Communicating with the Entire Population: Machine Translation of Emergency Alerts and Warnings

Jen Wilson*, **Vipasha Bansal**, **William D. Lewis**
Computational Linguistics Program
University of Washington

Abstract

It is essential that we communicate with the entirety of an affected population during a crisis, including those who do not speak a region's dominant language(s). Commercial Translation tools offer a tempting solution to broaden access and can provide a lifeline when no other resources are available. However, Emergency response managers often operate in the dark regarding the quality and adequacy of these automated outputs. In this paper, we conduct a nationwide survey of emergency managers to understand their current workflows, perspectives, and operational challenges. We then evaluate commercial translation tools against these needs, shedding light on their utility and offering practical guidance on how and when these tools can be safely used in emergency response.

1 Introduction

Communication during emergencies is a core part of public safety. Access to timely information during a crisis is as essential as access to "food, water, shelter, and medication" (Meier, 2015). Emergency alerts and warnings systems, such as Wireless Emergency Alerts (WEA)[1]—must reach the *entirety* of an affected population, rather than an English-fluent subset. Failing to achieve equitable coverage puts lives directly at risk.

This outreach is difficult in practice. Local emergency managers, public information managers, and communication staff often serve multilingual communities while working with limited time, staffing and budgets, sometimes across dozens or even hundreds of languages (Ciribuco et al., 2025). In regions where alerts are distributed only in the dominant language(s), non-fluent residents are inevitably excluded, which can have devastating consequences in time-sensitive crises (Villarreal et al., 2025; Pabón et al., 2025; O'Brien and Federici, 2019).

Because human translation is often cost-prohibitive or too slow for local jurisdictions, practitioners increasingly turn to automated solutions. Commercial Machine Translation (MT) systems, such as Google Translate[2] and Microsoft Translator[3], are widely available and support a large number of languages. Large Language Models (LLMs) such as ChatGPT[4], Claude[5], and Google Gemini[6] are also increasingly used for translation.

These tools may offer speed and convenience, but emergency messaging demands accuracy and usability. Inaccurate translations risk delivering alerts that are uninterpretable or dangerously misleading. Crucially, strong MT performance in one domain or language pair does not guarantee cross-lingual or cross-domain reliability (Bogoychev and Chen, 2021; Zouhar et al., 2024), requiring comprehensive multi-target evaluation to ensure safety.

Although this paper focuses on emergency information translation in the United States, the challenges we describe are not unique to any one country and are likely familiar to officials in many multilingual settings. As claims regarding Artificial Intelligence (AI)[7] expand across crisis management, practitioners face a rapidly evolving landscape of software tools and vendor claims (Foubert et al., 2026; Whyte and Haupt, 2026). In this paper, we seek to shed light on automated translation in emergency response from two angles.[8] First, we survey emergency managers across the United States to

*Corresponding author: jenwils@uw.edu

[1]https://www.fema.gov/emergency-managers/practitioners/integrated-public-alert-warning-system/public/wireless-emergency-alerts

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

[3]https://www.microsoft.com/en-us/translator/

[4]https://chatgpt.com/

[5]https://claude.com/

[6]https://gemini.google.com/app

[7]Machine Translation being one of many forms of AI.

[8]In this work, we chose to focus on the quality and utility of commercially available tools that do translation. We leave the analysis of MT workflows, logistics of their implementation and use in the field, etc., to others.

Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response, pages 46–64
November 9, 2026 ©2026 Association for Computational Linguistics

understand current translation practices, needs, and challenges.

Second, we examine the translation tools available to practitioners and review existing research on their evaluation and use in crisis settings. By bringing together the perspectives of emergency management practitioners and research on translation technologies, we hope to help bridge the gap between the development and evaluation of these tools and the contexts in which they are used.

2 Wireless Emergency Alerts

In the United States, the Wireless Emergency Alerts (WEA)[1] system broadcasts alerts to mobile and other WEA-enabled devices, providing actionable information in times of crisis. However, native support is restricted to English, leaving non-English speakers at a disadvantage unless they proactively sign up for translations.

In 2023, the US Federal Communications Commission (FCC, 2023) recognized that "it is essential that the public be able to receive in accessible language and format WEA Messages that are intended for them...to help the millions of people with access and functional needs, including people who primarily speak a language other than English or Spanish...better understand and take protective actions." Despite this, the ability of agencies to meet multilingual and accessible alerting needs varies widely depending on local demographics, staffing, funding, and available translation infrastructure.

3 Establishing the Need: A Survey of Emergency Managers

The role of language technology in emergency management is underexplored and deserves attention. Foubert et al. (2026) surveyed emergency managers to identify general perceptions and potential use cases for emerging AI. The specific adoption and efficacy of automated translation tools in crisis contexts require dedicated study.

To examine current translation practices, needs, and operational hurdles, we surveyed members of the Language & Accessibility for Alert & Warning Workgroup (LAAWW)[9] in December 2025. LAAWW members represent 42 U.S. states, territories, and Tribal Nations who are "working toward greater language and disability access in emergency alert and warning systems."

We used purposive sampling strategy to reach practitioners actively involved in language access.[10] Therefore, this survey is not intended to be statistically representative of all U.S. emergency management agencies; rather, it provides insight into the tools, decision-making processes, and challenges involved in producing multilingual alerts. Survey details are in Appendix F.

The survey asked about respondents' roles and responsibilities, languages spoken in their communities, translation tools and practices, and challenges encountered in multilingual communication. We received 25 responses, approximately 25% of meeting attendees.[11]

More than half of respondents identified themselves as a *Local Emergency Manager* or *Public Information Officer*, while 36% identified as an *Operations, Logistics, or Communications Manager* or *Language Coordinator or Interpreter*.[12]

Participants noted that while many jurisdictions have residents who speak a variety of languages, only a subset of these languages are consistently supported for time-critical alerts. Agencies must prioritize languages while balancing accuracy, usability, speed, and cost. Translations should be usable and accurate, while also affordably and quickly produced.

Although participants prioritized quality translations, for those who shared their challenges, 100% cited money, time, and/or limited resources.

3.1 Translation Needs Across LAAWW Membership

Translation needs vary substantially across jurisdictions represented in the survey. Respondents reported serving communities where between two and more than 200 languages are spoken. English was the primary language across all reported jurisdictions and Spanish was the most consistently supported additional language. However, the remaining language needs varied considerably by region. A partial list of languages that survey participants work with is available in Appendix D.

The scale of these differences creates distinct operational requirements. Smaller jurisdictions may

[9] https://sites.google.com/view/laaww

[10] This approach surfaces qualitative operational insights; future work with randomized sampling would be required for broader statistical generalization.

[11] Attendance at LAAWW meetings is approximately 100 people per month, representing approximately 80 unique entities (agencies, vendors, cities, counties, etc).

[12] The survey allowed respondents to select more than one role.

be able to reach most residents by translating alerts into only a few languages, while larger and more linguistically diverse jurisdictions may need to support many more. For example, one respondent from Maryland supports a community of approximately 5,500 people and three commonly spoken languages. Another works in New York City, where more than 200 languages are spoken across a population of more than 8.5 million. While we can look at average language populations, nearly every area has populations that deviate from that national average. These jurisdictions face fundamentally different translation-planning challenges.[13] Survey respondents support over 36 million people across the United States[14].

Agencies must identify the languages most important for their communities and develop processes for producing translations that are accurate, usable, affordable, and fast enough for time-critical alerts. The central need is not simply translation into more languages, but scalable and locally informed translation capacity. For more details on translation needs and speaker populations, as reported by the LAAWW membership, please see Appendix C.

3.2 Translation Tools and Practices

The majority of respondents (55%) relied exclusively on human translation and did not use automated tools. Another 27% used a mix of human translators and automated tools, while 18% relied exclusively on automated tools (Figure 1). For those using human translation, work was performed by organization staff (45%), community members or volunteers (26%), and contractors or commercial vendors (29%). These results show that jurisdictions use a range of staffing models to produce multilingual alerts.

