Towards Trustworthy AI-Mediated Communication Across Languages and Cultures

Dayeon Ki
University of Maryland
dayeonki@umd.edu

Abstract

A socio-technical gap exists between how NLP systems are developed and evaluated and how people use them in practice. To help close this gap, I propose a direction for scientific progress in NLP centered on advancing *trustworthy* AI-mediated communication between humans, using cross-lingual and cross-cultural interaction as a stress test for this goal—settings where common ground is hard-won, miscommunication can go unnoticed, and human users often lack the means to independently evaluate AI outputs. I outline a research agenda emphasizing two complementary requirements spanning both sides of the interaction. On the model side, I study how multilingual systems access and use knowledge across languages, and when they systematically privilege sources in certain languages. On the user side, I design decision-support mechanisms and evaluate how they shape user's reliance on imperfect outputs. Taken together, these results motivate future work for aligning multilingual NLP with real communicative practice, with the goal of building AI systems that more reliably serve diverse communities. This paper summarizes and draws heavily on my PhD thesis proposal.

1 Introduction

Communication across linguistic and cultural boundaries has long been central to human society—enabling trade, diplomacy, and everyday relationships (Pratt, 1991). Throughout history, humans have devised various ways to bridge these divides, including more recent developments in Natural Language Processing (NLP) for contributing multilingual Artificial Intelligence (AI) systems that mediate communication at scale. As these systems have been trained on increasingly diverse multilingual corpora (Conneau and Lample, 2019; Conneau et al., 2020), they have grown from supporting simple Machine Translation (MT) into *general*-purpose models used for a wider range of cross-lingual and cross-cultural support. People now use them in diverse communicative settings: from direct mediation of conversations between speakers without a shared language to more indirect assistance for interpreting unfamiliar cultural practices (Tamkin et al., 2024).

This shift also creates a mismatch with how progress in multilingual NLP is often measured. Despite widespread real-world use, many systems are still optimized and validated primarily via average performance on *decontextualized* benchmarks (Ackerman, 2000), which tend to reward fluent and adequate-sounding outputs (Bender et al., 2021a). Yet uneven capabilities and behaviors across languages remain visible (e.g., MT (Goyal et al., 2022), Question-Answering (QA, Li et al. (2025b))), and these systems are still under-tested in more open-ended *communicative* settings users now attempt, such as language learning and multi-turn interaction. Therefore, it remains difficult to characterize their impact in practice: how users perceive, interpret, and act on system outputs in real-world downstream decisions (Lee and See (2004)).

With this motivation, I argue that advancing trustworthy AI-mediated communication across languages and cultures is needed (§2.1). Here, *trustworthiness* is not achieved by building more capable models alone, nor by placing the burden on users to independently vet system outputs (§2.2). Instead, it requires complementary progress on both sides of the interaction: models that can help users establish common ground across differences in backgrounds and communicative goals (Clark and Schaefer, 1989), and user-facing designs and interaction strategies that support users' informed reliance on inevitably imperfect outputs. This paper investigates work along both threads.

On the model side, I begin with a core requirement for multilingual systems acting as communication mediators: knowledge parity across *any* languages and cultures (§3.1-3.2). The claim

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 45–59

is not simply that systems should achieve high task performance, but that they should *ground* their outputs in evidence in ways that give users equitable access to multilingual knowledge sources (Blasi et al., 2022), without systematically privileging certain languages during generation. To examine this, we develop a framework for measuring how models rely on multilingual evidence in retrieval-augmented generation (RAG) pipelines and empirically show a strong preference for English sources, even when they are irrelevant to user query (§3.3).

On the user side, trustworthy mediation also requires empowering the communication participants (§4.1). Yet using AI reliably is uniquely challenging in cross-lingual settings: users often lack practical ways to independently assess system outputs (e.g., evaluating a translation in a language they do not understand; Mehandru et al. (2023)), and these outputs are typically *not* direct predictions for downstream decisions, instead serving as inputs to many *implicit* judgments (e.g., Is the translation good enough to share with a friend? To translate an official document?) (§4.2). Accordingly, we propose a new decision support and design human-subject studies to compare it with alternatives, examining how each shapes users' decision-making and reliance on MT outputs (§4.3).

Together, these two threads point to a broader takeaway: advancing AI-mediated communication helps bring empirical findings from the model-side analysis to motivate the questions we ask in human studies, and interaction signals from human studies surface what model-side evaluations are missing, which creates a feedback loop between AI system development and real-world communicative practices. I conclude by distilling lessons from both threads and outlining forward-looking research directions toward multilingual AI systems that more reliably serve diverse communities (§5).

2 Background

I first establish the conceptual foundations for AI-mediated cross-lingual and cross-cultural communication (§2.1) and define trustworthiness in this context, framing it as a property jointly shaped by both the AI system and the human user (§2.2).

