Translation in the Hands of Many: Centering Lay Users in Machine Translation Interactions

Beatrice Savoldi, Alan Ramponi, Matteo Negri, Luisa Bentivogli
Fondazione Bruno Kessler, Italy
{bsavoldi,alramponi,negri,bentivo}@fbk.eu

Abstract

Converging societal and technical factors have transformed language technologies into user-facing applications used by the general public across languages. Machine Translation (MT) has become a global tool, with cross-lingual services now also supported by dialogue systems powered by multilingual Large Language Models (LLMs). Widespread accessibility has extended MT's reach to a vast base of *lay users*, many with little to no expertise in the languages or the technology itself. And yet, the understanding of MT consumed by such a diverse group of users—their needs, experiences, and interactions with multilingual systems—remains limited. In our position paper, we first trace the evolution of MT user profiles, focusing on non-experts and how their engagement with technology may shift with the rise of LLMs. Building on an interdisciplinary body of work, we identify three factors—usability, trust, and literacy—that are central to shaping user interactions and must be addressed to align MT with user needs. By examining these dimensions, we provide insights to guide the progress of more user-centered MT.

1 Introduction

The success of technology hinges on its ability to serve users, and Natural Language Processing (NLP) confronts this challenge as it transitions from an academic pursuit to a set of impactful tools. Among them, MT stands out as a cornerstone application, with current breadth and quality that fostered wider adoption (Wang et al., 2022). Multilingual demands (Moorkens and Arenas, 2024), paired with the accessibility of online systems, has put MT at the forefront of user-facing language technologies. Once confined to professional settings, MT is now used by millions (Pitman, 2021), bringing into its fold an array of *lay users* in contexts ranging from casual interactions (Gao et al., 2015) to critical domains such as healthcare and employment (Patil and Davies, 2014; Dew et al., 2018; Liebling et al., 2022; Valdez et al., 2023).

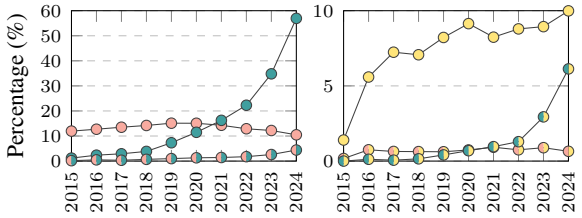

Figure 1: Trend of interest in *machine translation* MT, *language models* LM, *users* U, and combinations thereof in the ACL community over the last 10 years. Besides illustrating the rapid growth of LLM studies, the left panel highlights the increase in MT research incorporating LLMs (MT + LM), while the right panel shows rising attention to users, particularly in LLM-related work (LM + U).[1]

Despite MT's broad reach and potential for social impact in sensitive scenarios (Vieira et al., 2021), still little is known about its evolving relationship with the general public, how non-expert users interact with it, or how it caters to their needs (Carpuat et al., 2025). MT research has mainly focused on modeling advancements and—although translation studies have called for greater attention to end-user perspectives (Guerberof-Arenas and Moorkens, 2023) and related efforts from human-computer interaction (Zhang et al., 2021, 2022)—MT works that actively involve lay people and their experiences are still rare (Mehandru et al., 2023; Briakou et al., 2023).

In the wake of broader calls to bridge MT (Liebling et al., 2021) and language technologies with user-centered research (Heuer and Buschek, 2021; Kotnis et al., 2022), we posit that it is time to fill this gap and focus on how to support interactions between systems and lay people. Arguably,

[1]Details on the ACL Anthology queries are provided in Appendix A. For a complementary view, Figure 2 in the appendix also shows absolute counts in the trends of interest over the last ten years.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13888–13901
November 4-9, 2025

the rise of powerful, instruction-following LLMs (Touvron et al., 2023; Achiam et al., 2023; Gemini et al., 2024; Üstün et al., 2024, *inter alia*) engaging non-experts via chat interfaces has heightened user concerns (see Figure 1, LM + U on the *right*) and underscores the urgency to align with real-world interactions (Haque et al., 2022; Liao and Vaughan, 2023; Szymanski et al., 2024).[2] As MT moves towards LLM-based solutions (see Figure 1, MT + LM *vs* MT on the *left*), these have the potential to redefine how people engage with multilingual systems, challenging traditional task divisions with new paradigms for cross-lingual communication (Ouyang et al., 2023; Lyu et al., 2024).

To set the stage for this shift towards lay users' perspective, we examine the evolution of MT from professional settings to its wide general adoption (§2). We then identify three key factors—*usability* 👆, *trust* ✔, and *literacy* 📖—to ground user interactions with automatic translation tools (§3). Through this lens, we take stock of the current landscape to guide MT research in tandem with users (§4).

We release a curated list of the works discussed in the paper at: https://github.com/hlt-mt/awesome-human-centered-MT.