Among respondents using translation-specific commercial tools, Google Translate[2] was the choice in 90% of cases, followed by DeepL Translator[15] at 20%.[16] A minority (and we suspect growing) number of respondents use LLM–based

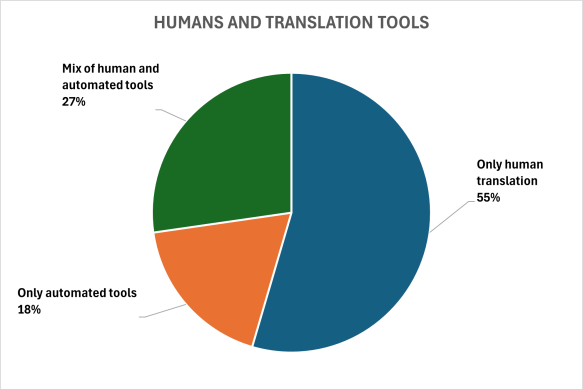


Figure 1: Human Translation and Translation Tools

tools, with ChatGPT[4] used in 30% of cases and Google Gemini[6] used in 10% of cases.[16] Given the small sample size, reported use of LLM-based tools likely reflects exploratory or supplemental use rather than standardized, organization-wide deployment. These results are consistent with broader interest in AI among emergency managers. Foubert et al. (2026) found that 75% of their respondents were interested in using AI for public and stakeholder messaging, including multilingual translations. Together, these findings suggest that automated tools are already part of some emergency communication workflows, while their organizational adoption remains uneven.

3.3 Factors Influencing Tool Selection

Respondents selected translation tools primarily on the basis of practical considerations rather than formal requirements. Availability or ease of access was a criteria for 56% of respondents, and perceived accuracy or quality was important for 52%. Cost or budget constraints influenced 40% of respondents' choices, while ease of use or staff comfort influenced 32%. These results suggest that agencies favor tools that are accessible, affordable, and sufficiently reliable for everyday workflows.

Compliance requirements (16%) and "other" factors (12%) were cited less often, indicating that regulatory or external mandates play a smaller role in shaping tool choices.

Tool adoption therefore appears to be driven more by immediacy and feasibility than by formal risk assessment or regulatory guidance (Figure 2). This pattern is consequential because a tool that is convenient to access may still require systematic evaluation before it is used for high-stakes messages

[13] Appendix D displays this data as percentage off expected.

[14] The population information reported here reflects the total population of respondents' jurisdictions and do not imply that all residents require translated alerts or that language services are uniformly provided across all languages in each jurisdiction. Rather, they illustrate the scale and diversity of contexts in which respondents operate.

[15] https://www.deepl.com/en/translator

[16] Respondents could specify more than one tool and the percentages do not necessarily add to 100%.

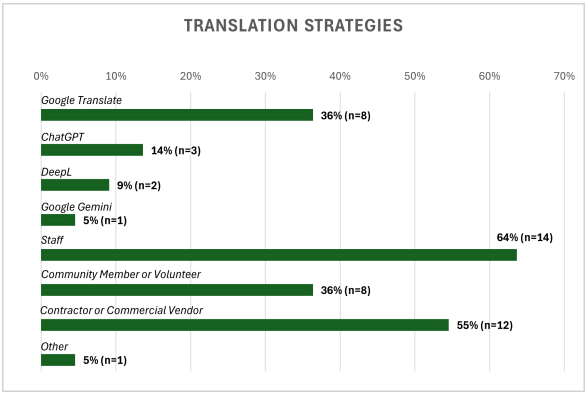


Figure 2: Strategies of respondents to obtain translations.

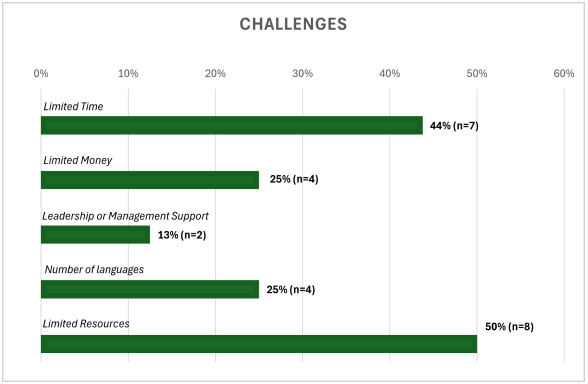


Figure 3: Operational challenges faced by survey respondents.

3.4 Evaluation and Review Practices

Formal evaluation of translation tool output was uncommon. Only 10% of respondents reported having a formal process for assessing accuracy, usability, or effectiveness. Sixty percent used an informal evaluation process, while 30% reported having no evaluation process.

Processes for reviewing or updating tools were similarly inconsistent. Forty percent of respondents reported reviewing or updating tools as needed, 25% reported having no formal review process and another 25% were unsure whether such a process existed. The remaining 10% reported that their tools were not reviewed at all. Thus, many jurisdictions may be using translation tools in high-stakes contexts without systematic procedures for evaluating or monitoring their performance, creating a risk of uneven message quality across languages.

3.5 Operational Challenges

Respondents described multiple barriers to producing multilingual alerts, including limited time, funding, staff, and management support.

Translator time and availability were also recurring concerns. Participant L1 emphasized the need to "minimiz[e] time delays in alerts and warnings...especially with translations into various languages." Participant P1 similarly described the "challenge [with] human translation is always time and availability" of translators.

Participant H1 reported that their state maintained a registry of approved interpreters, but very few were available in their rural county. As linguistic diversity increases, agencies may need to support additional languages even when human translators and external vendors are slow or unavailable.

Respondents also reported that the usefulness of translation tools varied by language. Participant R1 explained "multiple languages in the state are very unique thus singular translation services for only a few languages [have] proven ineffective in [the] past." This aligns with our CHRF++ findings (Section 5), which showed significant variation across languages and systems

Beyond operational barriers, participants also noted that language access does not guarantee message comprehension, particularly for populations with limited literacy or differing communication preferences. These constraints show a deeper challenge: language access alone is insufficient. The findings point to the need for complementary accessibility strategies, such as plain-language messaging, audio or pictorial alerts, and community-based dissemination, especially in high-risk populations (Villarreal et al., 2025).

4 Differences between MT and LLM use by Emergency Managers

Our survey demonstrated that respondents in emergency management and communications use both MT and LLM-based tools. There are many differences between the two technologies, but for the purposes of this paper, we will discuss the expected outputs and the types of data used in the training of Google Translate and OpenAI's ChatGPT.[17]

MT is traditionally trained on a corpus of aligned sentence pairs (Bapna et al., 2022), where the sentences corresponding to one another in each language have the same meaning. The MT models learn how text in one language corresponds

[17] Discussion of other similarities and differences, such as model architecture, training objectives, training process, inference mechanisms, etc. is beyond the scope of this paper.

to equivalent text in another language. Commercial tools, such as Google Translate and Microsoft Translator, are trained in this manner and publish lists of languages that are supported by their services.[18]

While LLMs such as ChatGPT are pre-trained on publicly available data (websites, news, etc.), multimodal data, non-English data, and data licensed for training (OpenAI, 2024), the specific data sources are rarely disclosed. The data may also include aligned sentence pairs and text in multiple languages (Chua et al., 2025). However, unlike the MT services discussed in the previous paragraph, LLMs generally do not list what languages they support for MT. In fact, MT may arise as an emergent property rather than as a training goal (Wei et al., 2022). LLMs are trained to generate plausible output, and will provide output even when lacking sufficient data to provide one that is correct. This can lead "models to convert uncertainty into content that is often plausible but wrong" (Kalai et al., 2026).