2.1 Why AI-Mediated Communication?

Communication may seem like a simple exchange of ideas through words that carry meaning (Reddy, 1979), yet it is fundamentally a collaborative process (Grice, 1975; Allwood, 1976; Bohm and Weinberg, 2004). As Clark and Brennan (1991) describe, successful communication requires coordination between interlocutors, both in content and in process, through the continual establishment of *common ground*: a set of mutual beliefs, presuppositions, and shared background knowledge that provide the necessary context for understanding (Stalnaker, 1972; Clark and Schaefer, 1989).

Consequently, when the interlocutors come from different linguistic or cultural backgrounds, establishing this common ground becomes substantially harder (Hall, 1959; Thomas, 1983; Hershcovich et al., 2022). For instance, when a Korean speaker wishes to communicate their plans for making *Songpyeon* at their grandparents' house for the upcoming *Chuseok*[1] with their American friend, several challenges arise. If they do not share a language, they would likely rely on translation tools even to initiate the conversation. Even when they understand the words, the concept of *Songpyeon* or *Chuseok* carry culturally specific associations with no direct counterpart in American culture, which requires the Korean speaker to explain or provide analogies to help their friend understand. In such cases, human interlocutors often construct a "third space," a communicative middle ground between two cultures (Planken, 2005), by, for example, describing *Chuseok* as a "Korean Thanksgiving" or using pragmatic strategies such as paraphrasing or providing brief clarifications (House, 2003).

As more complex cross-lingual and cross-cultural interactions grow common, NLP has researched to build multilingual AI systems as mediators that help people navigate these communicative gaps once bridged by human strategies at scale. However, these systems do more than transmit information, they actively *shape* how meaning is constructed and whose perspectives are amplified, and as they increasingly participate in human sense-making in the real-world, ensuring that AI-mediated cross-lingual and cross-cultural communication is trustworthy becomes an important goal.

2.2 What Constitutes Trustworthiness?

The notion of trustworthiness is not unique to AI; similar concepts appear across diverse disciplines, though their emphases differ. Across these fields, scholars generally agree perceived trustworthiness,

[1] *Chuseok* is the Korean harvest festival, which occurs on the 15th day of the 8th lunar month. *Songpyeon* is a traditional food eaten during *Chuseok*.

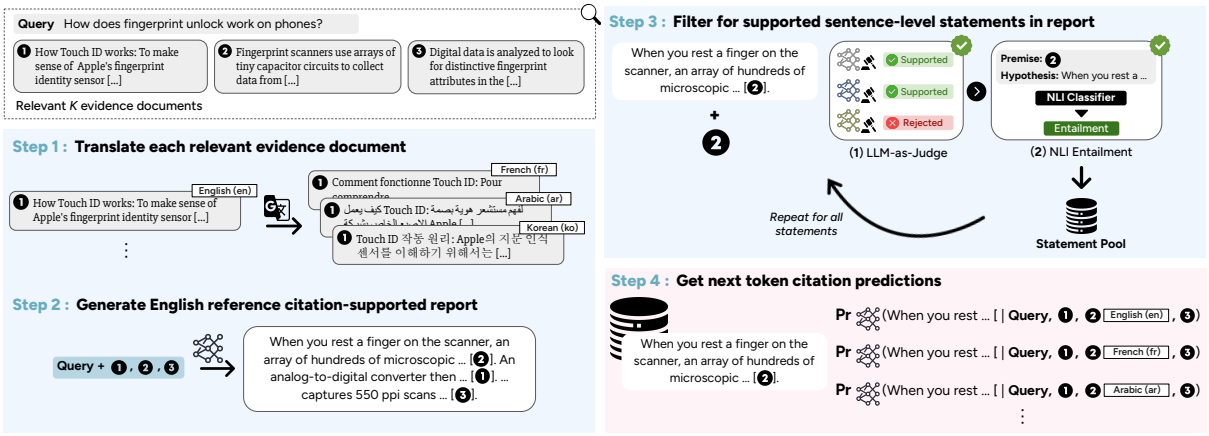

Figure 1: Overview of our approach measuring language preference using model internals. Synthetic data generation: Given an English query and its relevant evidence documents, we translate the documents into target languages and generate reference citation-supported reports (sentence-level statements with citation IDs). Measurement method: We detect language preference when next-token prediction accuracy for the correct citation ID drops as the language of the cited document varies.

whether of a person or a system, is multidimensional, encompassing *competence*, benevolence, and integrity: the ability to perform well, goodwill to act in others' interests, and adherence to ethical principles such as honesty and fairness (Mayer et al., 1995; Mishra, 1996; Rousseau et al., 1998). In sociolinguistics, particularly in linguistic anthropology and intercultural communication, trustworthiness is closely tied to authentic representation—accurately conveying social and cultural meanings—as well as to communicative competence, the ability to use language appropriately in specific social contexts (Gumperz, 1970, 1982; Fox, 1997). In Human-Computer Interaction (HCI), researchers have developed various ways to measure human trust in AI systems, often through trust-related behaviors (Vereschak et al., 2021).[2] One central indicator for trust is *appropriate reliance*, where a user's level of trust matches the system's true capabilities (Lee and See, 2004).