2 MT and User Evolution

Although online systems have existed for some time (Yang and Lange, 1998; McCarthy, 2004; Somers, 2005), we are now seeing unprecedented volumes of unrevised MT outputs being directly consumed by the public.[3] **Historically, real-world applications of MT often regarded so-called "mixed MT" workflows** (Wagner, 1983), where human intervention serves to revise—i.e. post-edit (Li, 2023)—MT to produce a reliable final translation. Attention to this usage scenario (Church and Hovy, 1993) is reflected in MT development (Green et al., 2014; Bentivogli et al., 2015; Daems and Macken, 2019), interfaces (Vieira and Specia, 2011; Vela et al., 2019), and evaluation (Popović and Ney, 2011; Bentivogli et al., 2016) using *professional* translators as a target. Such a trajectory was also paired with empirical experiments on when MT could support (Koponen, 2016; Moorkens and O'Brien, 2017) or interfere (Federico et al., 2014; Daems et al., 2017) with translators' activity.

The advent of stronger models with expanded language coverage—along with the rise of the Web and personal devices—progressively altered this landscape. **MT consumption has now reached wider adoption by the general public, who directly accesses *raw* MT output** in many diverse scenarios,[4] e.g. to gist content, for multilingual conversations (Pituxcoosuvarn et al., 2020; Pombal et al., 2025), in education (Yang et al., 2021; Yang, 2024), but also in high-stakes domains such as healthcare (Khoong et al., 2019; Valdez and Guerberof-Arenas, 2025), migration (Liebling et al., 2022), and emergency services (Turner et al., 2015).[5] **This shift to unmediated MT has led to a vast, heterogeneous base of *lay users* and, with it, novel desiderata and concerns**. For one, since lay users may have limited to no proficiency in at least one of the involved languages,[6] they are more vulnerable to errors. Mistranslations can lead to discomfort, misunderstandings, and even life-threatening errors (Taira et al., 2021) and arrests (The Guardian, 2017). Besides, non-experts can have requirements and expectations of which little is known, and that cannot be directly informed by existing research on professionals, as shown in the context of LLMs—e.g. Szymanski et al. (2024), see also Figure 1, Appendix B).

Indeed, general-purpose LLMs are calling for more considerations of users and real-world contexts of use, as demonstrated by surveys to understand how people interact with technologies, for which purposes and needs (Tao et al., 2024; Skjuve et al., 2024; Kim et al., 2024b; Stojanov et al., 2024; Bodonhelyi et al., 2024; Wang et al., 2024; Hyun Baek and Kim, 2023, *inter alia*). **Chat-based LLMs** have drawn in millions of users,[7] with their impressive **versatility and engaging interfaces that allow verbalizing requests, also for automatic translation** (Ouyang et al., 2023). As the MT field explores such LLM-based solutions (Zhu et al., 2024; Lyu et al., 2024; Alves et al., 2024, *inter alia*) and integrates MT into more com-

[2] For wider initiatives towards human-centered approaches in the *CL community, we notice the introduction of the Human-Centered NLP track since 2023, as well the HCI+NLP workshop (Blodgett et al., 2024a) and the tutorial on Human-Centered Evaluation of Language Technologies (Blodgett et al., 2024b).

[3] See the rising volume of Google Translate app downloads and words translated with it (Pitman, 2021).

[4] Also leading to a decrease in the demand for language skills and professional work (Frey and Llanos-Paredes, 2025).

[5] e.g. with COVID to compensate for interpreters shortages (Khoong and Rodriguez, 2022; Anastasopoulos et al., 2020).

[6] e.g. the *source* language in gisting and the *target* in communication contexts. See also Nurminen and Papula (2018).

[7] According to OpenAI, in the summer of 2024 ChatGPT reached 200 millions weekly active users.

plex systems, these solutions have the potential to reshape cross-lingual services and user engagement.

While this transition unfolds, overdue research on the experiences of lay users in cross-lingual and MT settings is gaining urgency. We map this gap and call for first steps to fill it.

3 Three Factors for MT Lay Users

Usability The usability of MT systems—how effectively, efficiently, and satisfactorily users can achieve their goals in a given context (ISO, 2018)—is informed and guided by how these systems are evaluated. The field, however, tends towards performance-driven leaderboards (Rogers, 2019), which have been criticized for pursuing abstract notions of *accuracy* and *quality* above the practical *utility* of a model or other relevant values (Ethayarajh and Jurafsky, 2020). These values are often contextual: Parthasarathi et al. (2021) discuss how *robustness* to misspellings might be detrimental if using MT for learning. Also, *faithfulness* is normally key to "MT quality", but in creative contexts like subtitling, *enjoyability* may take precedence over fidelity (Guerberof-Arenas and Toral, 2024).