5 Quality of Commercial MT Tools in Crisis Response

The survey of LAAWW membership shows a clear interest in automated tools for translation. Several studies have evaluated the use and performance of MT in crisis contexts, such as for emergencies (Turner et al., 2019), multilingual crisis-related communication (Rossetti et al., 2020), or pandemic-related communication across multiple languages (Anastasopoulos et al., 2020). More recent work has evaluated both MT systems and LLMs on emergency-response texts and multilingual crisis data, finding limitations in areas including domain-specific terminology, contextual ambiguity, register, and the preservation of urgency (Staiano et al., 2025; Ticona and Anastasopoulos, 2026; Court et al., 2026). Less attention has been given to systematically comparing currently available commercial translation tools on emergency alert content across multiple languages. This distinction is important for emergency managers, who are more likely to encounter translation tools as commercially available services than as models requiring specialized technical expertise to use or adapt.

[18]For a list of languages supported by Google Translate, see: https://docs.cloud.google.com/translate/docs/languages. For Microsoft Translator, see: https://www.microsoft.com/en-us/translator/languages/.

5.1 Quantitative Evaluation

Bansal et al. (2026) evaluated seven commercially available MT systems on the TICO-19 pandemic-related benchmark dataset (Anastasopoulos et al., 2020), which contains English-language COVID-19 material translated into 36 languages, and assessed translation usability based on whether the output was sufficiently understandable and adequately conveyed the relevant information. Their results demonstrate substantial variation in translation performance across languages and systems, particularly for lower-resource languages.

Pang et al. (2025) find that LLMs trained across multiple domains perform consistently across a range of tasks but can fall short of domain-specific models. To investigate whether this pattern extends to crisis communication, we compare scores on the full TICO-19 dataset with scores on its News and Announcements (N+A) subset, whose short, information-rich texts more closely resemble Wireless Emergency Alerts (WEAs). Because WEAs are issued in English, we focus on English-to-target-language (EX) translation scores.

LAAWW members identified 24 supported languages that overlap with the languages represented in TICO-19. This overlap enables a unique domain-specific evaluation of emergency-relevant text for those languages, allowing a comparison with general-purpose content. Following Anastasopoulos et al.'s (2020) categorization, we group the languages into three tiers: *Pivot* (high-resource), *Important* (mid-resource), and *Priority* (low-resource).

We examine CHRF++ scores for both the overall and N+A subset of TICO-19. CHRF++ is a quality metric widely used in the MT field for measuring quality (Popović, 2017). The differences between these scores for some language-system combinations, suggests that performance on a broader evaluation set does not necessarily predict performance on text with characteristics more similar to emergency alerts.

Pivot languages generally achieved the strongest Overall scores, while Important and Priority languages showed lower and more variable performance. No system was consistently strongest across languages, and the best-performing system sometimes changed between the Overall and N+A evaluations. These scores suggest that overall benchmark scores are insufficient for predicting performance on emergency-relevant text and that

evaluation should account for both language and domain.[19]

For example, Microsoft's Hindi N+A CHRF++ score was 13.16 points *lower* than its Overall score, Claude was 12.24 points lower, and GPT-5.4-mini was 10.33 points lower. In fact, all of the Pivot languages showed lower scores for the N+A subset, except for Swahili, for which all services were between 1.54 and 3.13 *higher* (Appendix G, Table 3). For most language–system pairs, CHRF++ scores were higher on Overall than on the N+A subset, with particularly large differences for Hindi and Arabic. However, this pattern was not universal: Bengali, Khmer, Kinyarwanda, Swahili, Somali, Tamil, and Zulu had lower Overall than N+A scores for at least one system.

5.2 Qualitative Evaluation

Qualitative analyses by Hajek et al. (2024) identify errors that may sound fluent, but substantially alter the intended meaning of emergency instructions. For example, the English instruction "Try to position the car *towards* the approaching fire" was translated to Chinese and then back into English as "Try to park your car *near* the fire," reversing the intended instruction and potentially creating a very serious safety risk.[20]

Hajek et al. (2024) also describe problems with lexical ambiguity. The instruction "shelter in a room" was translated into Spanish as "refugio en una habitación," which when backtranslated, means "a refuge in a room." The word *shelter* in this case was translated as a noun (*refugio*), when actually it was a verb (a command to take shelter). It may be tempting to assume that *something* is better than *nothing* when it comes to emergency messaging. However, these examples demonstrate that that is not always the case.

These examples illustrate a type of MT failure that is dangerous in emergency contexts: mistranslations that are fluent, possibly plausible-sounding are therefore unlikely to be flagged by non-expert reviewers, but convey meaningfully different instructions (Šorak, 2024).

Additionally, Ticona and Anastasopoulos (2026) demonstrate that LLMs inconsistently translate urgency, such as "a gas leak scenario classified as Critical in English is downgraded to Low in Marathi and Not Urgent in Amharic."

Qualitative evaluations provide examples that show where MT has gone wrong and can be useful in providing actionable content that can help guide how translations are done, used, or evaluated. However, they do not give an overall picture of MT output but only specific instances identified by human translators and may be, or may not be, representative of broader issues. This limitation underscores the need for domain-specific and systematic evaluation frameworks for MT in emergency settings.

6 What to Do and What Not to Do with Commercial MT Tools

The gold standard for translations of emergency alerts and messaging is to translate all anticipated messages by a human before they are needed (Federici, 2022; Villarreal et al., 2025). When an organization has limited resources, or an unanticipated crisis occurs, or where professional translators are not available, this process may not be possible. Commercial machine translation can bridge the gap and get information out to people quickly, while buying time to have a more thorough translation prepared.

Based on work by Ticona and Anastasopoulos (2026) and Bansal et al. (2026), we can see that there are gaps in the quality of commercial and open-source MT tools in the crisis domain cross-lingually, and potentially undesirable errors in output, such as those noted by Hajek et al. (2024). Knowing the quality gaps can help emergency managers integrate translation tools into their workflows appropriately. It is important to know how one can improve the quality of MT and how much a human should be in the loop. In this section, we will look at ideas and rules of thumb that emergency managers can adopt for using such tools.

6.1 The Epic Highs and Lows of Raw Machine Translation

As MT quality has improved, especially for high resource languages, the use of raw MT has become more prevalent, especially in contexts where human translators are not available (Nurminen, 2021; Nurminen and Papula, 2018; Kasperė et al., 2021; Vieira et al., 2022; Kenny et al., 2022)—a constraint reflecting the standard operating environment of many LAAWW practitioners.

[19] See Appendix G, Table 4 for full scores.

[20] They did not note which system produced this translation. Hajek et al. (2024) gathered their translations in 2023. Upon subsequently examining Google Translate (captured on 14 Jan, 2026), the Chinese translation is 着将汽向逼近的火焰。

However, there is a significant literature on the use of MT in crises, with the general recommendation is that humans should be in the loop to prevent severe errors with potentially catastrophic consequences[21] (Cadwell et al., 2019; O'Brien and Federici, 2019; Federici and Cadwell, 2018; Federici, 2022). Reflecting these risks, many Language Access Policies discourage the use of raw MT, *i.e.,* using the direct output of MT engines without any post-editing or human-in-the-loop (City of LA, 2022; U.S. GSA, 2026; US DOJ LEP Cmte, 2021).

Strictly avoiding raw MT may not be feasible when contending with budget limits, as highlighted by our survey respondents. But there are things that can be done to help mitigate some of these dangers, either through simplifying text, preparing messages in advance, use of templates, or doing as much as possible in advance of crises.

6.2 Translated Templates

The Federal Communication Commission (FCC) provides templates for WEAs, translated into the 13 most commonly spoken languages in the United States.[22] By using templates like these, one can sidestep the use of MT altogether in certain crisis scenarios.

We recommend creating or repuposing templates in advance of crises that may occur regularly. Use the following steps to create templates that can be repurposed:

1. **Pre-approval:** Draft templates, using short, plain-language messages for hazards or actions.

2. **Placeholder management:** Use clearly labeled and unambiguous placeholders in the templates for event specific details (*e.g.,* location and place names, dates, times, etc.). Ensure the placeholders are untranslated, and have the correct placement for the target languages.

