LLM-Powered Automatic Translation and Urgency in Crisis Scenarios

Belu Ticona*
George Mason University
United States
mticonao@gmu.edu

Antonios Anastasopoulos
George Mason University
United States
antonis@gmu.edu

ABSTRACT

Large language models (LLMs) are increasingly proposed for crisis preparedness and response, particularly for multilingual communication. However, their suitability for high-stakes crisis contexts remains insufficiently evaluated. This work examines the performance of state-of-the-art LLMs and machine translation systems in crisis-domain translation, with a focus on preserving urgency, a critical property for effective crisis communication and triage. Using multilingual crisis data (TICO-19, 30 languages) and a newly introduced urgency-annotated dataset of 100 scenarios translated into 29 languages, we show that dedicated translation models and LLMs exhibit substantial quality degradation, particularly for low-resource languages. Beyond translation quality, we conduct a human annotation study revealing a striking asymmetry: human assessors maintain consistent urgency judgments regardless of prompt language, while LLM-based urgency classifications vary widely across languages for identical scenarios, at times spanning the full range from *Not Urgent* to *Critical*. These findings highlight significant risks in deploying general-purpose language technologies for crisis triage and underscore the need for multilingual, human-centered evaluation frameworks.

Keywords

AI-mediated communication, crisis translation, large language models

INTRODUCTION

Large language models (LLMs) and related automated language technologies (LTs) have attracted growing scholarly and practical interest across all stages of crisis management (Lei et al., 2025). Beyond traditional natural language processing (NLP) tasks (such as classification, information extraction, and question answering), LLMs are increasingly capable of more complex operations, including summarization, text generation, and multistep reasoning.

These capabilities substantially broaden their potential applications in situational awareness, resource and needs coordination, risk assessment, and early warning, among others (Odubola et al., 2025). Furthermore, LLMs hold promise not only for managing and processing complex, heterogeneous data streams, but also for delivering operational support to the diverse stakeholders involved across the crisis management ecosystem (F. Xu et al., 2025).

However, the suitability of LLMs for multilingual crisis contexts remains largely underexamined. Most existing work draws on social media data, which is mostly available in a limited number of languages (D. Lewis et al., 2025). The few multilingual works that cover typologically and linguistically diverse languages focus only on specific tasks and types of crises (Abdul-Mageed et al., 2021; Imran et al., 2022). More specifically, machine translation evaluation in the crisis domain using LLMs and translation models (e.g., (Al Amer et al., 2023; Casacuberta et al., 2024; Lankford & Way, 2024)) remains scarce and often fails to cover languages from underserved communities and regions across the globe.

The rapid popularization of commercial LLMs and their availability in smaller, more affordable variants (e.g., GPT-4o mini, Gemini Flash Lite) has accelerated their widespread adoption as general-purpose systems. However,

*corresponding author

WiP Paper – Social Media & Crisis Communication: Narratives, Signals, and Sentiments
Proceedings of the 23rd ISCRAM Conference – The Hague, the Netherlands June 2026
Caroline Rizza, Apoorva Chauhan, Amy Matser, Joyce Kox, Willem Treurniet and Jeroen Wolbers eds.

these models are predominantly instruction-tuned on English data, and their multilingual capabilities emerge incidentally from large-scale pretraining rather than from explicit multilingual design.

Furthermore, it remains unclear how well LLMs align with human judgment across dimensions such as expertise, cultural context, and domain-specific values, a concern that is particularly critical in high-stakes domains such as crisis management, where expert knowledge is essential. In this context, human-AI collaboration emerges as a promising avenue for responsible LLM adoption, underscoring the need for robust, domain-specific, and human-centered evaluation frameworks.

In this sense, our work addresses the following research questions:

RQ1. Crosslingual Performance Variability. How does the performance of LLMs and translation models vary on crisis-related data across linguistically diverse languages? Are there systematic trends associated with geographic region?

RQ2. Human-LLM Alignment in Urgency Perception. To what extent does LLM performance on crisis-specific tasks align with human judgment? Are LLMs consistent across languages, and how do language and cultural context influence this alignment?

To answer these questions, in the MT For Crisis Domain Section (§3) we evaluate LLMs and dedicated translation models using COVID-19 human-translated parallel corpora spanning 30 geographically diverse languages. We focus primarily on open-weight models, as these are more accessible and adaptable than large commercial systems, a distinction that is particularly relevant for humanitarian NGOs, local governments, and crisis responders who serve affected communities in underserved regions, often under significant resource constraints.

In the Human and LLM Perception of Urgency Section (§4), we examine how LLM performance on urgency classification, a crisis-relevant task in which translation operates as an underlying process, varies across languages. We further conduct a small-scale human evaluation to assess the degree to which LLM urgency judgments align with those of human assessors in crisis contexts.

Our results show that both dedicated machine translation systems and state-of-the-art LLMs exhibit highly variable performance in the crisis domain across diverse languages. Critically, we demonstrate that translations that appear linguistically adequate can nonetheless distort the conveyed level of urgency, and that urgency judgments produced by LLMs vary substantially depending on the language of the prompt and input content.