Standard MT metrics offer coarse scores of generic performance to rank and compare *models*, but are opaque and only assume to inform how useful the model is when embedded within the system the user interacts with (Liebling et al., 2022). And yet, lay people are only involved as evaluators to provide model-centric insights, rather than to inform their experiences (Saldías Fuentes et al., 2022; Savoldi et al., 2024).[8] Furthermore, general-purpose LLMs now confront us with an "evaluation crisis" (Liao and Xiao, 2023), where existing methods and predefined benchmarks for modular tasks may be obsolete, failing to capture real-world downstream contexts. This raises the risk of widening the socio-technical gap, where evaluation practices lack validity and might diverge from human requirements in realistic settings.

Trust To prevent over-reliance on automatic translations, lay users must calibrate an appropriate level of (dis)*trust*. Indeed, they risk accepting potentially flawed translations at face value, and trust may be misplaced when an output appears believable but is inaccurate—an issue that is especially harmful in high-stakes contexts (Mehandru et al., 2023). Prior research on MT has shown that *fluency* and *dialogue flow* can falsely signal reliability (Martindale et al., 2021; Robertson and Díaz, 2022), and LLMs amplify this issue with their overly confident tone, even when incorrect (Xiong et al., 2024; Kim et al., 2024a). As general-purpose models increasingly replace domain-specific applications, providing mechanisms for trust calibration becomes even more urgent (Deng et al., 2022; Litschko et al., 2023). To harness the benefits of MT systems while avoiding over-reliance on flawed translations, lay users often resort to back-translation[9] as a strategy to improve confidence (Shigenobu, 2007; Zouhar et al., 2021; Mehandru et al., 2023). However, back-translation is often performed manually due to the lack of dedicated functionalities, and its soundness remains debated. Another critical factor in fostering appropriate trust is *transparency*—e.g. communicating uncertainty and providing explanations (Liao and Vaughan, 2023). While explainability work is growing (Ferrando et al., 2024), ensuring that explanations are informative and digestible to lay users rather than just developers is not trivial. Moreover, how to effectively integrate such uncertainty signals into the development of translation systems and their user interfaces is still an open question.

Literacy MT-mediation, as a form of human-machine interaction (Green et al., 2015; O'Brien, 2012), should also regard how lay users themselves play a role in improving interactions and apply control strategies to overcome MT limitations. This requires critical agency rather than passive consumption. In this area, prior work (Miyabe and Yoshino, 2010) has shown that preventing the display of potentially flawed translations causes discomfort to users, indicating that they prefer warnings and guidance over outright blocks. But warnings serve as an initial signal; then users should know how to proceed in recovering from MT errors (Shin et al., 2013). To address this, Bowker and Ciro (2019) introduce the concept of MT *literacy*, a digital skill to equip users with the knowledge to interact more effectively with MT.[10] This includes pre-editing input text to mitigate common failures (e.g. us-

[8] This trend might be exacerbated by AI surrogates, which have been suggested as a "replacement" for human participants (Wang et al., 2025; Agnew et al., 2024).

[9] i.e. automatically translating a text to a target language and then back to the source language.

[10] For online materials, see https://sites.google.com/view/machinetranslationliteracy/.

ing short sentences). While literacy workshops proved beneficial to students (Bowker, 2020),[11] reaching more vulnerable populations and underserved languages remains a challenge (Liebling et al., 2020).[12] Focusing on target comprehension, Liebling et al. (2021) explore interfaces with dictionary access and assistive bots.[13] While LLMs encourage participation through chat and interactive queries (Qian and Kong, 2024), their reliability in this role remains uncertain, as LLM-powered systems may impact cognitive attention required for critical engagement (Zhai et al., 2024; Lee et al., 2025). Also, MT literacy must evolve to address new opportunities and failures introduced by LLMs, such as cascading errors across multiple requests.

4 Future Directions and Conclusion

To conclude, we examine directions for future research in traditional or LLM-based MT that integrates lay user perspectives. We map such directions and corresponding recommendations to the three factors outlined in Section §3.

Consider Lay People As Users (👆, ✔, 📖) To gauge how/when users interact with MT as well as current blindspots we should consider their experiences rather than just involve them as manual evaluators. Inspired by monolingual work (Handa et al., 2025), analyzing user logs can help us observe real engagement and preferences. Surveys and *in vivo* research offer qualitative insights into users' perceptions (Zheng et al., 2019; Robertson and Díaz, 2022). To this aim, it is essential to avoid two main pitfalls: *i)* exploiting participants (see §6) and *ii)* treating them as a homogeneous group: factors like sociodemographics, education, and stress levels can greatly influence their expectations and interactions (Rooein et al., 2023; Ge et al., 2024).