3. **Rapid deployment:** When an incident occurs and a message template has already been prepared, it can be filled in and translated to additional languages through automated tools.

6.3 Text Simplification

Text simplification has long been used to improve MT output, and is discussed in the literature under various guises: text simplification (Štajner and Popovic, 2016), controlled language (Aikawa et al., 2007), and pre-editing (Sánchez-Gijón, 2022), among others. In the distributed messages setting, such as with WEAs, all messages should be as clear and direct as possible to increase the chance of information being translated accurately.

Štajner and Popovic (2016) find that text simplification can improve the fluency of the MT output. While their work focuses on automated simplification, the same principle can be extended to include manual simplification techniques, *i.e.,* where the authors of the messages simplify the text themselves. Hajek et al. (2024), for example, recommend avoiding culture-specific terms, avoiding two-part verbs (*e.g.,* look up, get along with, etc.), and avoiding elements with multiple grammatical functions. Recommendations to simplify crisis communications before translation have also been made by local governments (King County Executive, 2010), policy reports (Fordham, 2023), and government reports (Hajek et al., 2024).

LLMs can also be used in text simplification (Guidroz et al., 2025). For example, the following text could be used as a prompt to an LLM to simplify text supplied by the user:

Rewrite the following text in plain language at approximately a 6th-grade reading level. Use short sentences and everyday vocabulary. Keep all safety instructions, locations, and times accurate.

The text simplification process should be treated as a separate step from translation. Evaluating the simplified text before translation can help confirm that the source message remains accurate, retains all essential information, and is adequately simplified and ready for the next stage. Separating these steps makes it possible to identify and correct problems with the source message before it is disseminated to the public.

6.4 Using LLMs for Translations

The majority of LAAWW members use human translation (Figure 1), but respondents also indicated interest in translating with LLMs. The use of general-purpose LLMs, such as GPT and Gemini, in high-stakes communication contexts introduces risks that go beyond translation quality as measured by evaluation metrics. LLMs are known to produce fluent but semantically incorrect outputs — a phenomenon called "hallucination" — at rates that are difficult to predict and vary by language, domain,

[21] We discuss notable examples in Section 5.2.

[22] https://www.fcc.gov/multilingual-wireless-emergency-alerts hosts example templates.

and prompt formulation (Ji et al., 2023). In crisis communication, this unpredictability is a serious concern.

To reduce the risk of hallucinations, we recommend using a prompt such as the one below, which instructs the LLM to leave text untranslated when it is uncertain. Although this prompt may lower the likelihood of hallucinations, it is not guaranteed. For languages not well supported by the model, LLM-based translation should be avoided. See Appendix G for additional guidance.

Translate this message into [TARGET LANGUAGE]. Be literal. Do not add, omit or invent information. If a phrase is ambiguous, mark it as [UNCERTAIN] and leave it untranslated. Preserve names, dates, and numbers exactly. After the translation, provide a one-line confidence: High / Medium / Low and list any parts you were unsure about.

The authors acknowledge the mixed messaging this section provides. While we suggest that LLMs not be used for translations, we also give guidance on how to use LLMs for translations. As researchers, we are hesitant to recommend the use of LLMs for the reasons stated above. As people who have some understanding of the community that uses these tools, we know that LLMs may be used for translations regardless of what an academic paper may recommend. To that end, we also recommend strategies to use LLMs as effectively as possible.

The models used in commercial LLMs can change with little notice to end users, making it difficult to track how system behavior evolves over time. Because these system updates may be undocumented or only partially disclosed, a prompt or output that is usable at one point in time may produce different results later, even when the input remains unchanged. For emergency management practitioners who use LLMs for translation, these changes can be especially consequential. Small shifts in wording, terminology, or tone may affect whether a message is interpreted accurately and in a way that supports timely response. This potential inconsistency in output should be considered if choosing to use an LLM instead of a commerial MT tool, such as Google Translate or DeepL.

6.5 Post-Editing and Usage

It is challenging for organizations to determine how to best allocate limited resources. One option for translation, beyond the methods discussed previously (*e.g.,* using templates, simplification, etc.) is to spend resources on post-editing rather than on message creation and translation. In this scenario, human translators rely on MT to generate a translated output, and then subsequently correct any errors in the output. This allows translators to spend their time reviewing and editing translated messages rather than translating messages from scratch. The use of MT in workflows has been shown to reduce costs.

To make messages broadly consumable, in addition to the steps outlined in Sections 6.2 and 6.3, we recommend that the messages are shared on available public and social platforms (Purohit et al., 2025). Some platforms provide built-in translation features, allowing users to translate messages without copying the text into a separate program or application. However, this approach removes the ability to monitor quality of the translations and increases the risk that people will get misleading, incorrect, or harmful information. It will be up to the user to determine whether this will work for their situation.

MT engines can produce errors, some of which can have catastrophic consequences, such as examples from Hajek et al. (2024). Although quality evaluation is certainly possible, it is useful to determine *adequacy* for the specific need versus absolute *quality*. In other words, a translation that has some limited errors or is ungrammatical may still be of sufficient quality and good enough to be used. For suggestions on how to determine the adequacy of translations, see Appendix E.

Ad hoc translations likely occur regularly during crisis situations, given the ubiquity of translation tools and people's familiarity with them. In urgent situations, messages are often issued first in the dominant language(s) and translated for dissemination in other languages later. For those who do not speak the dominant language, urgency will drive them to quickly obtain their own translations, using some of the same commercial translation tools described in this paper. In some cases, a translated message may not be disseminated in a specific language at all, forcing recipients to translate the message themselves. Organizations should therefore consider not only the quality of translations *they* produce, but also the translations that recipients are likely to obtain independently when official translations are unavailable.

7 Conclusion

We recommend having humans in the loop in translating, editing, and evaluation whenever resources permit. In practice, however, jurisdictions that communicate in dozens of languages may not be able to review each message by hand. Automated workflows can expand translation capacity, but should be designed around the risks of the message, language, and tool rather than treated as a universal replacement for human review.

An important point to make is that MT engines will make errors, and not all errors are created equally. The types and severity of errors can vary greatly depending on the content, *i.e.,* the complexity of the original text or the amount of resources for the target language. As we have shown, MT tends to do better with higher-resource languages such as Indonesian or Spanish than for lower-resource languages such as Amharic or Pashto. Performance on the N+A subset did not always follow performance on the broader evaluation. These differences make language- and domain-specific evaluation essential.

In a world where communities are becoming ever more multilingual, translation will remain an important component of emergency and crisis communication. Fortunately, the quality and utility of MT is improving, but its use should be paired with realistic expectations, along with human oversight. In crisis settings, the goal is not simply to produce a translation, but rather to deliver information that affected communities can understand and act on safely.

Limitations

The survey done for this paper captures the languages that US-based agencies *report* that they need to support. It does not systematically compare those responses against actual distribution of non-English speakers in the respondents' jurisdictions as measured by US Census data or available American Community Survey data.

Ethical Considerations

Considerable care was taken in designing and administering the survey to ensure that the questions were appropriate to the study objectives, respectful of participants, and attentive to the needs and perspectives of the communities being studied. The survey was developed after careful consideration of the roles and responsibilities of LAAWW members. The methodologies and goals of the study were presented during a LAAWW meeting and members were able to ask questions. This process helped ensure that the survey questions, framing, and proposed use of the data were clearly understood and acceptable to the participating community.

The study was also submitted for review through the University of Washington (Seattle) Human Subjects Division Institutional Review Board (IRB). The IRB determined that the proposed activity constituted human subjects research and qualified for exempt status. Data collection proceeded in accordance with the requirements and conditions of that determination.

Participation in the survey was designed to protect participants' privacy. Participants were able to respond anonymously and were not required to provide identifying information. Participants were given the option to voluntarily provide contact information and geographic location if they wished to do so. These fields were optional and were not required for participation. The study therefore collected only information that participants chose to provide, with identifying information separated from the anonymous survey responses as appropriate for the purposes of analysis and reporting.