Building on these perspectives, I define trustworthy AI mediators in cross-lingual and cross-cultural communication as systems that (1) possess and appropriately use the knowledge needed to establish common ground between users of different languages and cultures, and (2) support users in making informed decisions from system outputs, even when those are in languages they are not fluent in.

The following sections study each thread in turn. I first lay the groundwork for what hinders and what is required for balanced knowledge coverage across languages and cultures. I then introduce a framework for measuring models' language preferences in evidence selection during RAG (§3). Next, I discuss how user reliance on AI systems has been conceptualized, and propose and evaluate decision supports in helping users engage more reliably with imperfect outputs (§4).

3 Model-Side Thread

I review the prevailing paradigm for understanding knowledge disparities across languages and its limitations (§3.1-3.2), then introduce a framework that probes model internals to examine language preference in multilingual knowledge access (§3.3).

3.1 Understanding Knowledge Disparities

Despite training data becoming increasingly linguistically diverse, large-scale pre-training remains heavily Anglocentric, with English comprising 80–90% of the corpora (Touvron et al., 2023; Grattafiori et al., 2024). As Joshi et al. (2020) point out, current AI systems still disproportionately center on a small subset of languages—often geographically clustered and drawn from a few dominant language families, such as English, Chinese, and Spanish—in both training and evaluation. Meanwhile, many others, such as Zulu or Fijian, remain excluded, creating a typological echo chamber. This persistent "data problem" (Aharoni et al., 2019) is represented well in the language distributions of pre-training corpora for widely used models, which remain heavily skewed toward English

[2] In HCI, "trust" refers to human users' reactions or attitudes toward an AI system, whereas "trustworthiness" concerns the properties of the system itself (Vereschak et al., 2024).

and other high-resource languages (e.g., Chinese).[3]

This disparity in language representation within training data leads to performance gaps across tasks when queries are posed in different languages. Accuracy and perplexity often worsen for certain languages, even when models are evaluated on the same examples translated into different languages (Zhang et al., 2023; Jin et al., 2024; Li et al., 2025b). Beyond performance disparities, uneven language coverage further results in unequal access to knowledge across languages (Bender et al., 2021b; Yu et al., 2022; Feng et al., 2024), wherein information expressed in higher-resource languages becomes more accessible and frequently amplified (Phillipson, 2018). As a result, models exhibit systematic language preference—a tendency to favor certain languages when accessing or eliciting knowledge—which ultimately creates differences in response quality (Boughorbel and Hawasly, 2023), consistency (Dong et al., 2025), and dispute resolution (Li et al., 2024) for users across languages.

3.2 Limitations of Prior Work

Prior work has examined language preference in RAG pipelines: whether models tend to retrieve (Telemala and Suleman, 2022; Yang et al., 2024; Amiraz et al., 2025) or rely on evidence written in certain languages during generation (Park and Lee, 2025; Sharma et al., 2025; Li et al., 2025a). Existing approaches to measure language preference in multilingual RAG (mRAG), however, often fail to capture citation correctness. In short-form RAG, preference has been estimated via information overlap (Sharma et al., 2025) or embedding similarity (Park and Lee, 2025), which do not directly account for correctness. In long-form RAG, where outputs contain in-line citations (Zheng et al., 2025; Xu and Peng, 2025), preference has typically been measured by comparing citation frequencies against the language distribution of retrieved documents. This signal is coarse and confounded by the relevance and informativeness of multilingual sources (C1) and in-line citations are prone to hallucinations (Gao et al., 2023; Zhang et al., 2025), making it unclear whether observed preferences reflect true attribution or spurious citations (C2).

[3]For instance, GPT-3 (Brown et al., 2020) has 92.7% of the training tokens in English; LLaMA-2 (Touvron et al., 2023) and LLaMA-3 (Grattafiori et al., 2024) has 89.7% and approximately 92% of pre-training data in English, respectively.

3.3 Our Approach

Method. In Ki et al. (2026), we address both challenges by proposing a controlled methodology for measuring language preference using model internal metrics. As illustrated in Figure 1, we first construct a synthetic multi-parallel dataset of relevant documents, which allows us to isolate the effect of language while controlling for other factors such as document content and relevance (**Step 1+2**; addresses C1). Citation correctness is then verified through a two-step filtering process (**Step 3**; addresses C2). Next, we compare the accuracy of next token citation predictions (e.g., predicting "2" for document ID 2) while varying the language of the same cited document and keeping other variables fixed, including the language of remaining documents, document positions in the input context, and the query language (**Step 4**). Differences in citation accuracy between languages indicate a preference for the higher-accuracy language.

