to pick up and use when updating their mental models of MT, allowing them to respond more effectively when encountering similar cases in the future. In particular, the salience of

[5]The slope, as described in Section 5.1, reflects the average change in performance per question, with negative values indicating a desirable trend toward optimal rewards.



(a) Slope of simple regret across question order.

(b) Overall average simple regret for each feature type.

Figure 4: Simple regret (maximum possible reward – actual reward) trends within the questions associated with each input/output feature of MT errors. The slope value (left) reflects the average change with each additional question. The *incompleteness/unnaturalness* and *phonetic similarity or noisy speech* with the steeper slope values are salient features of MT errors that users readily pick up on to update their mental models, whereas *topic-related errors* are the hardest for users to detect and leverage in this process.

incompleteness/unnaturalness aligns with Martindale and Carpuat (2018), which shows users are particularly reactive to disfluent translations.

Rare words or named entities show the flattest slopes; however, their mean value (right-most bars in Figure 4b) is already the lowest for basic and intermediate users, suggesting that users relied on this feature from the beginning.

The topic-related feature shows the second-flattest slopes and the highest overall average regret. We conjecture that this is because, even though participants are primed that the topic will be either sports or science, it is difficult for them to recognize the specific topic of a passage during the experiment unless it is explicitly indicated. As a result, topic-related errors are the hardest feature for users to detect and leverage when updating their mental models, while ***incompleteness/unnaturalness* and *phonetic noise* are the most salient to users**.

	Final Score			
Condition (↓)	Basic	Intermediate	Fluent	All
Default	39.60	30.00	71.00	46.87
Transcription	34.00	56.67	54.20	49.80
Error Span	41.75	34.14	43.00	38.53
All	38.54	40.50	57.00	45.07
	Overall Accuracy (%)			
Condition (↓)	Basic	Intermediate	Fluent	All
Default	63.75	60.00	70.00	64.58
Transcription	60.94	71.02	70.00	67.99
Error Span	68.75	71.43	68.75	70.00
All	64.42	68.12	69.64	67.52

Table 2: Comparison of final scores and overall accuracy across three explanation conditions: default, transcription, and error span. Transcription provides the greatest benefit for intermediate users, while error spans increase accuracy but reduce rewards due to over-selection, limiting effective mental model development.

5.4 Useful Explanation for Better MM

Table 2 presents the results of different types of explanation—default (no additional support), transcription, and error span—on participants' final accumulated scores and overall QA accuracy.[6]

For final scores (upper), ASR transcription yielded the highest average across all users (49.80), while also improving overall accuracy (from 64.58% to 67.99%). This suggests that presenting the transcription alongside the translation provides users with additional clues about how translation errors may stem from incorrect audio representations, which in turn helps them better understand the speech translation system's behavior and develop more refined mental models. The improvements are most prominent for intermediate French proficiency users, while performance for fluent users even decreases. One possible explanation is that achieving the highest level of performance requires careful listening to the audio, but the presence of ASR transcriptions may reduce the tendency to play and attend to the audio. As a result, fluent users might rely too heavily on the transcription, which limits their performance gains.

Error span explanation, on the other hand, yields the highest accuracy (70%) but the lowest final score (38.53). This pattern arises because users

[6]The plots with statistical significance analysis are in Appendix B, supporting meaningful difference between conditions and proficiency.

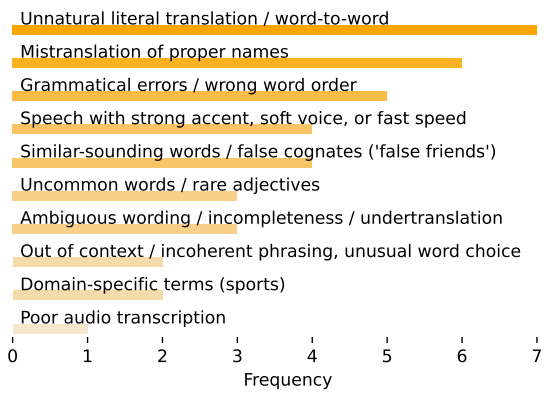


Figure 5: Key phrases from participants' reflections on the speech translation system's error patterns and their identifying cues for updating their MMs. These align with our predefined input/output feature categories.

often select any highlighted span, which increases correctness but also leads to frequent re-translation requests and consequently lower rewards, as similarly noted by Sarti et al. (2025). While error spans help users achieve higher accuracy in this setup, they do not necessarily support the development of a robust mental model of the MT system.