These findings raise serious concerns about the consistency and reliability of LLMs for crisis-relevant tasks, not only in multilingual settings, but more broadly as systems whose alignment with human judgment remains poorly

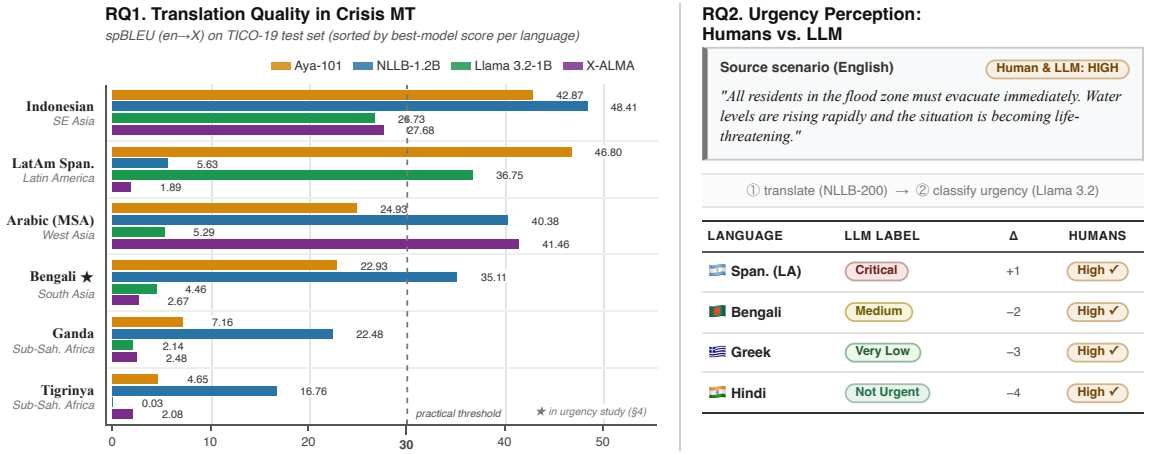

Figure 1. Left (RQ1): spBLEU scores (*en*→*X*) on TICO-19 crisis data across four model families. All models collapse for Sub-Saharan African languages (Ganda, Tigrinya); Latin Am. Spanish exposes extreme inter-model inconsistency (Aya-101: 46.80 vs. NLLB-1.2B: 5.63). The dashed line marks the practical threshold (≥30). Right (RQ2): Urgency classification of a `UrgencyScenarios` scenario translated into three languages via NLLB-200 and classified by Llama 3.2 across 30 languages. Human annotators agree across all languages (High); LLM labels span four distinct categories for the same source sentence, ranging from *Not Urgent* to *Critical*.

understood. This is especially consequential for crisis practitioners and humanitarian actors seeking to deploy such systems, even within human-AI collaboration frameworks, to serve affected populations globally.

Our work underscore the risks of adopting general-purpose LLMs in high-stakes crisis environments without rigorous, domain-specific evaluation. We therefore advocate for multilingual, human-centered evaluation of LLMs on crisis tasks, and for deployment practices grounded in the operational realities of emergency response. Figure 1 provides a joint preview of our main findings: translation quality collapses for low-resource languages (RQ1), and LLM urgency labels are highly unstable across languages for the same source scenario while human assessments remain consistent (RQ2).

RELATED WORK: LANGUAGE TECHNOLOGIES FOR CRISIS RESPONSE

Automatic Translation for Crises MT for crisis has progressed from early rule-based approaches, such as the Apertium initiative for Kurdish languages (Forcada & Tyers, 2016; Global CLEAR, 2016), to the rapid-response statistical MT frameworks established during the 2010 Haiti earthquake (Lewis, 2010). This led to large-scale, cross-institutional programs like DARPA LORELEI, which advanced neural MT specifically for health-related crisis communication in languages such as Arabic and Swahili (Strassel & Tracey, 2016; Tracey et al., 2019).

The COVID-19 crisis motivated numerous parallel data collection initiatives across multiple languages (Anastasopoulos et al., 2020; Casacuberta et al., 2024), producing some of the few crisis-domain parallel corpora available to date. However, fine-tuning efforts built on this data have largely focused on a small number of European languages (Lankford et al., 2024; Roussis, n.d.; Way et al., 2020).

More recent studies have explored the potential of LLMs for crisis MT in low-resource languages, comparing vanilla and fine-tuned versions of commercial LLMs with MT models specifically developed for low-resource settings, such as NLLB (Team et al., 2022). For instance, Lankford and Way show that fine-tuned NLLB on in-domain data outperforms custom versions of GPT-4 for Marathi and Irish.

Concurrent with our work, Merx et al. (2025) introduce OpenWHO, a parallel corpus drawn from the World Health Organization's e-learning platform, and evaluate a set of MT and LLM models including NLLB-54B, Gemini, DeepSeek-v3, and Gemma 3. While this represents a meaningful contribution to health-domain MT, the low-resource subset covers 9 languages with approximately 200 parallel sentences each, and the evaluation focuses on large commercial models.