Experiment Setup. We use ELI5 dataset (Fan et al., 2019) of long-form questions from the Reddit forum "Explain Like I'm Five". We study eight languages representing diverse range of resource levels and number of speakers: Arabic (ar), Bengali (bn), Spanish (es), French (fr), Korean (ko), Russian (ru), Swahili (sw), and Chinese (zh). We use six open-weight models varying in degree of multilinguality: LLAMA-3.1 8B and LLAMA-3.3 70B (Grattafiori et al., 2024), QWEN-3 8B and 14B (Yang et al., 2025), GEMMA-3 27B (Team et al., 2025), and AYA23 8B (Aryabumi et al., 2024).

Results. We address the overarching question: Do models preferentially cite documents in certain languages during long-form mRAG? To further inform building more robust systems, we empirically address two questions: (1) What factors amplify language preference? and (2) Is citation behavior driven more by document relevance or language? Our main findings can be summarized as follows:

- **Evidence of an English preference:** As shown in Table 1, we observe a pronounced tendency to cite English documents (with the highest citation accuracy) when the query is in English across all models. This preference amplifies when the cited document is in a lower-resource language (e.g., Bengali, Swahili).
- **Language outweigh relevance:** We show that models frequently cite English documents even when they are irrelevant to the query as

Language	LLaMA-3.1 8B	Qwen-3 8B	Aya23 8B	Qwen-3 14B	Gemma-3 27B	LLaMA-3.3 70B
English	67.4	62.6	60.0	83.0	86.2	85.9
French	62.9 (-4.49)	48.4 (-14.2)***	48.5 (-11.5)***	76.0 (-7.04)***	79.0 (-7.21)**	77.4 (-8.50)***
Russian	62.1 (-5.30)*	50.4 (-12.2)***	48.1 (-11.9)***	74.8 (-8.17)***	77.1 (-9.12)***	74.5 (-11.4)***
Spanish	62.1 (-5.32)*	51.9 (-10.7)***	49.1 (-10.9)***	77.4 (-5.61)*	80.2 (-6.04)**	76.0 (-9.90)***
Korean	61.7 (-5.68)*	49.7 (-12.9)***	42.2 (-17.8)***	70.3 (-12.7)***	77.5 (-8.71)***	69.2 (-16.7)***
Chinese	59.9 (-7.51)*	49.2 (-13.4)***	46.3 (-13.7)***	73.5 (-9.49)***	75.4 (-10.8)***	74.1 (-11.8)***
Arabic	59.5 (-7.91)**	47.6 (-15.0)***	43.2 (-16.8)***	72.6 (-10.4)***	78.4 (-7.82)***	67.3 (-18.6)***
Bengali	56.6 (-10.8)***	41.3 (-21.3)***	27.2 (-32.8)***	65.4 (-17.6)***	77.9 (-8.33)***	68.8 (-17.1)***
Swahili	53.0 (-14.4)***	30.4 (-32.2)***	22.4 (-37.6)***	54.7 (-28.3)***	74.0 (-12.2)***	67.3 (-18.6)***

Table 1: Citation accuracies (%) by model and language. We present mean accuracy values with the difference to English accuracy in subscript. Pairwise two-sided t-tests with Bonferroni correction are performed to compare accuracy between English and the target language, with the null hypothesis that the mean citation accuracy is equal across languages. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant. Color indicates the magnitude of accuracy difference: largest, second largest, others. Columns: increasing model size; rows: decreasing accuracy difference (of first model).

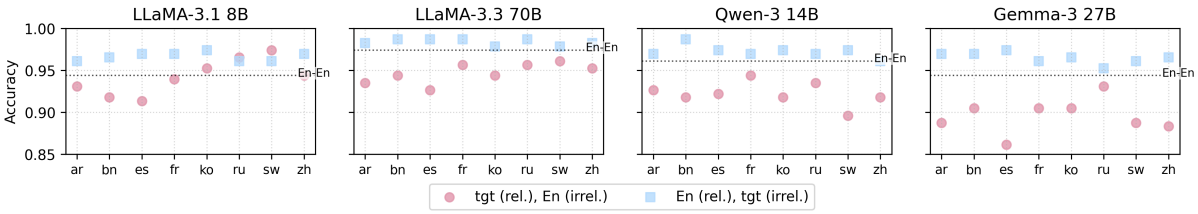

Figure 2: Citation accuracy per model with one relevant (rel.) and one irrelevant evidence document (irrel.) in different languages. We test three different conditions: (1) **En-En**: Both relevant and irrelevant documents are in English; (2) **tgt-En** (●): Relevant document in the target language, irrelevant document in English; (3) **En-tgt** (■): Relevant document in English, irrelevant document in the target language. Models trade off document relevance for language preference.

shown in Figure 2, suggesting that language itself exerts a stronger influence than document relevance in long-form mRAG.

Takeaways. Taken together, our findings show that model internals can expose systematic citation behaviors in mRAG systems, raising inclusivity concerns in multilingual knowledge access, where models not only *favor* certain languages, but may also *trade-off* evidence quality in doing so. These disparities illustrate how imbalanced language representation in training data shapes what knowledge multilingual AI systems access, prioritize, and ultimately use to justify their outputs.