Taken together, these steps were intended to ensure that the survey process was conducted thoughtfully and responsibly, from initial community consultation through IRB review, data collection, privacy protection, and subsequent analysis.

Acknowledgments

This work was partially supported by the National Science Foundation under the award CIRC 2346335. We would like to thank the two LT4CPR workshop reviewers for their very helpful comments and suggestions, all of which helped improve this paper. We also extend our warmest THANK YOU to the members of LAWWW, whose contributions to the survey, including their suggestions, questions, and support, made this paper what it is today. Without LAWWW, this paper could not have been written.

References

Takako Aikawa, Lee Schwartz, Ronit King, Mo Corson-Oliver, and Carmen Lozano. 2007. Impact of Controlled Language on Translation Quality and Post-editing in a Statistical Machine Translation Environ-

ment. In *Proceedings of the Machine Translation Summit XI*.

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.

Vipasha Bansal, William D. Lewis, Nicole Mihaylova, Elizabeth Brown, and Benjamin Pong. 2026. Is MT Ready for the Next Crisis or Pandemic? In *Proceedings of the Language Technologies for Crisis Preparedness and Response (LT4CPR Workshop)*, Hengqin, Zhuhai, China. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, and 5 others. 2022. Building machine translation systems for the next thousand languages. *Preprint*, arXiv:2205.03983.

Nikolay Bogoychev and Pinzhen Chen. 2021. The highs and lows of simple lexical domain adaptation approaches for neural machine translation. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 74–80, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Cadwell, Sharon O'Brien, and Eric DeLuca. 2019. More than tweets: A critical reflection on developing and testing crisis machine translation technology. *Translation Spaces*, 8(2):300–333.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual capabilities and knowledge barriers in multilingual large language models. *Preprint*, arXiv:2406.16135.

Andrea Ciribuco, Federico M. Federici, Lorenzo Guadagno, Brigid Maher, Loredana Polezzi, and Rita Wilson. 2025. *Migrants, multilingual communication and cascading crises*, volume 1, pages 87–102. Routledge, Abingdon, Oxfordshire, UK.

City of LA. 2022. Citywide language access plan fy 22-24. https://communityinvestment.lacity.gov/sites/g/files/wph2001/files/2025-04/Citywide%20Language%20Access%20Plan%20FY%2022-24%2C%20City%20of%20Los%20Angeles%20%281%29.pdf. [Accessed 16 Jul 2026].

Sara Court, Lara Downing, and Micha Elsner. 2026. LLMs in the real world: Evaluating "AI" in emergency contexts. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35748–35760, San Diego, California, United States. Association for Computational Linguistics.

FCC. 2023. Emergency Alert System; Wireless Emergency Alerts. Final Rule.

Federico Federici and Patrick Cadwell. 2018. Training citizen translators: Design and delivery of bespoke training on the fundamentals of translation for New Zealand Red Cross. *Translation Spaces*, 7(1):20–43.

Federico M. Federici. 2022. Translating hazards: multilingual concerns in risk and emergency communication. *The Translator*, 28(4):375–398.

Anni Fordham. 2023. Beyond barriers. https://churchilltrust.my.salesforce-sites.com/api/services/apexrest/v1/image/?Id=069RF00000NxDTtYAN. [Accessed 01 Jun 2026].

Samsara Foubert, Minjung Park, Kanad Pardeshi, Thomas Manzini, Robert McDaniel, David Merrick, Robin Murphy, Aarti Singh, and Hoda Heidari. 2026. Investigating the role of AI in emergency management: Use cases, challenges, and opportunities. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, page 2055–2091, New York, NY, USA. Association for Computing Machinery.

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric (Yifan) Cao, and 7 others. 2025. LLM-based Text Simplification and its Effect on user Comprehension and Cognitive Load. arXiv:2505.01980.

John Hajek, Anthony Pym, Yu Hao, Maria Karidakis, Ambrin Hasnain, Anila Hasnain, Juerong Qiu, Ke Hu, and Rachel Macreadie. 2024. Understanding and improving machine translations for emergency communications. https://works.hcommons.org/records/j3fba-nsn06. [Accessed 01 Jun 2026].

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and 1 others. 2026. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653:1047–1051.

Ramunė Kasperė, Jolita Horbačauskienė, Jurgita Motiejūnienė, Vilmantė Liubinienė, Irena Patašienė, and Martynas Patašius. 2021. Towards Sustainable Use of Machine Translation: Usability and Perceived Quality from the End-User Perspective. *Sustainability*, 13(23):13430.

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

King County Executive. 2010. Written language translation process king county written language translation process manual guidelines, effective practices, maps, resources. https://kingcounty.gov/exec/styleguide/~/media/cbf01a3bf9644c69b9ca5af88abb1202.ashx?la=en. [Accessed 01 Jul 2026].

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control. *Preprint*, arXiv:2405.16969.

Patrick Meier. 2015. *Digital Humanitarians*. CRC Press, Boca Raton.

Mary Nurminen. 2021. *Investigating the influence of context in the use and reception of raw machine translation*. **PhD**, Tampere University.

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.

OpenAI. 2024. Gpt-4o system card.

Sharon O'Brien and Federico Federici. 2019. Crisis translation: Considering language needs in multilingual disaster settings. *Disaster Prevention and Management*, 29(2):129–143.

América R. Gaviria Pabón, Makenzie J. Krocak, Joseph T. Ripberger, and Joseph E. Trujillo-Falcón. 2025. Linguistic challenges in severe weather communication: Comparing english- and spanish-speaking populations. *Bulletin of the American Meteorological Society*, 106(9):E1940 – E1954.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark.

Hemant Purohit, Cody Buntain, Amanda Lee Hughes, Steve Peterson, Valerio Lorini, and Carlos Castillo. 2025. Engage and mobilize! understanding evolving patterns of social media usage in emergency management. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Alessandra Rossetti, Sharon O'Brien, and Patrick Cadwell. 2020. Comprehension and trust in crises: Investigating the impact of machine translation and post-editing. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 9–18, Lisboa, Portugal. European Association for Machine Translation.

Pilar Sánchez-Gijón. 2022. Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience.

Vanessa Šorak. 2024. From terminological inconsistencies to major mistranslations: A qualitative analysis of nmt errors in public health communication. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 1(1).

Maria Carmen Staiano, Lifeng Han, Johanna Monti, and Francesca Chiusaroli. 2025. ITALERT: Assessing the quality of LLMs and NMT in translating Italian emergency response text. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 566–577, Geneva, Switzerland. European Association for Machine Translation.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Belu Ticona and Antonios Anastasopoulos. 2026. Llm-powered automatic translation and urgency in crisis scenarios. *Proceedings of the International ISCRAM Conference*, 23.

Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR Public Health Surveill*, 5(1):e11171.

U.S. Census Bureau. 2024. Language spoken at home by ability to speak english for the population 5 years and over. U.S. Census Bureau. Accessed 27 August 2026.

US DOJ LEP Cmte. 2021. Improving access to public websites and digital services for limited english proficient (lep) persons. https://niwaplibrary.wcl.american.edu/wp-content/uploads/LEP_Website_Language_Access_Guide_508.pdf. [Accessed 16 Jul 2026].

U.S. GSA. 2026. Introduction to translation technology — digital.gov. https://web.archive.org/web/20260625210818/https://digital.gov/resources/introduction-to-translation-technology. [Accessed 16 Jul 2026].