Our work is complementary in this regard, offering a broader multilingual scope across the 30 typologically diverse languages evaluated from TICO-19, with a focus on open-weight models more accessible to the humanitarian organizations and low-resource institutions most likely to serve crisis-affected communities. Collectively, these efforts highlight meaningful progress in crisis MT, yet parallel corpora for other crisis types remain absent from the literature, and coverage of typologically diverse and low-resource languages is still limited.

Urgency Assessment Among the applications of LLMs in crisis settings, automatically assessing the urgency of incoming information has emerged as a key task for disaster situation assessment and response prioritization (Lei et al., 2025). Studies have shown that LLM-based urgency assessment can match untrained emergency health personnel but still fall short of professionally trained doctors, such as emergency department staff (Masanneck et al., 2024).

Moreover, Lee et al. evaluated commercial LLMs in real-life clinical conversations, showing that they can accurately classify the urgency of emergency department patients. However, these evaluations are conducted solely in English, lacking a multilingual assessment. Khullar et al. tackle this gap by evaluating LLMs (such as GPT-4-o, Claude 4.5, and Qwen3) in 6 South Asian languages, studying the effect of orthographic variations on real-world data in the health domain. Their results show a degradation in performance for Romanized messages, revealing a blind spot in LLM-based urgency classification systems. Our work expands the scope of prior work by broadening the range of languages covered and including open-weight LLMs.

MT FOR CRISIS DOMAIN: THE COVID-19 CASE STUDY

In this section, we evaluate the performance of language models in MT in languages linguistically and geographically diverse from different regions of the world. Our goal is to evaluate these models specifically in the *crisis domain*, since their performance generally depends on the amount and type of data in which they are trained on.

We understand by *crisis domain* to all tasks that involve the process and management of data related to crisis, covering their wide definitions according to the theoretical frameworks in used[1].

[1] In the crisis literature, there exist frameworks that define *crisis* in an event-based (Lerbinger, 2012), attribution-based (Coombs & Holladay, 2010), and temporal-dynamics (Paraskevas, 2013) perspective, among others

In this work, we focus our study in public health crisis, using COVID-19 data as a representative case. The scale and globally implications of this event motivated the creation of parallel corpora in multiligual settings, which makes it more suitable for our evaluation. Furthermore, this is one of the only crises type for which parallel corpora is available across a wide range of languages globally.

With regard to the models, our evaluation primarily considers open-weight models, as their accessibility and potential for fine-tuning make them more suitable stakeholders with limited resources that served in vulnerable regions of the globe (e.g. humanitarian NGOs). Accordingly, our evaluation covers languages less-represented in the Internet from Africa and Asia mainly.

Methodology

Data We use the TICO-19 dataset in our evaluations, a publicly available crisis-domain parallel corpus covering linguistically and geographically diverse languages from Africa and Asia mainly, as well as few languages from Latin America and Europe (Anastasopoulos et al., 2020). This dataset encompasses a range of sources, including Wikipedia articles, PubMed publications, news reports, and NGO communications, human-translated from English into 39 low- and medium-resource languages.

For our experiments, we use the test set, which contains approximately 2,000 sentences per language, comprising a total of 70,000 sentences.

Models We evaluate a set of open-weight models spanning two complementary paradigms: dedicated multilingual MT systems and general-purpose LLMs. This selection reflects both the current landscape of low-resource MT and the growing role of LLMs in translation tasks for specialized domains such as crisis response.

We include `NLLB-200` as a state-of-the-art multilingual MT system built explicitly for low-resource languages; `X-ALMA` as an LLM-based MT model with a modular multilingual architecture; `Aya-101` as a multilingual LLM trained with broad language diversity as a core design objective; and `Llama` 3.2 as a widely adopted general-purpose LLM that, despite limited official multilingual support, serves as a common backbone for domain- and language-specific fine-tuning.

The evaluations were conducted in June 2025 for `NLLB-200` and `Aya-101`, and October 2025 for `X-ALMA` and `Llama` 3.2. We report the exact versions used for each model below.

`NLLB-200` is a suite of neural MT models covering more than 200 languages, developed by Meta as part of the *No Language Left Behind* initiative,[2] with an explicit focus on lesser-resourced languages underserved by existing MT systems. As a dedicated multilingual MT system, it constitutes a natural reference point for low-resource translation evaluation, and is widely used as a backbone for fine-tuned MT systems targeting specific low-resource language families, such as indigenous languages of the Americas (De Gibert et al., 2025) and languages of India (Pakray et al., 2025). We use the open-weight variant released through the Open Neural Machine Translation project.[3]

`Aya-101` is a multilingual generative model developed by Cohere Labs, supporting 101 languages of which approximately 50% are considered low-resourced (Üstün et al., 2024). Unlike most multilingual instruction-following models, Aya-101 was explicitly designed to counteract English-centricity in training data, with only 21.5% of its instruction data in English. We include it as a representative multilingual LLM with broad cross-lingual coverage and a strong orientation toward underrepresented languages. We use the official open-weight model available on Hugging Face.[4]

`Llama` 3.2 is a family of multilingual generative models, available in 1B and 3B parameter sizes, released by Meta in both pretrained and instruction-tuned variants (Grattafiori et al., 2024). Although the model officially supports only eight high-resource languages, it is widely adopted across applied NLP tasks due to its open-weight availability and strong general-purpose performance, making it a common backbone for language- and domain-specific fine-tuning. We include Llama 3.2 to assess the zero-shot translation capabilities of a general-purpose LLM not explicitly optimised for low-resource or multilingual scenarios, providing a contrasting reference point to the other models in our evaluation.