4 User-Side Thread

The model-side thread reveals systematic imperfections in how multilingual AI systems access and use knowledge, but whether and how these impact real users attempting to communicate across languages remains an open question that controlled benchmarks cannot answer alone. Addressing this requires shifting focus from model to user behavior—from *what* the system outputs to *how* users interpret, act on, and are misled by those outputs.

I first review how prior work has measured and designed human interaction with, and reliance on, AI systems (§4.1), and how cross-lingual communication setup poses unique challenges (§4.2). I then propose a new form of decision support and evaluate its impact through human studies (§4.3).

4.1 Human Reliance on AI Systems

People increasingly use AI as decision support in everyday settings, such as video recommendation or search autocompletion (Bunt et al., 2012), and photo organization (Amershi et al., 2019). In these contexts, AI typically plays a supportive role by providing *direct* predictions or explanations in various formats (Bussone et al., 2015; Buçinca et al., 2021; Wang et al., 2022). Such feedback is intended to help users calibrate when and how much to rely on system outputs (Lai et al., 2023).

A growing body of work therefore examines the nature of human reliance on AI systems (Lai et al., 2023), especially in decisions involving risk and uncertainty (Jacovi et al., 2021). To operationalize trust in measurable terms, prior works often study users' reliance behavior (de Fine Licht and Brülde, 2021), commonly defined as the "decision to follow someone's recommendation" (Vereschak et al., 2021). As such, a central challenge in human-AI

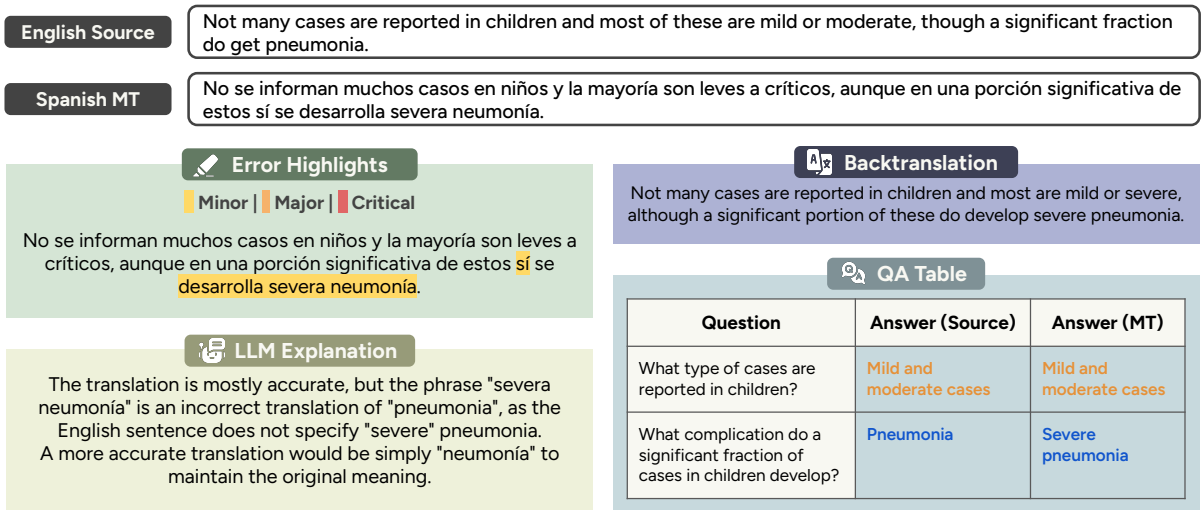

Figure 3: Overview of our tested quality feedback in human study. During the AI-Assisted step, each treatment group participant is presented with an English source, Spanish MT, and one of four randomly assigned quality feedback types. Error Highlights: We adopt a QE system, xCOMET-XXL (Guerreiro et al., 2024), to generate error annotations and display error spans with color-coded legend; LLM Explanation: We use LLAMA-3.3 70B-generated textual explanations of overall MT quality; Backtranslation: We use Google Translate to backtranslate Spanish MT to English; QA Table: We use LLAMA-3.3 70B.

interaction is achieving *appropriate* reliance: helping users accept correct AI advice while rejecting incorrect ones (Eckhardt et al., 2025).

4.2 Limitations of Prior Work

While human reliance on AI systems has been extensively studied in traditional AI-assisted decision-making tasks, extending this concept to cross-lingual communication introduces new challenges. For example, in the context of MT, the role of the AI system (i.e., MT system) takes on a different character since (1) monolingual users often lack the mechanisms to reliably assess MT quality, and (2) the AI prediction (i.e., MT output) is *not* a direct prediction for the user's decision-making task. Given this difference, we focus 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 backtranslation (Agrawal et al., 2022), error highlighting that flags problematic spans in MT (Rubino et al., 2021; Briakou et al., 2023), and textual explanations (Fomicheva et al., 2022; Xu et al., 2023). However, such feedback can be hard to interpret and often fails to convey how mistranslations affect real users (C1). Moreover, only a small number of human studies have evaluated how quality feedback influences user decision-making and reliance in MT (Zouhar et al., 2021; Mehandru et al., 2023), where findings remain mixed, and systematic comparisons against newer feedback mechanisms are still lacking (C2).