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

Melissa Villarreal, Carson MacPherson-Krutsky, and Mary Angelica Painter. 2025. Barriers and best practices for inclusive emergency alerts and warnings. *International Journal of Disaster Risk Reduction*, 125:105581.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Christopher Whyte and Brittany 'Brie' Haupt. 2026. Building centaur responders: is emergency management ready for artificial intelligence? *Disasters*, 50(3):e70054.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

APPENDIX

A Translation Dataset

As part of our collaboration with LAAWW, we are compiling a collection of pre-translated emergency alerts from multiple agencies across the United States. The alerts span various emergency types, including (*e.g.,* weather, public health, hazmat, etc.). Each alert includes the original English message, professional human translations into dozens of languages, and metadata such as agency, date issued, language pair, alert type, etc.

We intend to make this dataset available to all those in the emergency management community who wish to use these pre-translated alerts. We also intend to make the dataset available to the research community to support future work on translation quality evaluation, multilingual emergency communication, etc. We are actively growing this data set. If you have data to share, please contact the authors.

B Participant Information

Table 1 is intended to give the reader further insight into participant comments. Only survey participants who are quoted in this paper are included in this table.

Type of EM	ID	Juridisction Scope	Location
Local Emergency Manager	L1	County	West Coast
Public Health Coordinator	H1	County	West Coast
Public Information Officer	P1	City	East Coast
Advocate (Retired Researcher)	R1	State	West Coast

Table 1: Demographic information of quoted survey participants.

C Translation Needs and Speaker Populations

The language needs of each geographic area vary because populations are not evenly distributed across the nation. Figure 4 shows the difference between the national and NY/NJ populations who speak a language other than English as their primary language. The population of NY/NJ is approximately 5.85% of the total population of the United States. The trend line is the expected result if the percentages in NY/NJ were the same nationally. Languages that appear above the line are more heavily represented in NY/NJ while languages that appear below the line are more heavily represented nationally.

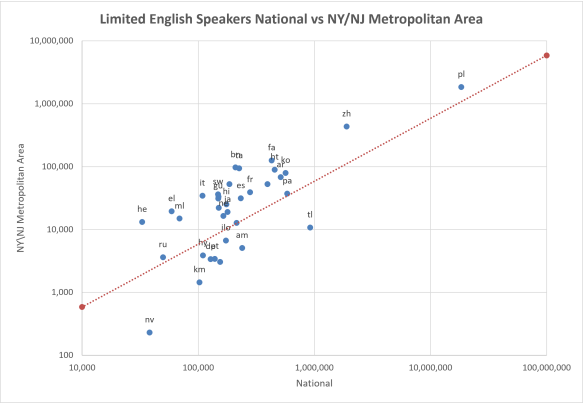


Figure 4: Count of Language Speakers in the NY/NJ Metropolitan area vs Nationally. Trend line represents the expected population if the percentages were identical in the NY/NJ Metropolitan area vs Nationally. (U.S. Census Bureau, 2024)

This data allows us to see some striking outliers in the populations. There are approximately 435,000 people who speak Chinese (zh) and "Speak English Less than Well" in NY/NJ accounting for nearly 23% of the approximately 1,895,000 people nationally, much higher than the expected 5.85%.

Table 2 reports a similar analysis across major metropolitan areas in the U.S, where the top row lists expected demographic share based on national population proportions. For example, Seattle-Tacoma-Bellevue, WA Metropolitan Area represents 1.22% of the U.S. population, but contains substantial demographic skew: eighteen language communities are over-represented (>2 x baseline) and eight are under-represented (<0.5 x baseline).

We can use Haitian Creole[23] as a specific language example. In Table 2 we see that it is underrepresented in nearly every area. This is because Haitian Creole speakers are disproportionally represented in Florida, were approximately 50% of Haitian Creole speakers live.

The data also shows that there is no language group within these areas that is spoken by the expected proportion of speakers. Each county, metropolitan area, or state has its own unique mix of languages, thus each area must to do their own language analysis to determine their community's language and communication needs.

D Languages and Language Varieties

If they chose, survey participants were able to provide the languages spoken by the communities they serve. The list below is an example of the

[23]The US Census uses the language name Haitian.

	National	Atlanta	Chicago	Dallas	Los Angeles	Minneapolis	San Francisco	Seattle
Total	321,745,943	1.88%	2.77%	2.43%	3.82%	1.10%	1.37%	1.22%
Amharic, Somali, or other Afro-Asiatic languages	238,986	3.73%	3.56%	4.96%	2.26%	14.53%	1.84%	6.31%
Arabic	513,293	0.48%	4.20%	3.23%	5.14%	0.65%	1.86%	1.11%
Armenian	109,630	0.00%	0.54%	0.00%	83.24%	0.00%	0.40%	0.65%
Bengali	209,543	1.66%	1.74%	2.17%	3.35%	0.32%	0.17%	0.71%
Chinese	1,895,001	1.36%	2.76%	1.33%	14.23%	0.54%	10.76%	3.22%
French (incl. Cajun)	279,230	2.76%	2.01%	3.11%	2.36%	1.22%	1.38%	1.61%
German	127,650	0.83%	1.98%	0.96%	1.86%	2.07%	0.79%	1.31%
Greek	59,054	0.73%	8.86%	0.10%	2.49%	0.00%	0.73%	0.28%
Gujarati	148,495	5.28%	12.67%	3.60%	2.88%	0.48%	1.35%	0.45%
Haitian	454,970	1.73%	0.18%	0.21%	0.13%	0.26%	0.03%	0.10%
Hebrew	32,924	1.66%	1.64%	0.60%	9.32%	0.29%	1.25%	2.37%
Hindi	174,630	3.09%	4.63%	4.39%	4.79%	0.45%	7.12%	2.66%
Hmong	91,284	0.69%	0.09%	0.38%	0.69%	28.45%	0.00%	0.19%
Ilocano, Samoan, Hawaiian, or other Austronesian languages	173,239	1.06%	2.40%	1.09%	6.79%	0.25%	3.78%	4.88%
Italian	108,644	0.52%	5.52%	1.42%	2.43%	0.35%	1.55%	0.08%
Japanese	179,098	1.20%	2.39%	1.51%	13.42%	0.54%	4.20%	3.72%
Khmer	102,600	2.19%	0.75%	2.36%	14.97%	2.36%	3.74%	7.49%
Korean	565,989	4.23%	3.60%	3.16%	22.66%	0.32%	2.30%	4.41%
Malayalam, Kannada, or other Dravidian languages	69,100	1.23%	6.26%	12.09%	1.32%	0.73%	1.99%	2.43%
Navajo	38,127	0.00%	0.00%	0.30%	0.00%	0.03%	0.00%	0.12%
Nepali, Marathi, or other Indic languages	164,869	2.94%	1.51%	7.76%	3.21%	0.64%	4.47%	1.40%
Persian (incl. Farsi, Dari)	213,566	2.43%	1.45%	2.27%	20.99%	0.37%	5.89%	5.02%
Polish	185,755	0.37%	35.58%	0.28%	0.99%	0.21%	0.37%	0.48%
Portuguese	394,010	1.63%	0.53%	1.20%	1.26%	0.18%	1.05%	1.22%
Punjabi	150,447	0.31%	2.01%	1.11%	2.46%	0.22%	5.77%	4.43%
Russian	429,378	1.01%	4.26%	0.57%	7.49%	1.32%	2.25%	3.48%
Serbo-Croatian	79,111	1.93%	13.69%	1.77%	2.22%	0.85%	0.72%	2.13%
Spanish	18,432,221	1.41%	3.46%	3.73%	9.51%	0.39%	1.71%	0.62%
Swahili or other languages of Central, Eastern, and Southern Africa	138,888	7.02%	0.97%	8.13%	0.84%	2.94%	0.25%	3.17%
Tagalog (incl. Filipino)	584,724	0.49%	3.09%	1.40%	14.52%	0.58%	9.23%	2.85%
Tamil	49,660	3.16%	3.91%	5.99%	2.76%	1.25%	5.10%	3.43%
Telugu	104,199	3.69%	2.32%	12.31%	0.84%	0.82%	3.73%	3.69%
Thai, Lao, or other Tai-Kadai languages	154,016	1.17%	2.80%	3.08%	9.62%	1.73%	3.13%	4.25%
Ukrainian or other Slavic languages	232,585	0.67%	15.75%	0.42%	2.69%	1.28%	1.03%	7.71%
Urdu	148,178	1.97%	9.39%	6.34%	3.96%	0.09%	1.92%	1.04%
Vietnamese	918,210	3.23%	1.37%	4.87%	18.03%	0.84%	3.11%	3.93%
Yiddish, Pennsylvania Dutch or other West Germanic languages	225,109	0.05%	0.21%	0.12%	0.17%	0.10%	0.00%	0.09%
Yoruba, Twi, Igbo, or other languages of Western Africa	149,225	2.09%	4.83%	2.93%	2.08%	4.24%	0.78%	2.53%