Overall, ***transcription explanation is more effective, especially for intermediate proficiency users***, while error span is less suitable for supporting the development of well-formed MMs.

6 Analysis on User's Reflection

At the end of the survey, participants were asked to reflect on the types of errors the MT system is likely to make and the characteristics they learned to look for when identifying incorrect translations. We manually analyzed the key phrases in these reflection notes and grouped them into categories, shown in Figure 5. The most frequently mentioned issue was unnatural literal translation, aligning with our earlier analysis of the most salient features (§5.3). Overall, we find that the traits users identify when updating their mental models align closely with our predefined categories of I/O features.

7 Related Work

Mental Model. A mental model can be understood as a general cognitive construct that people form through their interactions with the world, including others, themselves, the tasks they do, and the topics they learn. These models provide both predictive and explanatory power for making sense of such interactions (Norman, 1983; Kumar et al., 2023). In explainable AI, a mental model refers to how users understand both the functioning of the AI system and the conditions under which it is applied (Miller, 2018; Mueller et al., 2019).

Mental models are closely linked to trust and transparency in AI systems, highlighting the importance of users' understanding in anticipating system failures (Vats et al., 2025). Strengthening this mutual understanding between humans and AI has been shown to improve collaborative performance (Tabrez et al., 2020; Hoffman et al., 2023). Especially in recent days, as model complexity increases, the explainability of predictions—through transparency, interpretability, and accessibility—has become increasingly important for enabling users to build accurate mental models that closely reflect the systems' actual capabilities (Anderson et al., 2020; Feng and Boyd-Graber, 2022; Sieker et al., 2024).

Human-MT Collaboration. As a widely used AI technology, machine translation (MT) is a common case of human-AI collaboration in the real world, facilitating communication in multilingual settings (Calefato et al., 2016), and even in high-stakes domains like healthcare, law, and immigration (Liebling et al., 2020; Nunes Vieira, 2024; Vieira et al., 2021). However, studies of human-MT collaboration have historically focused on professional translators (Hutchins, 2001; Cadwell et al., 2016; O'Brien, 2024; Scansani et al., 2019; Sarti et al., 2025), and the growing population of lay users remains understudied (Savoldi et al., 2025; Kenny et al., 2022). Recent work in this space has focused on exploring the specificities of post-editing by non-experts (Koehn, 2010; Bawden et al., 2024; O'Brien et al., 2018), and on developing quality feedback mechanisms to help users decide whether a given output is safe to use or not (Zouhar et al., 2021; Mehandru et al., 2023; Ki et al., 2025). However, these feedback mechanisms remain imperfect, are not available in generic translation apps, and are almost exclusively focused on text translation. This paper thus takes a complementary approach and seeks to get to the root of the problem by first measuring how people understand MT behavior, focusing on the challenging case of speech translation.

Furthermore, the cross-lingual QA framework introduced here could motivate future interventions to promote MT literacy. While existing MT literacy education focuses on telling users where MT can

go wrong (Bowker and Ciro, 2019), our framework lets them experience errors and their consequences for themselves in controlled settings, and calibrate their MM through interactions.

8 Conclusion

To understand and improve users' mental models of machine translation systems, we propose a cross-lingual QA framework investigating how users perceive MT error tendencies and how their mental models evolve through interaction. Our findings show that users refine their mental models over time, with fluent and intermediate users demonstrating consistent improvements while basic users struggle to adapt. Error cues that are used to update their mental models, such as incompleteness and phonetic noise, were the most salient features, whereas topic-related errors remained the hardest to detect. Finally, transcription explanation is effective in supporting accurate mental model development, while error span highlighting encourages over-reliance.