4.3 Our Approach

Method. In Ki et al. (2025a), we propose a question generation and answering (QG/QA) framework grounded in the idea that a translation is unreliable if key questions about the source yield different answers when derived from the source text or the backtranslated MT. We hypothesize that these QA pairs foreground the *functional* consequences of potential errors, rather than offering a mechanistic account of what is wrong (Lombrozo and Wilkenfeld (2019); addresses C1). This design also aligns with the view of explanations as *social*—facilitating knowledge transfer through interaction, where users can weigh evidence in light of their existing beliefs (Miller, 2019).

We then conduct a between-subjects human study in Ki et al. (2025b) 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 (i.e., "Is the Spanish translation good enough to safely share with your Spanish neighbor?"). For each of 20 examples, participants first make a binary shareability judgment (Safe to share as-is/Needs bilingual review before sharing)[4] and self-

[4]We use the notion of shareability to capture not only perceived MT quality but also the potential *risk* of miscommu-

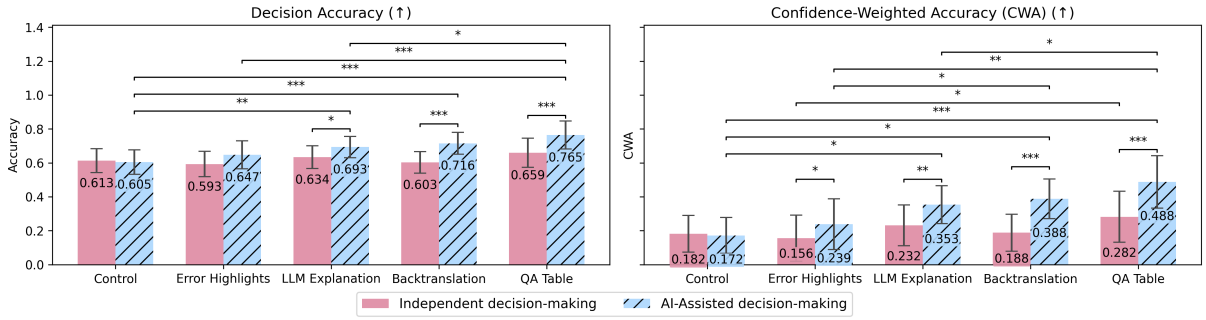

Figure 4: Average decision accuracy (*left*) and CWA (*right*) for each condition. Paired-sample t-tests are performed to compare independent and AI-assisted performance and linear mixed-effects ANOVA with Bonferroni corrections. *: significant with $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; non-marked: not statistically significant.

report confidence (Independent) and then reassess the same example with a randomly assigned quality feedback (AI-Assisted). We compare four types of quality feedback as shown in Figure 3, grouped by their degree of explicitness: error highlights and LLM explanation provide *explicit* assessments of MT quality, whereas backtranslation and QA table offer *implicit* support by guiding participants to compare MT inputs and outputs (addresses C2).

Experiment Setup. We construct 20 English-Spanish examples in the COVID-19 domain by introducing targeted linguistic perturbations into Spanish MT outputs. We categorize each perturbations as either minor or critical based on the potential real-world impact of the resulting error. We recruit 91 U.S.-based participants who self-identify English as their first, primary, and fluent language. Self-reported monthly MT usage varies: 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.

For dependent variables, we measure (1) decision accuracy, by comparing participants' shareability judgment to the gold label; (2) confidence-weighted accuracy (CWA), which combines accuracy with self-reported confidence via confidence weighting (Ebel, 1965; Mehandru et al., 2023) to capture whether participants made the correct decision weighted by their confidence in that decision; and (3) switch percentage, the rate at which participants change their decision after viewing AI feedback (Srivastava et al., 2022; He et al., 2023). Following Schemmer et al. (2023), we compute

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

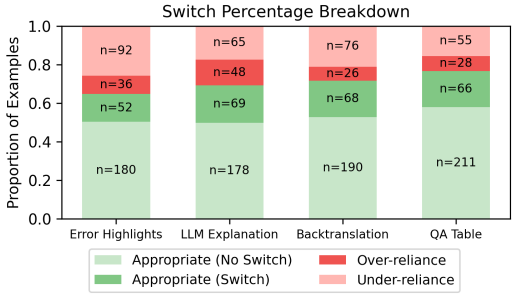

Figure 5: Switch percentage breakdown per quality feedback type. Each shows appropriate, over-, and under-reliance.

three reliance outcomes: (i) over-reliance, switching from a correct to an incorrect decision after feedback; (ii) under-reliance, failing to switch from an incorrect to a correct decision after feedback; and (iii) appropriate reliance, either correcting an incorrect decision after feedback (switch) or maintaining a correct one (no switch).