`X-ALMA` is an open-weight LLM-based MT model with a modular architecture that groups linguistically similar languages into dedicated modules, officially supporting 50 languages (H. Xu et al., 2025). Its plug-and-play design enables language-specific components to be integrated or extended without retraining the full model, making

[2] https://ai.meta.com/research/no-language-left-behind/
[3] https://github.com/gordicaleksa/Open-NLLB
[4] https://huggingface.co/CohereLabs/aya-101

it well-suited for adaptation to new languages within an existing linguistic family. We include X-ALMA as a representative of the emerging class of LLM-based MT systems that combine the generative capabilities of large language models with translation-specific training objectives. We use the official release available on GitHub.[5]

Metrics We evaluate translation performances using SacreBleu (Post, 2018), a well-known MT evaluation framework that computes automatic metrics based on the lexical distance between the hypothesis sentence and its reference translation.

We report two complementary metrics: 1) spBLEU, a variation of the BLEU score that operates on sub-words units to better handle diverse scripts (Goyal et al., 2022), and 2) chrF++ (Popović, 2017), which improves on upon standard-level evaluation by incorporating word n-grams to increase robustness for morphological variations.

Analysis of Crosslingual Performance

RQ1. Crosslingual performance variability. How does the performance of LLMs and translation models vary on crisis-related data across linguistically diverse languages, and what systematic trends emerge regarding directionality and geographic regions?

In this section, we discuss our results in response to our primary research question. Our findings indicate that machine translation (MT) in the crisis domain remains a significant challenge for the evaluated LLMs and specialized models.

We identify a systematic trend of **asymmetry in translation direction**: quality from English to other languages ($en \rightarrow X$) is consistently lower than translation quality into English ($X \rightarrow en$). This pattern suggests that while these tools may assist in information gathering, they are less reliable for disseminating critical instructions to affected communities, thereby diminishing the utility of crisis communication technology in multilingual settings.

To facilitate metric interpretability, we apply a general heuristic: spBLEU values < 20 indicate significant information loss; values between 20–29 indicate that the translation conveys general meaning but contains major grammatical and semantic errors; and values > 35 represent understandable translations. For practical crisis applications, we consider scores > 40 suitable for situational awareness and > 50 for general purposes. While `NLLB-200` demonstrates the highest consistency across both directions, maintaining a mean spBLEU > 30 and winning in 18 out of 30 languages, other models exhibit total failure in specific geographic regions.

Specifically, we observe high performance volatility in `X-ALMA` and `Llama` 3.2. These models perform well for certain varieties, such as Brazilian Portuguese or Latin American Spanish, but **collapse** for languages from other regions, such as Ganda, Dari, and Marathi (spBLEU < 15). This geographic trend highlights the necessity of preliminary model assessment on local linguistic varieties before deployment. A model that functions for one ethnic group may fail for another, leading to disparate impacts on aid delivery and evacuation safety.

More concerningly, for several languages spoken in conflict-prone or disaster-relevant regions, all evaluated models exhibit significant degradation (e.g., a maximum spBLEU of 16.76 for Tigrinya and 20.82 for Sorani Kurdish). This systematic limitation suggests that for these specific linguistic domains, machine translation is currently not recommended for critical applications, leaving human translation as the only viable and safe option.

Limitations

Our evaluation is limited in the number of languages and models covered, which is mainly due to the limited existing human-translated parallel corpora in multiple languages to conduct a comparative analysis. Our work focuses only on vanilla versions[6] of translation models commonly used in academic spaces as based models for domain and language/region specific versions (e.g. Sánchez et al. (2025)). Further improvement could be achieved if they were fine-tuned in crisis health data.

Moreover, our evaluation only covers two of the most traditional automatic string metrics, leaving others for further studies. Future work could focus on language-specific evaluations, covering specific metrics according to the language and translation direction. For instance, Li et al. (2025) introduce SSA-COMET which outperforms other semantic metrics thanks to metric training on large-scale human annotated MT evaluations in African languages. Furthermore, other work analyze limitations of semantic metrics when used in unseen language in training (Zebaze et al., 2025; Zouhar et al., 2024).

[5] https://github.com/fe1ixxu/ALMA

[6] Models without adaptation to specific-domain data (e.g. without fine-tuning)

Other limitation of our work is that we only explore a zero-shot prompt technique and models are known for their high variability to prompting techniques. A future improvement could explore multiple executions and example selection, which have especially shown better results in the high-to-low resource translation direction (Zebaze et al., 2025).

Discussion and Future Work

In summary, while models specialized in language diversity and translation show promise for local situational monitoring, our results confirm that they remain unreliable for crisis communication dissemination. Deploying off the shelf models in highly specialized domain scenarios requires dedicated assessment on the local languages of the deployment context, as performance inconsistencies can lead to the spread of life threatening misinformation, the deterioration of institutional trust, or the exclusion of linguistic communities from essential aid.