Design for Usability and Utility (👆) Achieving human-like translations should not be blindly viewed as the ultimate goal—automated text is a means to serve a broader purpose, not an end in itself (Caselli et al., 2021). Prior work has evaluated systems based on their success in guiding human decision-making (Zhao et al., 2024) or by assessing gender bias in MT via user-relevant measures, like time, effort, or economic costs (Savoldi et al., 2024). Research could focus on making measurements more actionable (Delobelle et al., 2024), e.g. to identify usability thresholds below which MT is no longer beneficial. Therefore, we should aim to correlate automated approaches with human-centered measurements to harness the benefits of both.[14] However, this is challenging due to the variability of utility values among users and usages. Multi-metric and multifaceted approaches like HELM (Bommasani et al., 2023) show promise in this area, but future work could further align MT evaluation and design with socio-requirements and prototypical use cases (Liao and Xiao, 2023).

Enrich MT Outputs (✔) In user-facing systems, it is crucial to not only focus on generated translations but also to develop methods for *estimating* and *conveying* uncertainty, ambiguities, and errors to ensure reliable usage (Xu et al., 2023; Zaranis et al., 2024). For instance, Briakou et al. (2023) use contrastive explanations to help users understand cross-linguistic differences, but it is unclear how to disentangle when their approach captures critical errors or simple meaning nuances in the wild. Quality estimation can also warn users in real time about flawed translations, though numeric indicators are hard to interpret to lay users (Miyabe and Yoshino, 2010). Indeed, a key area of future research is how to best communicate digestible information to lay users, e.g. via visualizations.[15] Textual explanations show promise in communicating uncertainty and avoiding over-reliance in LLMs, but the exact language used is relevant (Kim et al., 2024a), and we thus advocate for MT work in this area.

Foster Transparency (✔ 📖) and Agency (📖) Users should have the *option* to be active participants when interacting with MT. In addition to real-time explanations, they could receive clear information about MT's strengths and limitations (e.g. support across languages). The field might adapt transparency tools like *model cards* (Mitchell et al., 2019) into simplified, public-facing versions and support literacy efforts around emerging technologies.[16] To foster *user agency*—the ability to make informed, intentional decisions about MT use—approaches such as gamification (Chen,

[11] For other data literacy initiatives targeting students, see the DataLitMT project (Hackenbuchner and Krüger, 2023).

[12] e.g. see BabelDr for a case of MT design for healthcare involving migrant populations: https://babeldr.unige.ch/.

[13] See the Lara system, integrating the two-box interfaces with a bot: https://laratranslate.com/translate.

[14] e.g. replicability and ecological validity, respectively.

[15] e.g. by highlighting errors or reliable keywords.

[16] For example, see the Elements of AI program: https://www.elementsofai.com/.

2023) could help promote literacy and lightweight critical engagement. Yet, since MT often serves immediate, time-sensitive needs, it remains uncertain whether users always want or are able to engage critically (Buçinca et al., 2021). MT experts are well placed to advance these efforts through interdisciplinary collaboration. For instance, Xiao et al. (2025) investigate how to sustain non-native speakers in influencing the production of their message in MT-mediated communication.

Bridge Interdisciplinary Avenues (👆, ✔, 📖) Incorporating user needs, values, desiderata, and human factors is still in its early stages in NLP. However, disciplines like human-computer interaction (HCI), experimental psychology, and social sciences have established practices to draw from (Liao and Xiao, 2023). These methodologies may take longer to implement, but they yield useful insights, e.g. on people cognition and trust, or to implement user studies. Besides, they offer methods that approximate real-world interactions cost-effectively, e.g. Wizard of Oz tests prior to developing a new method (Goyal et al., 2023), or simulating user actions based on past user data (Zhang et al., 2021). These approaches can be highly useful, but—circling back this section—the fundamental first step remains engaging with end users to understand their needs and behaviors first.

5 Limitations

Factors Our analysis centers on three key criteria. These are not exhaustive of all user-centered concerns, but they serve as a starting point for a research agenda on human engagement with MT. The selection of these criteria was guided, first, by the aim to capture complementary and distinct dimensions of MT usage, namely: *i) usability*—how to align technology with users through model/system adaptation; *ii) trust*—how to calibrate user-MT interactions by addressing dynamics of reliance and confidence; and *iii) literacy*—how to empower users by fostering their ability to engage with and learn about MT. Second, our choice reflects their recurrence in the literature we reviewed and discussed throughout the paper. At the same time, they resonate with broader debates in adjacent fields: in HCI and translation studies, usability is an established quality characteristic (Guerberof-Arenas et al., 2019); in AI governance, the EU AI Act explicitly foregrounds trust/trustworthiness as a core principle[17] and introduces literacy as a requirement for fostering awareness and competence.[18]

Text-to-Text MT We do not unpack the differences between text-to-text MT and other modalities, such as speech translation and multimodal cross-lingual tasks (Papi et al., 2025). While we acknowledge the relevance of these distinctions, we chose to focus on the broadest and most established MT technology. Expanding to other modalities is an important avenue for future work, but our scope was limited by space and focus.