Table 2: Language Variation of actual vs. expected representation for select Metropolitan areas in the United States (U.S. Census Bureau, 2024). underrepresented = expected/2 overrepresented = expected*2
Languages spoken by fewer than 10,000 people are grouped with other languages by the U.S. Census Bureau. More information can be found here: https://usa.ipums.org/usa/volii/00language.shtml.

data they shared. If participants shared a language variety (as they defined it), it is included here. The data is formatted showing the principal language, followed by varieties (if provided), *e.g.,* **Language**: variety.

Ahtna; **Alutiiq / Sugpiaq**; **Amdo Tibetan**; **Amharic**; **Arabic**: Algerian, Chadian, Classical, Egyptian, Gulf, North Levantine, South Levantine, Yemeni; **Armenian (Western)**; **Aromanian**; **Ashanti**; **Avestan**; **Azeri**;

Bahamian Creole; **Bambara**; **Barbadian Cre-**

ole; **Bavarian**; **Belarusian**; **Bengali**; **Bislama**; **Brazilian**; **Bukhori**; **Burmese (Intha)**;

Calabrese; **Cambodian / Khmer**; **Cebuano**; **Chinese**: Cantonese, Mandarin, Simplified, Fujianese, Hainanese (Southern Min), Hokkien, Hunanese, Han, Shanghainese, Taishanese, Taiwanese, Wenzhounese; **Coptic**; **Crimean Tatar**; **Croatian**;

Dari; **Deg Xinag**; **Dena'ina**; **Dutch**; **Ecuadorian Kichwa**; **Efik**; **English**: African-American, Australian, Canadian, Ghanaian, Hawaiian, Indian, Irish, Italian, Jewish, Liberian, Native American, New Zealand, Newfoundland, Nigerian, Pakistani, Philippine; **Ewe**; **Eyak**; **Fante**; **Farsi**; **Finnish**; **French**; **Fulani**;

Galician; **Garifuna**; **Geg Albanian**; **Georgian**; **German**: Austrian, Gottscheerisch, Swabian; **Greek**: Cretan, Cypriot, Koine; **Grenadian Creole**; **Gujarati**; **Guyanese Creole**; **Gwich'in**; **Haida**; **Haitian Creole**; **Hakha Chin**; **Hamgyŏng Korean**: Joseonjok; **Hebrew**: Biblical; **Hindi**; **Hmong**; **Holikachuk**; **Hungarian**;

Icelandic; **Igbo**; **Ilocano**; **Indonesian**; **Inupiaq**; **Irpino**; **Italian**; **Iu Mien**; **Jamaican Patois**; **Japanese**; **Jewish Babylonian Aramaic**; **Judeo-Shirazi**; **Judeo-Spanish**; **Juhuri**; **Kalderash Romani**; **Kannada**; **Karen**; **Kinyarwanda**; **Kirundi**; **Konkanoi**; **Korean**; **Koyukon**;

Lahu; **Lao**; **Lemko**; **Lenape**; **Lithuanian**; **Lukumí**; **Maithili**; **Malayalam**; **Maltese**; **Mandinka**; **Maranao**; **Marathi**; **Marshallese**; **Marwari**; **Matinecock**; **Mixtec**; **Mohican**; **Mon**; **Montaukett**; **Montserrat Creole**;

Nahuatl; **Nauruan**; **Navajo**; **Ndebele**; **Neapolitan**; **Nepali**; **Newari**; **Noakhailla**; **North Frisian**; **Nuer**; **Pashto**; **Polish**; **Portuguese**: Brazilian, Portuguese; **Pugliese**; **Punjabi**; **Quiripi**; **Ramaluk**; **Romagnol**; **Romanian**; **Russian**;

Sandwippa; **Sanskrit**; **Serer**; **Sherpa**; **Sicilian**; **Sinhalese**; **Somali**; **Soninke**; **Spanish**: Argentine, Bolivian, Castilian, Central American, Colombian, Cuban, Dominican, Ecuadorian, Mexican, Puerto Rican, Venezuelan; **Spanglish**; **St. Lawrence Island Yupik**; **St. Lucian Creole**; **Swahili**; **Sylheti**;

Tagalog; **Tajik**; **Tamang**; **Tamil**; **Tanacross**; **Tanana**; **Telugu**; **Tetum**; **Thai**; **Tibetan**: Classical, Tö; **Tlingit**; **Tosk Albanian**; **Trinidadian Creole**; **Tsimshian**; **Turkish**; **Twi**; **Ukrainian**; **Unangam Tunuu / Aleut**; **Upper Kuskokwim**; **Upper Tanana**; **Urdu**; **Uzbek**; **Yiddish**: Hasidic, Litvish; **Yup'ik**; **Zulu**.

E More on Post-editing and Evaluating Translations

Here are some rules of thumb for determining the adequacy of the translations provided by a commercial translation tool (or of any translations):

- Use Multidimensional Quality Metric (MQM) (Lommel et al., 2024) to define what *major errors* might exist. Ultimately, there should be no *major errors*, but minor errors may be acceptable.[24]
- Critical data, such as place names, should be kept in the original language because geographic place names tend to stay consistent and do not change based on language spoken. This information can be evaluated even by those who do not read the target language.

Post-editing is often done using Computer Assisted Translation (CAT) tools such as Trados or MemoQ[25], although it can be done in tools such as Word or Excel (with a little more difficulty). One of the advantages of using CAT tools is that they track translations and can reuse material from earlier translations (through *Translation Memories*). Many professional translators are familiar with these tools.

In scenarios where time is limited or it is not possible to do a formal evaluation, one can do round-trip translation and cross checking. For round-trip translation, use an MT tool to perform the initial translation into the target language. Once the translation output is generated, copy the translated text into a different MT tool and translate it back into the original language. If the new translation preserves the meaning, then the translation is likely usable. To use cross checking, use two different MT tools to produce a translation for the same text and compare the outputs. If the outputs are identical and likewise pass the back translation test, then the translation is likely usable.

There are many other aspects of translation evaluation, including fluency, tone, meaning, terminology, additions/omissions, cultural appropriateness, localization, accessibility, etc. These are beyond the scope of this paper.

[24] Examples of major errors may include incorrectly translated safety instructions, incorrect locations, or incorrect times.

[25] https://www.trados.com/, https://www.memoq.com/

F Survey Questions

The survey was presented to LAAWW membership during a monthly LAAWW meeting. Survey goals and methodologies were explained to the group and time was given for questions.
The survey questions below are reproduced as administered. Response choices and formats are provided. If *Other* was a Response Choice, participants were able to use a Free Response to answer.