These findings align with broader trends in low resource machine translation, where fine tuned systems remain the strongest baseline across diverse regions, including Indigenous languages of the Americas (De Gibert et al., 2025), low resource Indic languages (Pakray et al., 2025), and African languages, where even commercial LLMs fall short of fine tuned baselines (Ojo et al., 2025).

In the crisis domain, however, such fine tuning remains out of reach for most languages, since parallel crisis corpora are scarce and the infrastructure needed to build and evaluate domain adapted systems does not exist for many of the languages most exposed to humanitarian emergencies. Our evaluation of vanilla models therefore reflects a realistic picture of what is currently deployable in crisis affected regions.

Future work could extend this evaluation to include commercial and smaller instruction tuned models (Ojo et al., 2025), as well as more suitable metrics depending on the language. For instance, while some languages could be better assessed using semantic metrics due to the availability of MT human evaluation data, others could be better studied using aggregate approaches based on lexical metrics as used in the present work (Cavalin et al., 2025).

More broadly, most models are trained on general-domain data that does not reflect the linguistic and situational characteristics of crisis scenarios, and even state-of-the-art systems struggle with robustness to non-standard input and domain-specific terminology (Kocmi et al., 2025). In crisis settings, subtle changes in wording could alter risk perception, delay response, or erode trust, making it essential to go beyond translation accuracy and assess the faithful preservation of communicative intent.

This points to the need for human-centric evaluation frameworks that examine how practitioners actually rely on language technology to perform their tasks (Carpuat et al., 2025), a point we reinforce in the following section, where we study LLMs in urgency classification when translation is an underlying task.

HUMAN AND LLM PERCEPTIONS OF URGENCY

In this section, we investigate how translation quality influences crisis communication when LLMs are integrated into crisis-related applications. While Lei et al. (2025) analyze LLM utility across diverse tasks, including sentiment analysis, vulnerability detection, and situational assessment, as well as generative tasks like the automatic creation of evacuation plans, our work focuses on classification as a primary approach.

Classification represents a lower-complexity task that allows for a direct comparison between human and LLM distribution patterns, avoiding the high degree of nuance and discrepancy often found in generative outputs.

Specifically, we center our analysis on the dimension of *urgency*. We explore the similarities and differences in how humans and LLMs perceive urgency when exposed to content conveying varying levels of gravity.

Although this study is limited to a single dimension, this methodology is extensible to other dimensions such as severity, factuality, or sentiment, among others. Through this approach, we aim to determine how translation quality impacts the conveyance of meaning for critical *proxies* of the semantic dimensions essential to effective crisis translation.

Methodology

There exist limited human-translated parallel corpora to conduct fair evaluation in crisis domain, as discussed in the previous section. In this sense, our approach explores the use of LLMs to create synthetic data that span a wide level of urgency.

We expect real-world data to reflect even more complex nuances on urgency assessment, and therefore, being a more complicated task that the one addressed in this work. In other words, if LLMs' performance on urgency assessment using synthetic data shows poor alignment with humans, the use of real-word crisis data would perish even more this alignment.

Table 1. Translation Quality ($en \rightarrow X$) Across Geographic Groups. Bold indicates the highest spBLEU score per language. NLLB-1.2B remains the most consistent for generation, while X-ALMA shows surprising strength in specific languages like Somali and Kinyarwanda.