Results. Our main findings are as follows:

- **Providing any feedback helps:** As shown in Figure 4, all four treatment conditions improve decision accuracy and CWA in the AI-Assisted step relative to the Independent step.
- **QA table yields the largest gains:** Among the four feedback types, QA table shows the strongest and most consistent effects overall.
- **Implicit feedback has lower over-reliance:** As illustrated in Figure 5, both implicit quality feedback (backtranslation and QA table) yield higher appropriate reliance and lower over-reliance than the explicit feedback types. Across conditions, participants are more likely to *maintain* their initial decisions than to switch: under-reliance consistently exceeds over-reliance, and appropriate reliance (no switch) exceeds (switch).

Takeaways. Together, our findings suggest that using QA pairs as quality feedback is a promising direction for supporting cross-lingual communication. More broadly, these results underscore the value of interpretable, *user-driven* 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), rather than having the system explicitly prescribe what to do.

5 Concluding Thoughts

I propose a direction for scientific progress in NLP centered on advancing AI-mediated communication across languages and cultures. This agenda requires two complementary threads: (1) models that help users establish common ground across differences in background and communicative goals, and (2) interaction strategies that support users' informed reliance on inevitably imperfect outputs.

Lessons Learned. On the model side (§3), the central lesson is that a trustworthy AI mediator requires knowledge parity *in use*—counteracting the imbalanced language representation in training data. The core issue here is not simply whether a model can achieve high multilingual task performance, but whether it provides equitable access to multilingual sources when grounding its claims. By measuring language preference through model internal signals in controlled settings, this thread moves beyond coarse proxies (e.g., citation frequency or surface overlap) toward diagnostics that can isolate when language itself drives evidence selection and when it induces trade-offs with evidence quality. This makes visible a failure mode: even when multilingual knowledge sources exist, systems may still systematically privilege those written in particular languages, shaping which evidence is surfaced and which perspectives are amplified. Methodologically, the broader implication is that evaluating multilingual systems as communication mediators requires attention to the full pipeline, retrieval, selection, generation, and reasoning, to understand not only *what* a model outputs, but *how* it arrives at its predictions.

On the user side (§4), the central lesson is that trustworthy mediation is not achieved by model capability alone; it also requires interaction designs that preserve users' *agency* (Shneiderman, 2022): users need support to assess risks, goals, and consequences for themselves, rather than prescribing decisions. We study this in a realistic cross-lingual communication scenario, where reliable use is uniquely challenging since users often lack practical ways to assess outputs and system outputs do not map directly onto downstream decisions. We propose QA pair-based feedback and show, through human-subject studies, that this interpretable, user-driven support can improve decision accuracy with better-calibrated confidence and promote more appropriate reliance on MT outputs.

Looking Forward. These lessons motivate several future research directions that better aligns multilingual NLP with communicative practice:

- **Evaluate systems in communicative contexts, not just on benchmarks.** For mediator-like systems, we should complement model-centric metrics with user-centered outcomes (e.g., decision accuracy, over/under-reliance, and downstream risk) and explicitly measure communicative failure modes that benchmarks often miss (e.g., misunderstandings, repair behavior, and conversation breakdowns).
- **Understand systems' *reasoning* processes.** A promising next step is to characterize *why* models behave differently across languages, both in evidence use and final outputs, and to translate that understanding into user-facing guidance. Reasoning traces are one candidate bridge: if designed carefully, they could help users retrace and oversee why the model made a particular prediction, and potentially enable them to learn patterns they can extrapolate to future decisions (Holzinger et al., 2023).
- **Build AI literacy for human users.** While developers may be aware of a system's weaknesses, these limitations are rarely communicated to end users. To enable functional, user-driven feedback, interfaces should explicitly communicate the systems' capabilities and blind spots through actionable signals, clarifying what the system can and cannot be expected to do in the current setting.
- **Examine effects on interpersonal dynamics.** Beyond individual decision-making, AI-mediated communication can reshape how people coordinate with one another. Interpersonal adaptation—how interlocutors adjust phrasing, clarify intent, and negotiate meaning—is central to establishing common

ground, and AI mediation may shift when and how this occurs. A key direction is to evaluate mediator systems for their effects on conversational dynamics (e.g., attribution, accountability, perceived effort) and on longer-term interpersonal relationships.

The broader goal is to move multilingual NLP toward systems that reliably help people establish common ground and make reliable decisions across languages and cultures. This work develops methods, measurements, and interaction designs aimed at supporting diverse communities in practice.

Acknowledgments

I would first like to thank my PhD thesis advisor Marine Carpuat, as well as my committee members Kevin Duh, Rachel Rudinger, and Fumeng Yang. Many thanks also to all coauthors on the projects reviewed in this paper, including (in alphabetical order) Daniel Khashabi, Dawn Lawrie, Eugene Yang, and Paul McNamee. Last but not least, thanks to all the members of the CLIP lab at the University of Maryland for their constructive feedback and support in developing this work.