Together, our study highlights promising directions for designing MT systems and interfaces that better support users in building accurate and robust mental models, opening new possibilities for extending mental model research beyond MT and toward deeper aspects of user understanding. In particular, our experiments show that the QA-tasked game framework is a promising interactive way to improve people's MT literacy, especially for intermediate and bilingual users, and that existing tools can provide useful explanations in the form of transcriptions. Yet, it opens up future challenges: how to support users with no source-language proficiency and how to encourage attention to input features that strongly influence AI outputs.

Limitations

We only experimented with French as the source language. While French offers a useful test case due to its availability in multilingual benchmarks, our findings may not fully generalize to other languages, especially those that differ more significantly from English in terms of phonetics, morphology, or domain coverage. We consider it promising that our framework shows users' ability to build mental models in French, so future work should extend this framework to a broader set of languages to study the robustness of our conclusions.

Both the number of questions and the number of participants in our experiment were limited due to resource constraints. While the small sample size can limit statistical power and the diversity of strategies observed across users, our analysis in Appendix B shows that the key results are statistically significant and thus meaningful. Still, scaling up the dataset and recruiting a larger, more varied participant pool would enable more reliable estimates and a richer understanding of how different user groups develop mental models.

Ethics Statement

Participants were compensated $10 per set that takes up to 40 minutes to complete, and this rate is above our region's hourly minimum wage. The institutional review board (IRB) at our institution approved our study, confirming that potential risks were appropriately managed. Participants were explicitly informed that they would be part of a research study, and the study proceeded only after giving their consent. Recruitment was determined by participants' proficiency in English and French on the crowdsourcing platform, regardless of demographic or geographic characteristics. We will release no PII from participants. Appendix C has all instructions shown to participants. We use AI assistants to partially refine or polish writing at the sentence level (e.g., fixing grammar, re-wording sentences).

Acknowledgments

This material is based on work supported in part by the Institute for Trustworthy AI in Law and Society (TRAILS), which is supported by the National Science Foundation under Award No. 2229885. This material is also based upon work supported by the National Science Foundation under IIS-2403436 (Boyd-Graber) and DGE-2236417 (Balepur). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers. We also thank Yu Hou, Dayeon Ki, Fenfei Guo, Hiba El Oirghi, Julio Poveda, and the members of the CLIP Lab at the University of Maryland, College Park, for their insightful feedback.

References

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024. Assessing

the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267.

Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew L. Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Trans. Interact. Intell. Syst.*, 10(2):15:1–15:37.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation. In *Proc. INTERSPEECH 2023*, pages 4064–4068.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019a. Beyond accuracy: The role of mental models in human-ai team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):2–11.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019b. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2429–2437.

Rachel Bawden, Ziqian Peng, Maud Bénard, Éric Clergerie, Raphaël Esamotunu, Mathilde Huguin, Natalie Kübler, Alexandra Mestivier, Mona Michelot, Laurent Romary, Lichao Zhu, and François Yvon. 2024. Translate your own: a post-editing experiment in the NLP domain. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 431–443, Sheffield, UK. European Association for Machine Translation (EAMT).

Lynne Bowker. 2010. Can machine translation meet the needs of official language minority communities in canada? a recipient evaluation. *Evaluation of Translation Technology*, 8:123.

Lynne Bowker and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited.

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.

Michelle Brachman, Siya Kunde, Sarah Miller, Ana Fucs, Samantha Dempsey, Jamie Jabbour, and Werner Geyer. 2025. Building appropriate mental models: What users know and want to know about an agentic ai chatbot. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 247–264, New York, NY, USA. Association for Computing Machinery.

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.

Patrick Cadwell, Sheila Castilho, Sharon O'Brien, and Linda Mitchell. 2016. Human factors in machine translation and post-editing among institutional translators. *Translation Spaces*, 5(2):222–243.

Fabio Calefato, Filippo Lanubile, Tayana Conte, and Rafael Prikladnicki. 2016. Assessing the impact of real-time machine translation on multilingual meetings in global software projects. *Empirical Softw. Engg.*, 21(3):1002–1034.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Marta R. Costa-jussà, Bokai Yu, Pierre Andrews, Belen Alastruey, Necati Cihan Camgoz, Joe Chuang, Jean Maillard, Christophe Ropers, Arina Turkatenko, and Carleigh Wood. 2025. 2M-BELEBELE: Highly multilingual speech and American Sign Language comprehension dataset download PDF. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10893–10904, Vienna, Austria. Association for Computational Linguistics.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Huaixia Dou, Xinyu Tian, Xinglin Lyu, Jie Zhu, Junhui Li, and Lifan Guo. 2025. Speech translation refinement using large language models. *Preprint*, arXiv:2501.15090.