Language	Aya101		NLLB-1.2B		Llama3.2-1B		X-ALMA	
	spBLEU	chrF2++	spBLEU	chrF2++	spBLEU	chrF2++	spBLEU	chrF2++
Africa								
Amharic	13.23	23.80	**25.27**	**34.22**	0.49	2.15	1.31	2.08
Ganda	7.16	17.95	**22.48**	**42.66**	2.14	12.71	2.48	5.85
Hausa	**20.36**	**41.56**	18.12	36.70	1.60	12.07	2.29	3.28
Kinyarwanda	10.17	28.88	40.99	56.70	0.95	9.62	**55.92**	**70.54**
Lingala	6.20	20.41	**21.73**	**44.36**	1.96	13.26	2.64	5.13
Nigerian Fulfulde	3.70	11.33	**44.98**	**62.14**	0.93	8.11	33.61	51.30
Somali	8.98	27.01	24.97	47.53	0.56	6.85	**38.24**	**51.98**
Swahili	**29.60**	**51.92**	12.38	31.43	2.93	20.02	1.36	1.87
Tigrinya	4.65	11.33	**16.76**	**26.73**	0.03	1.28	2.08	2.97
West Cent. Oromo	4.03	18.54	34.91	53.10	0.04	0.18	**38.17**	**55.24**
Southern and Eastern Asia								
Bengali	22.93	39.81	**35.11**	**49.29**	4.46	19.94	2.67	4.02
Burmese	18.19	35.62	**51.79**	**70.84**	0.05	0.62	49.72	68.72
Hindi	**29.62**	47.92	29.39	**51.78**	13.23	32.31	2.02	3.11
Indonesian	42.87	64.22	**48.41**	48.41	26.73	50.52	27.68	44.66
Marathi	14.46	34.60	**23.73**	**48.92**	1.29	8.04	2.66	9.75
Nepali	**24.74**	**44.30**	24.19	40.16	2.98	16.86	1.57	2.26
Standard Malay	**41.63**	**63.46**	25.37	44.60	16.59	41.11	25.41	44.45
Tagalog	32.34	55.05	**47.88**	**67.97**	5.37	25.21	0.98	2.65
Urdu	22.40	41.17	**36.94**	36.94	2.22	15.99	1.80	2.82
Simp. Chinese	31.06	29.25	**33.54**	**53.47**	21.27	18.73	6.04	18.34
West, Eastern, and Central Asia								
Arabic (MSA)	24.93	40.80	40.38	54.30	5.29	19.80	**41.46**	**54.23**
Dari	**17.38**	**36.99**	3.23	15.52	1.33	5.95	1.27	2.64
Kurdish (Kurmanji)	17.38	36.87	**54.69**	**72.73**	0.38	3.08	53.81	71.48
Kurdish Sorani	1.97	2.90	**20.82**	**42.94**	0.30	4.17	2.27	9.82
Russian	30.05	48.18	**32.69**	32.69	13.70	29.91	2.85	15.01
Southern Pashto	**19.04**	35.10	16.30	**40.16**	0.25	5.00	2.39	5.22
Western Persian	27.39	44.50	**58.07**	**73.20**	5.04	20.05	56.47	71.55
Latin America								
Brazilian Portuguese	**46.99**	**64.65**	32.04	49.26	38.12	57.99	27.19	43.54
Latin American Spanish	**46.80**	**65.15**	5.63	22.84	36.75	57.10	1.89	3.14
Europe								
French	35.86	55.37	**37.22**	37.22	27.28	48.27	36.29	**52.68**
Mean	22.38	38.38	**32.40**	**48.40**	7.42	18.27	16.59	26.23
Std Dev	13.20	16.33	13.91	13.68	11.23	17.58	20.35	26.69
Wins (spBLEU)	9	—	**18**	—	0	—	4	—

The `UrgencyScenarios` Dataset

Due to the limited availability of human-translated multilingual data, we create our own dataset in English, automatically translating it into additional 29 languages. We use ChatGPT-4o-mini[7] to iteratively generate scenarios conveying 6 different levels of urgency.

We prompt to cover scenarios related to natural and climate disasters, manually selecting those who were different than the one already generated in previous iterations. This observation aligns with our assumptions that LLMs tend to generate similar content, conveying less variety of urgency (see discussion in Limitations).

[7] https://chatgpt.com

Table 2. Translation Quality ($X \rightarrow$ en) Across Geographic Groups. Bold indicates the highest spBLEU score per language. NLLB-1.2B shows robust performance in African and Southeast Asian languages, while Llama3.2-1B and X-ALMA exhibit significant failure modes in low-resource scripts.

Language	Aya101		NLLB-1.2B		Llama3.2-1B		X-ALMA	
	spBLEU	chrF2++	spBLEU	chrF2++	spBLEU	chrF2++	spBLEU	chrF2++
Africa								
Amharic	28.44	50.53	**36.75**	56.70	1.39	9.84	1.47	2.80
Ganda	27.47	44.60	**38.97**	57.70	5.60	18.39	1.51	14.50
Hausa	**30.62**	49.32	30.39	48.05	4.93	18.34	1.07	11.95
Kinyarwanda	24.69	45.48	**40.82**	61.37	3.39	16.10	22.83	38.43
Lingala	19.75	37.90	**37.40**	54.17	2.44	15.54	1.29	14.48
Nigerian Fulfulde	13.69	27.84	**45.84**	63.08	4.94	17.00	1.41	14.42
Somali	15.96	31.69	**32.00**	51.39	2.74	13.70	1.65	16.39
Swahili	**38.47**	58.46	19.58	34.78	15.68	35.73	1.45	13.73
Tigrinya	26.68	47.10	**32.94**	52.03	1.86	7.90	1.65	15.95
West Cent. Oromo	19.80	39.25	**54.82**	71.48	2.95	15.37	0.90	11.97
Southern and Eastern Asia								
Bengali	39.51	60.06	**51.13**	68.42	16.86	37.95	0.94	4.52
Burmese	29.69	51.55	**57.76**	73.17	0.61	4.52	0.76	2.51
Hindi	**44.28**	64.12	39.12	56.87	25.73	46.65	7.05	11.52
Indonesian	46.08	65.60	**56.56**	73.00	34.94	56.11	40.44	60.76
Marathi	**32.83**	55.36	31.43	49.74	12.79	32.93	2.15	6.80
Nepali	**42.53**	62.56	36.81	57.48	10.27	28.87	8.20	17.57
Standard Malay	**48.53**	67.25	43.81	63.15	30.24	50.72	39.86	56.92
Tagalog	54.51	70.52	**62.47**	75.73	28.37	46.23	0.61	2.39
Urdu	**33.26**	55.69	32.31	58.44	16.63	38.63	2.42	4.85
Simp. Chinese	25.45	51.76	**48.84**	64.79	13.06	36.67	14.35	37.06
West and Central Asia								
Arabic (MSA)	36.03	58.60	**46.31**	66.21	21.13	44.69	10.06	22.60
Dari	**34.39**	56.70	5.89	19.88	18.51	39.94	1.17	13.84
Kurdish (Kurmanji)	34.29	54.59	**54.56**	70.92	4.27	17.80	1.95	16.37
Kurdish Sorani	27.85	50.20	**36.29**	56.59	1.23	13.09	2.06	16.49
Russian	35.97	57.73	43.44	62.57	28.25	51.69	**53.51**	69.36
Southern Pashto	**33.70**	55.46	32.41	51.92	3.66	20.31	6.26	12.84
Western Persian	35.81	57.58	**57.37**	73.49	23.95	46.15	8.10	16.78
Latin America								
Brazilian Portuguese	**50.10**	68.99	43.44	62.68	43.85	63.64	0.64	6.09
Latin American Spanish	48.58	68.03	14.46	29.97	41.95	62.29	**53.06**	69.31
Europe								
French	39.49	59.09	**44.56**	63.60	34.52	54.56	40.99	58.30
Mean	33.36	53.69	**41.40**	**59.79**	14.15	31.52	10.03	19.34
Std Dev	9.94	10.97	12.59	13.08	13.11	18.06	16.14	19.08
Wins (spBLEU)	10	—	**19**	—	0	—	2	—