ACL Anthology Query Our trends assessment of prior work on MT, LLMs, and Users—reported in Figure 1—is based on papers published in the ACL Anthology (see Appendix A). While including other sources could have further enriched our trend overview, the Anthology remains the main historical reference point in NLP. Hence, it represents an optimal litmus test for assessing trajectories in the field. Still, throughout the paper, we engage with literature from diverse communities, primarily from *translation studies* and *human factors in computing*, to provide a broader interdisciplinary perspective.

Slower Science Our proposed future directions advocate for user-centered analyses and studies that require more time and resources compared to automated evaluations and *in vitro* experiments, potentially slowing down the research cycle. However, we argue that user-driven insights are crucial and can only yield benefits to align MT with real-world needs and users.

6 Ethics Statement

In this work, we advocate for user-centered MT research by focusing on lay users. First, unlike human-in-the-loop methods (Wang et al., 2021)—which rely on human contributions to enhance model functionality—we prioritize approaches and directions that are intended to serve and benefit users.

Second, we do not conduct experiments with participants in this paper. Hence, we do not discuss ethical best practices for research in this area, though we deem them as indispensable, e.g. obtaining proper ethical approval, securing informed

[17] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[18] https://artificialintelligenceact.eu/article/4/.

consent, and ensuring non-intrusive engagement when working with human participants.

Finally, while we broadly discuss lay users, we do recognize that they actually encompass diverse groups and communities. Many remain underserved by language technologies, particularly speakers of "low-resource" languages, and might face well-known biases in NLP tools related to gender (Savoldi et al., 2025), dialect (Blodgett et al., 2020), or social class (Cercas Curry et al., 2024). Especially when engaging with more vulnerable communities and user groups, it is important to respect their lived experiences, avoid exploitative research practices, and ensure they are not treated as mere data sources but as valued participants and users—see e.g. Bird and Yibarbuk (2024); Ramponi (2024) and Birhane et al. (2022).

7 Acknowledgments

Beatrice Savoldi is supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. The work presented in this paper is also funded by the Horizon Europe research and innovation programme, under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People), and the ERC Consolidator Grant No 101086819.

References

2018. ISO 9241-11:2018 - Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts. Accessed: 2025-02-10.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. The AI gap: How socioeconomic status affects language technology interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18647–18664, Vienna, Austria. Association for Computational Linguistics.

Luisa Bentivogli, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2015. On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):388–399.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao, editors. 2024a. *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Mexico City, Mexico.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024b. Human-centered evaluation of

language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 39–43, Miami, Florida, USA. Association for Computational Linguistics.

Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language models: A user study with chatgpt. *Preprint*, arXiv:2402.02136.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Lynne Bowker. 2020. Chinese speakers' use of machine translation as an aid for scholarly writing in english: a review of the literature and a report on a pilot workshop on machine translation literacy. *Asia Pacific Translation and Intercultural Studies*, 7(3):288–298.

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

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. 2023. Explaining with contrastive phrasal highlighting: A case study in assisting humans to detect translation differences. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21.

Marine Carpuat, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. An interdisciplinary approach to human-centered machine translation. *Preprint*, arXiv:2506.13468.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. Classist tools: Social class correlates with performance in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12643–12655, Bangkok, Thailand. Association for Computational Linguistics.

Yulin Chen. 2023. Using a game-based translation learning app and google apps to enhance translation skills: Amplification and omission. *International Journal of Human–Computer Interaction*, 39(20):3894–3908.

Kenneth W Church and Eduard H Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8:239–258.

Aslihan S. Cifter and Hua Dong. 2009. User characteristics: Professional vs. lay users. In *Include2009 Proceedings*, London. Royal College of Art. Include2009, April 8–10, 2009. ISBN: 978-1-905000-80-7.

Joke Daems and Lieve Macken. 2019. Interactive adaptive smt versus interactive adaptive nmt: a user experience evaluation. *Machine Translation*, 33(1):117–134.

Joke Daems, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in psychology*, 8:1282.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.

Wesley Hanwen Deng, Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Beyond general purpose machine translation: The need for context-specific empirical research to design for appropriate user trust. In *Proceedings of the Workshop on Trust and Reliance in AI-Human Teams at CHI 2022*, TRAIT 2022, New Orleans, LA, USA.

Kristin N Dew, Anne M Turner, Yong K Choi, Alyssa Bosold, and Katrin Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics*, 85:56–67.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. Workingpaper.