Aileen Edele, Julian Seuring, Cornelia Kristen, and Petra Stanat. 2015. Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research*, 52:99–123.

Shi Feng and Jordan Boyd-Graber. 2022. Learning to explain selectively: A case study on question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8372–8382, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Gaido, Susana Rodríguez, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2021. Is "moby dick" a whale or a bird? Named entities and terminology in speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1707–1716, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Mohamed Anwar, Juan Pino, Wei-Ning Hsu, Marine Carpuat, Bowen Shi, and Changhan Wang. 2024a. XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12896–12911, Bangkok, Thailand. Association for Computational Linguistics.

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 (EMNLP)*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024b. SpeechQE: Estimating the quality of direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21852–21867, Miami, Florida, USA. Association for Computational Linguistics.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5.

Eduard Hovy, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 17(1):43–75.

W John Hutchins. 2001. Machine translation over fifty years. *Histoire epistémologie langage*, 23(1):7–31.

Ayu Iida, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. 2024. Integrating large language model and mental model of others: Studies on dialogue communication based on implicature. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, HAI '24, page 260–269, New York, NY, USA. Association for Computing Machinery.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Douglas A. Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford J. Weinstein. 2005. Measuring translation quality by testing english speakers with a new defense language proficiency test for arabic.

Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing humans' mental models of ai: An item response theory approach. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1723–1734, New York, NY, USA. Association for Computing Machinery.

Dorothy Kenny, Olga Torres-Hostench, Caroline Rossi, Alice Carré, Pilar Sánchez-Gijón, Sharon O'Brien, Joss Moorkens, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Felipe Sánchez-Martínez, and Gema Ramírez-Sánchez. 2022. *Machine Translation for Everyone*. Language Science Press.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025. Should i share this translation? Evaluating quality feedback for user reliance on machine translation. *Preprint*, arXiv:2505.24683.

Gary Klein and Robert R Hoffman. 2008. Macrocognition, mental models, and cognitive task analysis methodology. *Naturalistic decision making and macrocognition*, pages 57–80.

Philipp Koehn. 2010. Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545, Los Angeles, California. Association for Computational Linguistics.

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.

Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Differentiating mental models of self and others: A hierarchical framework for knowledge assessment. *Psychological Review*, 130(6):1566–1591.

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.

Ting Liu, Chi-kiu Lo, Elizabeth Marshman, and Rebecca Knowles. 2024a. Evaluation briefs: Drawing on translation studies for human evaluation of MT. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 190–208, Chicago, USA. Association for Machine Translation in the Americas.

Zhongtao Liu, Parker Riley, Daniel Deutsch, Alison Lui, Mengmeng Niu, Apurva Shah, and Markus Freitag. 2024b. Beyond human-only: Evaluating human-machine collaboration for collecting high-quality translation data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1095–1106, Miami, Florida, USA. Association for Computational Linguistics.

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

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Preprint*, arXiv:1706.07269.

Elchin Mirzayev. 2025. Pronunciation issues in translation: challenges and implications. *Acta Globalis Humanitatis et Linguarum*, 2:41–50.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *Preprint*, arXiv:1902.01876.

Donald A Norman. 1983. Some observations on mental models. In *Mental models*, page 8. Psychology Press.

Lucas Nunes Vieira. 2024. *Uses of AI Translation in UK Public Service Contexts: A Preliminary Report*. Chartered Institute of Linguists.

Sharon O'Brien. 2024. Human-Centered augmented translation: Against antagonistic dualisms. *Perspectives*, 32(3):391–406.

Sharon O'Brien, Michel Simard, and Marie Josée Goulet. 2018. Machine Translation and Self-post-editing for Academic Writing Support: Quality Explorations. In *Translation Quality Assessment: From Principles to Practice, 2018, ISBN 978-3-030-08206-2, Págs. 237-262*, pages 237–262. Springer Suiza.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable ai and users' mental models. In *Where is the Human? Bridging the Gap Between AI and HCI*, United States. Association for Computing Machinery, Inc. CHI 2019 Workshop : Where is the Human? Bridging the Gap Between AI and HCI ; Conference date: 04-05-2019 Through 04-05-2019.