In total, we manually selected and curated 100 sentences, each of them describing a scenario with a level of urgency in the context of natural disaster. We translated this set into 29 medium and low-resource languages using a small version of a state-of-the-art MT model, `NLLB-200` (Team et al., 2022), creating a multilingual set of 3,000 scenarios; we refer to this dataset as `UrgencyScenarios`.

Annotation Setup First, we conduct a small-scale human study to understand how urgency perception fluctuates across languages in assessing crisis scenarios. We recruit two native speakers for each of the following four languages: Spanish, Greek, Bengali, and Hindi. Note that Greek is not part of the TICO-19 evaluation set used in §, so Figure 1 illustrates examples only from the three languages present in both studies (Spanish, Bengali, and

Hindi); the examples were selected among those showing the clearest label divergence between the LLM and human annotations. We focus on graduate students whose first language is the language of study and who professionally work in the United States, demonstrating a proficient level of English.

We split `UrgencyScenarios` into two, providing half set in English and the other half in the annotator's native language. In order to have the dataset fully annotated in all languages, we provide separate halves to each of the speakers of the same language, producing a single annotation per language per sentence.

We ask the annotators to classify each scenario choosing among six urgency categories: Non-critical, Very-Low, Low, Medium, High, Critical; along with an example for each one in a few-shot configuration (see prompt). In this way, we obtained a set of annotations per language, and four sets in English.

Then, we conduct LLM annotations using `Llama` 3.2 as a representative model. We provide the whole translated prompts, as was done with the human annotations. Due to the generative nature of the model's response, we capture model's urgency annotation for each scenario by analyzing the inclusion of the label provided in the prompt.[8]

Results and Analysis: Human Alignment and Urgency Case Study

RQ2. Human Alignment: Urgency Case Study To what extent does LLM performance on crisis-specific tasks align with human judgments? Are LLMs consistent across languages? How do language and cultural context influence this alignment?

Human annotation consistency. In response to **RQ2**, our results indicate that human participants, regardless of language or cultural background, maintain a high level of consensus on urgency assessments. In contrast, the LLM utilized as an annotator demonstrates significant inconsistency when classifying the same scenario across different languages (see Figure 2).

The human annotation distributions for the evaluated languages overlap significantly, and all groups concentrating on similar urgency levels (see Figure 2, left).

A detailed analysis of annotation shifts caused by translation (see Figure 3, left) reveals a high proportion of agreement for scenarios classified as "High" and "Low" urgency. Disagreements typically shift between adjacent classes; for example, 66.7% of scenarios classified as "Very Low" in English are annotated as "Low" when translated.

However, we also observe specific class transitions that severely alter the urgency assessment, such as a shift from "Critical" in English to "Not Urgent" in the target translation. We hypothesize that the translation quality of key terms associated with urgency, such as "immediately," which carries varying contextual weight, significantly impacts these assessments.

Overall, the harmonized distribution and robust agreement suggest that humans share a common understanding of urgency that transcends linguistic differences. We posit that urgency assessment is generally language-independent, despite minor variations.

Table 3. Cross-lingual urgency labels assigned by `Llama` 3.2 to the same gas leak scenario across six languages. Despite being *Critical* in English and Indonesian, the same text is classified as *Low* in Marathi and *Not Urgent* in Amharic. See Appendix for further examples.