Carl Benedikt Frey and Pedro Llanos-Paredes. 2025. Lost in translation: Ai's impact on translators and foreign language skills. Accessed: 2025-05-20.

Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is better than one: Improving multilingual collaboration by giving two machine translation outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 852–863, New York, NY, USA. Association for Computing Machinery.

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Team Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, and 1330 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare Voss, Marine Carpuat, and Hal Daumé III. 2023. What else do I need to know? the effect of background information on users' reliance on QA systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3330, Singapore. Association for Computational Linguistics.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2015. Natural Language Translation at the Intersection of AI and HCI. *Communications of the ACM*, 58(9):46–53.

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

Ana Guerberof-Arenas and Joss Moorkens. 2023. Ethics and machine translation: The end user perspective. In *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 113–133. Springer.

Ana Guerberof-Arenas, Joss Moorkens, and Sharon O'Brien. 2019. What is the impact of raw mt on japanese users of word: preliminary results of a usability study using eye-tracking. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 67–77.

Ana Guerberof-Arenas and Antonio Toral. 2024. To be or not to be: A translation reception study of a literary text translated into dutch and catalan using machine translation. *Target*, 36(2):215–244.

Janiça Hackenbuchner and Ralph Krüger. 2023. DataLitMT – teaching data literacy in the context of machine translation literacy. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 285–293, Tampere, Finland. European Association for Machine Translation.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *Preprint*, arXiv:2503.04761.

Amanul Haque, Vaibhav Garg, Hui Guo, and Munindar Singh. 2022. Pixie: Preference in implicit and explicit comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 106–112, Dublin, Ireland. Association for Computational Linguistics.

Hendrik Heuer and Daniel Buschek. 2021. Methods for the design and evaluation of hci+ nlp systems. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–33.

Tae Hyun Baek and Minseong Kim. 2023. Is chatgpt scary good? how user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics*, 83:102030.

Elaine C Khoong and Jorge A Rodriguez. 2022. A research agenda for using machine translation in clinical medicine. *Journal of General Internal Medicine*, 37(5):1275–1277.

Elaine C Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the use of google translate for spanish and chinese translations of emergency department discharge instructions. *JAMA internal medicine*, 179(4):580–582.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024a. " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024b. Evallm: Interactive evaluation of large language model prompts on user-defined

criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–21. ACM.

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25(2):131–148.

Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022. Human-centric research for nlp: Towards a definition and guiding questions. In *Proceedings of the Second Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, Seattle, Washington.

Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

May Li. 2023. Post-editing of machine translation. In *Routledge Encyclopedia of Translation Technology*, pages 582–600. Routledge.

Q. Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *Preprint*, arXiv:2306.01941.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Daniel Liebling, Katherine Heller, Samantha Robertson, and Wesley Deng. 2022. Opportunities for human-centered evaluation of machine translation systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 229–240, Seattle, United States. Association for Computational Linguistics.

Daniel J Liebling, Katherine Heller, Margaret Mitchell, Mark Díaz, Michal Lahav, Niloufar Salehi, Samantha Robertson, Samy Bengio, Timnit Gebru, and Wesley Deng. 2021. Three Directions for the Design of Human-Centered Machine Translation.

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.

Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber-Genzel, and Barbara Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Singapore. Association for Computational Linguistics.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352.

Marianna Martindale, Kevin Duh, and Marine Carpuat. 2021. Machine translation believability. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 88–95, Online. Association for Computational Linguistics.

Brian McCarthy. 2004. Does online machine translation spell the end of take-home translation assignments. *CALL-EJ Online*, 6(1):6–1.

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.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Mai Miyabe and Takashi Yoshino. 2010. Influence of detecting inaccurate messages in real-time remote text-based communication via machine translation. In *Proceedings of the 3rd International Conference on Intercultural Collaboration*, ICIC '10, page 59–68, New York, NY, USA. Association for Computing Machinery.

Joss Moorkens and Ana Guerberof Arenas. 2024. Artificial intelligence, automation and the language industry. *Handbook of the Language Industry: Contexts, Resources and Profiles*, 20:71.

Joss Moorkens and Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. In *Human issues in translation technology*, pages 127–148. Routledge.

Mary Nurminen and Niko Papula. 2018. Gist MT users: A snapshot of the use and users of one online MT tool. In *Proceedings of the 21st Annual Conference*

of the European Association for Machine Translation, pages 219–228, Alicante, Spain.

Sharon O'Brien. 2012. Translation as human–computer interaction. *Translation spaces*, 1(1):101–122.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2025. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *arXiv preprint arXiv:2507.19634*.

Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3205–3227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sumant Patil and Patrick Davies. 2014. Use of google translate in medical communication: evaluation of accuracy. *BMJ*, 349.