References

Mark S Ackerman. 2000. The intellectual challenge of cscw: the gap between social requirements and technical feasibility. *Human–Computer Interaction*, 15(2-3):179–203.

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via back-translation 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.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Jens Allwood. 1976. *Linguistic Communication as Action and Cooperation*. Ph.D. thesis, University of Göteborg, Department of Linguistics, Göteborg, Sweden.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan. 2025. The cross-lingual cost: Retrieval biases in RAG over Arabic-English corpora. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 69–83, Suzhou, China. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *Preprint*, arXiv:2405.15032.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021a. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

David Bohm and Rachel A. Weinberg. 2004. *On Dialogue*, 2nd edition. Routledge.

Sabri Boughorbel and Majd Hawasly. 2023. Analyzing multilingual competency of LLMs in multi-turn instruction following: A case study of Arabic. In *Proceedings of ArabicNLP 2023*, pages 128–139, Singapore (Hybrid). 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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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).

Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, page 169–178, New York, NY, USA. Association for Computing Machinery.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics*, ICHI '15, page 160–169, USA. IEEE Computer Society.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Karl de Fine Licht and Bengt Brülde. 2021. On Defining "Reliance" and "Trust": Purposes, Conditions of Adequacy, and New Definitions. *Philosophia*, 49:1981–2001.

Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang. 2025. Evaluating and mitigating linguistic discrimination in large language models: perspectives on safety equity and knowledge equity. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

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. 2025. A survey of ai reliance. *ACM Comput. Surv.*, 58(6).

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, Miami, Florida, USA. Association for Computational Linguistics.

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.

Christine Fox. 1997. The authenticity of intercultural communication. *International Journal of Intercultural Relations*, 21(1):85–103.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press, New York.

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.

John J. Gumperz. 1970. *Sociolinguistics and Communication in Small Groups*. University of California Language-Behavior Research Laboratory, Berkeley, CA. Language-Behavior Research Laboratory Report.

John J. Gumperz. 1982. *Discourse Strategies*. Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge. Originally published 1982; online publication November 2009.

Edward T. Hall. 1959. *The Silent Language*. Doubleday, Garden City, NY.

Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. 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).

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Bettina Finzel, Ute Schmid, and Heimo Mueller. 2023. Toward human-level concept learning: Pattern benchmarking for ai algorithms. *Patterns*, 4(8):100788.

Juliane House. 2003. English as a lingua franca and its influence on discourse norms in other languages. In Gunilla Anderman and Margaret Rogers, editors, *Translation Today: Trends and Perspectives*, pages 168–179. Multilingual Matters.

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*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn Lawrie, and Kevin Duh. 2026. Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG. In *Forty-third International Conference on Machine Learning*.

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

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025b. Should I Share this Translation? Evaluating Quality Feedback for User Reliance on Machine Translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12080–12103, Suzhou, China. Association for Computational Linguistics.

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*, FAccT '23, page 1369–1385, New York, NY, USA. Association for Computing Machinery.

John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80. Original work published 2004.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is your, my land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025a. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4215–4241, Vienna, Austria. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025b. Language ranker: a metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

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

Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

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. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Aneil K. Mishra. 1996. Organizational responses to crisis: The centrality of trust. In Roderick M. Kramer and Thomas R. Tyler, editors, *Trust in Organizations: Frontiers of Theory and Research*, pages 261–287. Sage Publications, Newbury Park, CA.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating Language Preference of Multilingual RAG Systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5647–5675, Vienna, Austria. Association for Computational Linguistics.

Robert Phillipson. 2018. Linguistic imperialism. *The Encyclopedia of Applied Linguistics*.

Brigitte Planken. 2005. Managing rapport in lingua franca sales negotiations: A comparison of professional and aspiring negotiators. *English for Specific Purposes*, 24(4):381–400.

Mary Louise Pratt. 1991. Arts of the contact zone. *Profession*, pages 33–40. Accessed 25 Oct. 2025.

Michael J. Reddy. 1979. The conduit metaphor. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press, Cambridge.

Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404.

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.

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.

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.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2025. Faux Polyglot: A Study on Information Disparity

in Multilingual Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8090–8107, Albuquerque, New Mexico. Association for Computational Linguistics.

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

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.

Robert C. Stalnaker. 1972. Assertion. In Peter Cole, editor, *Pragmatics*, pages 315–332. Brill.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, and Deep Ganguli. 2024. Clio: Privacy-preserving insights into real-world ai use. *Preprint*, arXiv:2412.13678.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Joseph P Telemala and Hussein Suleman. 2022. Language-preference-based re-ranking for multilingual swahili information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 144–152.

Jenny Thomas. 1983. Cross-Cultural Pragmatic Failure. *Applied Linguistics*, 4(2):91–112.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,

Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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).

Jialu Wang, Yang Liu, and Xin Wang. 2022. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland. Association for Computational Linguistics.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *Preprint*, arXiv:2506.12594.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024. Language fairness in multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2487–2491.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China. 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.