Gabriele Sarti, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberof-Arenas, Malvina Nissim, and Arianna Bisazza. 2025. Qe4pe: Word-level quality estimation for human post-editing. *Preprint*, arXiv:2503.03044.

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

Randy Scansani, Silvia Bernardini, Adriano Ferraresi, and Luisa Bentivogli. 2019. Do translator trainees trust machine translation? an experiment on post-editing and revision. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 73–79, Dublin, Ireland. European Association for Machine Translation.

Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 2024. The illusion of competence: Evaluating the effect of explanations on users' mental models of visual question answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19459–19475,

Miami, Florida, USA. Association for Computational Linguistics.

Hervé Spechbach, Johanna Gerlach, Sanae Mazouri Karker, Nikos Tsourakis, Christophe Combescure, Pierrette Bouillon, and 1 others. 2019. A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR medical informatics*, 7(2):e13167.

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

Aaquib Tabrez, Matthew B. Luebbers, and Bradley Hayes. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*.

Masaru Tomita, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki. 1993. Evaluation of MT systems by TOEFL. In *Proceedings of the Fifth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Kyoto, Japan.

Brendan Tomoschuk, Victor S Ferreira, and Tamar H Gollan. 2019. When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, 22(3):516–536.

Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, Lei Ding, Jay Mehta, Nathan Grinnell, Li Liu, Sijia Zhong, Devanathan Nallur Gandamani, Xinyi Tang, Rohan Ghosalkar, Celeste Shen, and 4 others. 2025. A survey on human-ai collaboration with large foundation models. *Preprint*, arXiv:2403.04931.

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

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*.

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 Martindale, Charlotte Vaughn, Ge Gao, and Marine Carpuat. 2025b. Toward machine translation literacy: How lay users perceive and rely on imperfect translations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kaige Xie, Sarah Wiegreffe, and Mark Riedl. 2022. Calibrating trust of multi-hop question answering systems with decompositional probes. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2888–2902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 Additional Details of Quality Estimation Setup and Proficiency Level

In our post-survey, we ask participants to self-assess their source language (French) proficiency on a five-point scale (Figure 7). We group the first three responses as basic, the fourth as intermediate, and the fifth as fluent. Since self-reported language proficiency is often unreliable (Tomoschuk et al., 2019; Edele et al., 2015), we adjust participants' proficiency levels based on their QE performance rather than using raw survey responses. As described in Section 4.3, each participant completes four QE tasks (one before every four QA tasks), in which they rated translation adequacy on a 1–6 scale for each segment (Figure 6). In parallel, we compute XCOMET scores for the gold English transcription and the translated text (Section 4.2), converted to a 1–6 scale (Table 3). Participants' accumulated differences between their ratings and the XCOMET scores serve as their QE performance.

Based on this measure, we adjust initial proficiency groups: advanced participants with accumulated differences above 30 are demoted to intermediate; intermediate participants above 40 are demoted to basic; intermediate participants below 20 are promoted to advanced; and basic participants below 30 are promoted to intermediate. The final distribution of adjusted proficiency groups appears in Table 1.

Range of x	Scale
$x < 0.60$	1
$0.60 \leq x < 0.80$	2
$0.80 \leq x < 0.90$	3
$0.90 \leq x < 0.95$	4
$0.95 \leq x < 0.98$	5
$x \geq 0.98$	6

Table 3: Mapping from XCOMET score x to 1–6 scale.

B Statistical Significance

We present statistical tests of user scores across proficiency levels (Figure 8). Scores for fluent users differ meaningfully from those of both basic and intermediate groups, while intermediate users perform slightly better than basic users, though the difference is not substantial. To examine the role of explanation, we compare user scores across assistance conditions (Figure 9), focusing on the intermediate group to control for proficiency effects. Transcription explanation shows clear improvement over the default condition and performs better than error span explanations, whereas error span yields higher scores than the no-explanation baseline, but the difference is not statistically meaningful. This supports our conclusion that transcription explanation effectively helps users refine their mental models, whereas error spans are less suitable for fostering well-formed ones.