Jeff Pitman. 2021. Google translate: One billion installs, one billion stories. `https://blog.google/products/translate/new-features-make/translate-more-accessible-for-its-1-bill/ion-users/`. Engineering Manager, Google Translate.

Mondheera Pituxcoosuvarn, Yohei Murakami, Donghui Lin, and Toru Ishida. 2020. Effect of cultural misunderstanding warning in mt-mediated communication. In *Collaboration Technologies and Social Computing: 26th International Conference, CollabTech 2020, Tartu, Estonia, September 8–11, 2020, Proceedings*, page 112–127, Berlin, Heidelberg. Springer-Verlag.

José Pombal, Sweta Agrawal, Patrick Fernandes, Emmanouil Zaranis, and André FT Martins. 2025. A context-aware framework for translation-mediated conversations. *Preprint*, arXiv:2412.04205.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Ming Qian and Chuiqing Kong. 2024. Enabling human-centered machine translation using concept-based large language model prompting and translation memory. In *Artificial Intelligence in HCI: 5th International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part III*, page 118–134, Berlin, Heidelberg. Springer-Verlag.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

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

Anna Rogers. 2019. How the transformers broke nlp leaderboards. Online. Accessed: 2020-05-20.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do llms adapt to different age and education levels? *arXiv preprint arXiv:2312.02065*.

Belén Saldías Fuentes, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89, Dublin, Ireland. Association for Computational Linguistics.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6).

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Alina Secara, Isabel Rivas Ginel, Antonio Toral, Ana Guerberof, Dragos Ioan Ciobanu, Justus Brockmann, Claudia Plieseis, Raluca-Maria Chereji, and Caroline Rossi. 2025. Lt-lider language technology map-technologies in translation practice and their impact on the skills needed. *report*.

Tomohiro Shigenobu. 2007. Evaluation and usability of back translation for intercultural communication. In *Proceedings of the 2nd International Conference on Usability and Internationalization*, UI-HCII'07, page 259–265, Berlin, Heidelberg. Springer-Verlag.

JongHo Shin, Panayiotis G Georgiou, and Shrikanth Narayanan. 2013. Enabling effective design of multimodal interfaces for speech-to-speech translation system: An empirical study of longitudinal user behaviors over time and user strategies for coping with errors. *Computer Speech & Language*, 27(2):554–571.

Marita Skjuve, Petter Bae Brandtzaeg, and Asbjørn Følstad. 2024. Why do people use chatgpt? exploring user motivations for generative conversational ai. *First Monday*, 29(1).

Harold Somers. 2005. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133.

Ana Stojanov, Qian Liu, and Joyce Hwee Ling Koh. 2024. University students' self-reported reliance on chatgpt for learning: A latent profile analysis. *Computers and Education: Artificial Intelligence*, 6:100243.

Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A. Metoyer, and Toby Jia-Jun Li. 2024. Comparing criteria development across domain experts, lay users, and models in large language model evaluation. *Preprint*, arXiv:2410.02054.

Breena R Taira, Vanessa Kreger, Aristides Orue, and Lisa C Diamond. 2021. A pragmatic assessment of google translate for emergency department instructions. *Journal of General Internal Medicine*, 36(11):3361–3365.

Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024. ChatGPT role-play dataset: Analysis of user motives and model naturalness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3133–3145, Torino, Italia. ELRA and ICCL.

The Guardian. 2017. Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Anne M. Turner, Megumu K. Brownstein, Kate Cole, Hilary Karasz, and Katrin Kirchhoff. 2015. Modeling Workflow to Design Machine Translation Applications for Public Health practice. *Journal of Biomedical Informatics*, 53:136–146.

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

Susana Valdez and Ana Guerberof-Arenas. 2025. "google translate is our best friend here" a vignette-based interview study on machine translation use for health communication. *Translation Spaces*.

Mihaela Vela, Santanu Pal, Marcos Zampieri, Sudip Naskar, and Josef van Genabith. 2019. Improving CAT tools in the translation workflow: New approaches and evaluation. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 8–15, Dublin, Ireland. European Association for Machine Translation.

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

Lucas Nunes Vieira and Lucia Specia. 2011. A review of translation tools from a post-editing perspective. In *Proceedings of the Third Joint EM+/CNGL Workshop Bringing MT to the Users: Research Meets Translators (JEC'11): Luxembourg, 14 October 2011*, pages 33–42.

Elizabeth Wagner. 1983. Rapid post-editing of systran. In *Proceedings of Translating and the Computer 5: Tools for the trade*, London, UK. Aslib.

Angelina Wang, Jamie Morgenstern, and John Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

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

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural

machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Jin Yang and Elke D Lange. 1998. Systran on altavista a user study on real-time machine translation on the internet. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, pages 275–285. Springer.

Yanxia Yang. 2024. Understanding machine translation fit for language learning: The mediating effect of machine translation literacy. *Education and Information Technologies*, pages 1–18.