1. **Which of the following best describes your current role(s) in emergency management?**
   Response format: Multiple Choice; Text
   Response choices: Local emergency manager; State emergency manager; Public information officer; Language coordinator / interpreter; Operations, logistics, or communications manager; Researcher; Other

2. **Please list the city, county, or region your agency or institution primarily serves.**
   Response format: Free response

3. **Approximately how many different languages are spoken by the communities you serve?**
   Response format: Free response

4. **If you have a convenient list of the languages, please paste it here or share a URL.**
   Response format: Free response

5. **Do you translate FROM languages other than English?**
   Response format: Yes-no style response
   Response choices: Yes; No

6. **What languages do you translate FROM other than English?**
   Response format: Free response

7. **Feel free to share more information about any of your answers from this section.**
   Response format: Free response

8. **How does your agency/institution identify the languages to prioritize for translation and outreach?**
   Response format: Multiple Choice; Text
   Response choices: US Census or American Community Survey data; Your State's Department of Education Data; Local community or needs assessments; Input from schools, health, or social service agencies; Feedback from community organizations; Staff or volunteer knowledge; Other

9. **What are the translation tools used by your agency/institution?**
   Response format: Multiple Choice; Text
   Response choices: Google Translate; Microsoft Translator; DeepL; ChatGPT; DeepSeek; Google Gemini; Staff member; Contractor; Community member or volunteer; Other

10. **What are the main reasons your agency/institution selected these tools or methods?**
    Response format: Multiple Choice; Text
    Response choices: Cost or budget constraints; Availability or ease of access; Accuracy or quality of translation; Easy to use or staff comfort; Compliance or policy requirements; Other

11. **Has your agency/institution evaluated how well these tools work for accuracy, usability, or effectiveness?**
    Response format: Yes-no style response
    Response choices: Yes - formal evaluation; Yes - informal evaluation; No

12. **Does your agency/institution review or update its translation tools or methods?**
    Response format: Yes-no style response
    Response choices: Yes - annually; Yes - as needed; No formal review process; Not sure; No

13. **Who usually handles translations or multilingual communication within your agency/institution?**
    Response format: Multiple Choice; Text
    Response choices: Dedicated translation team; Individual staff member; Shared among staff; External vendor or contractor; Volunteers / community partners; Other

14. **Feel free to share more information about tools, processes, personnel, and capacity here.**
    Response format: Free response

15. **What additional tools, training, or partnerships would most help your agency improve multilingual communication and translation capacity?**
    Response format: Free response

16. **What are some challenges?**
    Response format: Free response

17. **May we contact you if we have follow up questions? If so, please provide your preferred contact information.**
    Response format: Free response

G Evaluation Results

Language	Microsoft	Google	GPT-5.4-mini	Gemini 2026	Claude
Pivot					
Arabic	7.34	7.42	7.66	8.34	7.67
Chinese	11.97	10.66	13.02	11.10	12.67
French	5.67	4.85	6.03	4.89	5.47
Hindi	13.16	9.28	10.33	8.34	12.24
Indonesian	1.98	3.33	2.97	3.03	2.44
Portuguese (BR)	4.41	4.41	4.42	4.69	4.93
Russian	6.66	8.06	6.68	6.97	6.70
Spanish (LA)	4.04	6.14	3.94	4.03	3.75
Swahili	-2.96	-3.13	-3.04	-1.54	-2.30
Important					
Bengali	1.65	1.05	0.04	-0.66	-0.13
Farsi	1.99	3.53	4.78	4.17	3.96
Marathi	2.07	1.72	1.20	0.90	1.70
Myanmar	1.95	3.69	2.22	2.72	3.88
Tagalog	1.47	2.52	2.13	2.40	0.75
Tamil	0.03	0.28	-0.78	0.98	0.25
Urdu	1.34	2.96	2.56	2.59	2.04
Priority					
Amharic	4.61	4.47	1.96	3.73	1.41
Dari	1.25	–	1.18	0.80	2.29
Khmer	-6.54	-3.96	-3.40	-1.36	-2.54
Kinyarwanda	1.32	0.03	-0.39	-1.08	0.60
Nepali	3.45	3.91	4.65	4.35	3.76
Pashto	5.94	10.11	5.34	2.98	4.96
Somali	-0.21	-0.23	-1.01	0.07	-0.25
Zulu	-2.19	-2.93	-1.38	-2.17	-0.59

Table 3: Differences between Overall CHRF++ Score and News+Announcements CHRF++. Blue indicates Overall score is higher than N+A, orange indicates Overall is lower than N+A. Google Translate results were unavailable for Dari. Khmer, Myanmar, and Chinese do not use whitespace to mark word boundaries and CHRF was used instead. (Bansal et al., 2026)

	Microsoft		Google		GPT-5.4-mini		Gemini 2026		Claude	
Language	**Overall**	**N+A**	**Overall**	**N+A**	**Overall**	**N+A**	**Overall**	**N+A**	**Overall**	**N+A**
Pivot										
Arabic	55.08	47.73	**59.23**	**51.82**	57.86	50.20	57.27	48.93	55.64	47.97
Chinese	45.74	33.77	**50.48**	**39.82**	48.10	35.08	42.66	31.56	45.89	33.21
French	63.64	57.97	63.13	58.28	**65.52**	**59.49**	63.05	58.16	63.02	57.55
Hindi	**65.62**	52.46	61.76	52.48	62.81	52.48	61.75	**53.41**	62.42	50.18
Indonesian	74.72	**72.74**	74.88	71.54	**74.93**	71.95	73.92	70.88	73.74	71.30
Portuguese (BR)	72.06	67.65	72.05	67.64	**72.99**	**68.56**	**72.99**	68.30	72.62	67.69
Russian	56.13	49.47	**59.57**	**51.51**	58.11	51.43	58.19	51.23	57.14	50.44
Spanish (LA)	71.88	67.84	**74.17**	68.03	74.08	**70.14**	73.64	69.60	72.84	69.09
Swahili	59.96	62.92	**61.59**	**64.72**	60.44	63.48	60.16	61.70	55.86	58.16
Important										
Bengali	**51.40**	**49.75**	46.69	45.64	47.25	47.21	44.38	45.04	47.55	47.68
Farsi	52.32	50.33	**56.44**	**52.92**	52.71	47.93	51.80	47.63	51.38	47.42
Marathi	**46.80**	**44.73**	46.05	44.33	44.97	43.76	44.33	43.43	41.57	39.87
Myanmar	48.16	46.21	**54.56**	50.87	53.99	**51.77**	52.62	49.90	42.84	38.95
Tagalog	64.72	63.25	**68.27**	**65.75**	66.61	64.48	63.42	61.02	63.83	63.08
Tamil	**56.87**	**56.85**	53.11	52.84	48.75	49.53	50.68	49.70	46.48	46.23
Urdu	50.77	49.42	**52.80**	**49.84**	50.67	48.12	49.17	46.58	49.37	47.33
Priority										
Amharic	**35.84**	31.23	35.78	**31.31**	31.23	29.27	32.62	28.89	27.60	26.18
Dari	**48.15**	**46.90**	–	–	44.23	43.05	44.56	43.76	43.63	41.34
Khmer	60.53	**67.07**	**61.60**	65.66	57.52	60.92	56.86	58.22	46.66	49.20
Kinyarwanda	45.88	44.56	**47.87**	47.84	47.74	48.13	47.50	**48.59**	33.23	32.64
Nepali	53.42	49.98	**56.19**	**52.27**	52.39	47.75	49.35	45.00	50.25	46.49
Pashto	39.47	33.53	**48.52**	**38.40**	42.25	36.91	40.10	37.12	33.66	28.70
Somali	44.92	45.12	**46.62**	**46.85**	45.44	46.46	44.53	44.46	37.18	37.44
Zulu	52.76	54.95	**53.61**	**56.54**	53.34	54.73	50.31	52.49	42.54	43.13

Table 4: Overall and News+Announcements (N+A) CHRF++ scores for English→Target Language across translation systems and languages. Highest scoring language result for Overall and N+A is in bold. Shading indicates relative score ranges within the evaluation: blue represents higher scores (>60), pale green represents mid-range scores (45-60), and orange represents lower scores (<45). CHRF++ does not have a universal safety or usability threshold, so these colors indicate relative performance only. Google Translate results were unavailable for Dari. Khmer, Myanmar, and Chinese do not use whitespace to mark word boundaries and CHRF was used instead. (Bansal et al., 2026)