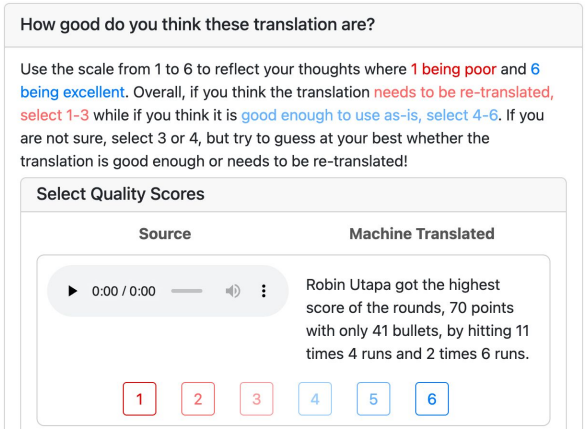


Figure 6: Interface of the additional quality estimation task. Every four QA items, we ask users to rate the speech translation before answering the QA task. These QE ratings are later used to estimate their source-language fluency.

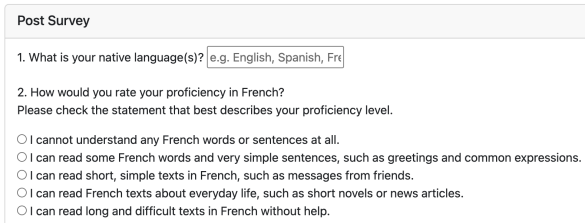


Figure 7: Part of our post-survey, asking about French proficiency.

C Details of Interface

We provide the main instructions shown to the users (D.1) and the instructions for collecting the reflection notes (D.2).

D Dataset Details

For our studies, we used 16 reading comprehension sets from 2M-BELEBELE (Costa-jussà et al., 2025). All data is in English or French. The dataset is publicly available and our usage in this paper is thus within its intended use. We did not collect the dataset ourselves, so we did not check it for PII.

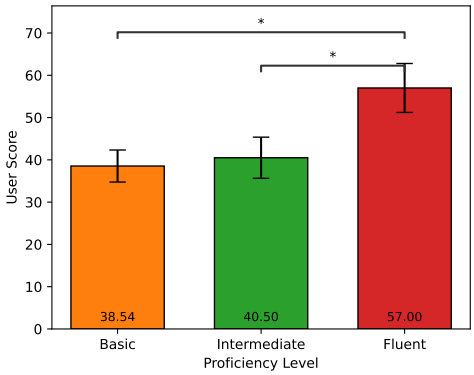


Figure 8: User scores across proficiency levels. * indicates p-value < 0.05.

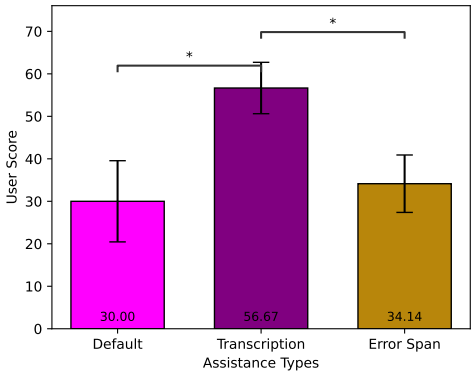


Figure 9: User scores across explanation conditions for intermediate participants. * indicates p-value < 0.05.

D.1: Main Instruction

Machine translated passage will be consumed by the question answering (QA) system to answer the question. Select only the sentences that you believe need to be re-translated by a professional translator because they may have MT errors in key information, so that the QA system can answer the question correctly with the final passage. If you think a machine translated sentence is good enough to use as-is, it's better not to select it to maximize your reward: even in some cases, selecting nothing is okay.

D.2: Instruction for Reflection Note

Playing the game with these questions in mind will help you make better decisions. We encourage you to jot down a brief note after each item as a reflection. You'll revisit these same questions at the end of the game.
1. What kinds of things do you think the MT system used here is likely to handle correctly or incorrectly?
2. What types of errors do you think this MT system is prone to make?
3. What characteristics have you learned to look for when identifying incorrect translations?