Yanxia Yang, Xiangling Wang, and Qingqing Yuan. 2021. Measuring the usability of machine translation in the classroom context. *Translation and Interpreting Studies*, 16(1):101–123.

Emmanouil Zaranis, Nuno M Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA. Association for Computational Linguistics.

Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):1–37.

Yongle Zhang, Dennis Asamoah Owusu, Marine Carpuat, and Ge Gao. 2022. Facilitating global team meetings between language-based subgroups: When and how can machine translation help? In *Proceedings of the ACM on Human-Computer Interaction 6, CSCW1*, pages 1–26, New York, NY, USA. Association for Computing Machinery.

Yongle Zhang, Dennis Asamoah Owusu, Emily Gong, Shaan Chopra, Marine Carpuat, and Ge Gao. 2021. Leveraging machine translation to support distributed teamwork between language-based subgroups: The effects of automated keyword tagging. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Lingjun Zhao, Khanh Xuan Nguyen, and Hal Daumé Iii. 2024. Successfully guiding humans with imperfect instructions by highlighting potential errors and suggesting corrections. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 719–736, Miami, Florida, USA. Association for Computational Linguistics.

Wujie Zheng, Wenyu Wang, Dian Liu, Changrong Zhang, Qinsong Zeng, Yuetang Deng, Wei Yang, Pinjia He, and Tao Xie. 2019. Testing untestable neural machine translation: An industrial case. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 314–315.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. 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.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A ACL Anthology Query

To identify research trends in the ACL community (Figure 1), we searched for specific keywords in either the title or abstract of research articles published from `2015-01-01` to `2024-12-31` and hosted in the ACL anthology repository.[19] Specifically, we use the following keywords in a case-insensitive fashion and including all grammatical numbers by means of regular expressions:

- **machine translation** (MT): *translation*, *machine translation*, *nmt*, and *mt*;
- **language models** (LM): *llm*, *language model*, *large language model*, and *foundation model*;
- **users** (U): *user*.

To reduce noise, we exclude editorials (i.e. those with a `proceedings` bibtex type) and rare instances of articles without any author from the matching documents. We obtain a total of 62,032 articles, of which 8,072 match MT keywords, 13,977 match LM keywords, and 5,084 match U keywords.

In Figure 2, we present the trends of interest over the last ten years in terms of absolute counts.

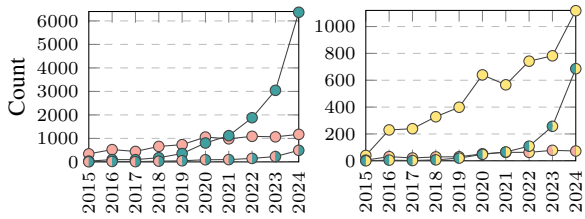

Figure 2: Trend of interest (absolute counts) in *machine translation* **MT**, *language models* **LM**, *users* **U**, and combinations thereof in the ACL community over the last 10 years. Besides illustrating the rapid growth of LLM studies, the left panel highlights the increase in MT research incorporating LLMs (**MT** + **LM**), while the right panel shows rising attention to users, particularly in LLM-related work (**LM** + **U**).

B Professional and Lay Users of MT

Table 1 offers a preliminary outline of some key differences between professional (i.e. translators or MT post-editors) and lay users of machine translation. While we acknowledge that these characteristics often exist along a continuum rather than as clear-cut categories, here we draw on prototypical positions inspired by Cifter and Dong (2009) to highlight contrasting tendencies. This distinction is useful for framing how different user profiles interact with MT systems, particularly in terms of expectations, urgency, error tolerance, and the ability to critically assess output.

A more fine-grained subclassification of user types (e.g. according to varying degrees of language proficiency or MT literacy) is currently hindered by the limited literature on these aspects. However, concurrent work by Bassignana et al. (2025) highlights digital literacy as a key potential divide in the use of language technologies. Along these lines, post-editors can develop higher digital literacy and skills through sustained engagement with language technologies as part of their professional growth (Secara et al., 2025), in contrast to lay users, whose levels of digital literacy or access can vary greatly.

[19] `https://aclanthology.org` (accessed: `2025-02-01`).

Aspect	**Professional Users**	**Lay Users**
Training	Specialized	None or limited training
Usage Context	Professional tasks	Personal or immediate needs
Terminology	Familiar with domain-specific terminology	Can be unfamiliar with specialized terminology translated
Awareness of Limitations	Familiarity with MT capabilities and limitations	Limited/None
Language Proficiency	Proficient in both source and target languages	Limited or no proficiency in at least one of the languages
Error Evaluation	Can effectively judge translation quality and identify errors	May struggle to spot errors

Table 1: Overview of Professional vs. Lay Users of Machine Translation.