Lang.	Translation	LLM Label
"A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed instantly."		
en	*A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed inst…*	Critical
es-LA	*Se ha detectado una importante fuga de gas en la zona residencial; equipos de evacuación y reparación deben…*	High
id	*Kebocoran gas besar telah terdeteksi di daerah perumahan; tim evakuasi dan perbaikan harus segera dikerah…*	Critical
mr	*निवासी भागात मोठ्या प्रमाणात गॅस गळती झाल्याचे आढळून आले आहे; स्थलांतर आणि दुरुस्ती पथके तातडीने तैनात करणे आवश्यक…*	Low
zh	*在住宅区发现大气泄漏, 必须立即派出疏散和维修小组.*	Medium
am	*በመኖሪያ አካባቢ አንድ ትልቅ የጋዝ ፍሳሽ ተገኝቷል፤ የማስወገጃና የጥገና ቡድኖች ወዲያውኑ መሰማራት አ…*	Not Urgent

[8] A further approach could semantically analyze the LLMs annotation, instead of simply pattern matching with the given label.

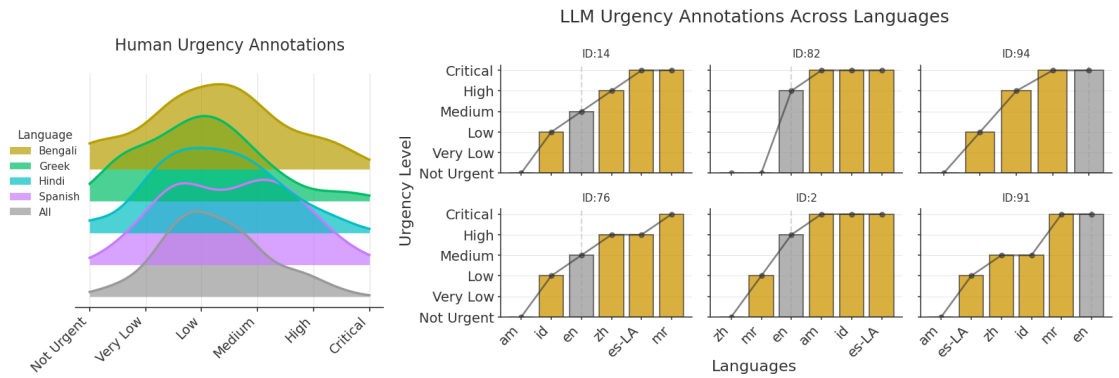

Figure 2. Distribution of urgency scores per language group across all annotated sentences, whether in English or the annotator's native language (left). Cross-lingual LLM urgency assessment for the same scenario — the model assigns different labels depending on the prompt language (right). While human annotators are overall consistent, the LLM shows substantial cross-lingual instability.

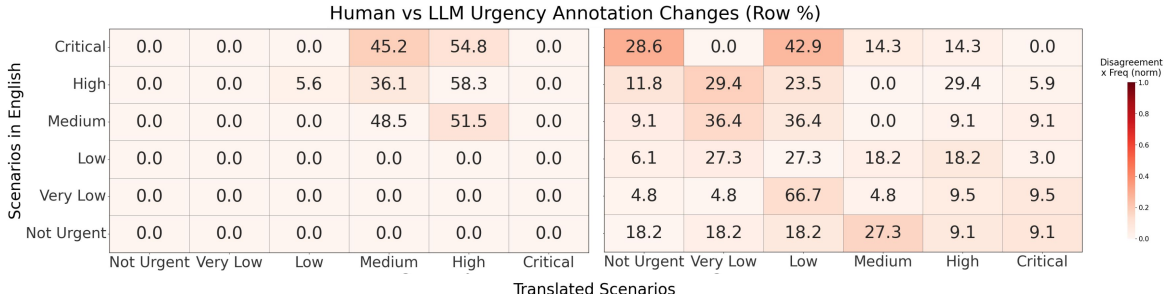

Figure 3. LLMs change their urgency assessment across our scenarios, largely due to translation quality.

LLM cross-lingual instability. Conversely, our findings reveal substantial cross-lingual instability in LLM-based annotations. Given an identical scenario, the LLM assigns disparate urgency levels based solely on the prompt language (see Figure 2, right). Table 3 illustrates this with a concrete example: a gas leak scenario classified as *Critical* in English is downgraded to *Low* in Marathi and *Not Urgent* in Amharic by the same model. Additional examples are provided in Appendix .

Furthermore, the LLM occasionally refuses to classify the input, hallucinates, or provides responses outside the provided categories. Out of the 30 languages evaluated, only 14 achieved a classification rate of at least 80% across all scenarios.

Within this subset, translation causes the model to systematically favor "Medium," "High," or "Critical" classifications, reflecting an overly cautious response strategy. This results in near-equal rates of agreement and disagreement: among scenarios classified as "Medium" in English, only 48.5% remain "Medium" upon translation, while 51.5% shift to "High." While such caution may appear to align with crisis safety, it remains unclear whether the model effectively grasps the underlying complexity of the scenario.

Implications for crisis triage. Our findings suggest that automated pipelines based on current state-of-the-art LLMs are not yet suitable for the automatic triage of real-life crisis scenarios. When queried in most languages, LLMs adopt cautious strategies that yield practically ineffective outputs for urgency classification. While a potential solution involves translating all inputs into English before querying the model, this approach remains vulnerable to cascading errors caused by mistranslations or subtle semantic shifts in urgency, as discussed in the preceding sections.

Limitations and Future Work

The findings of this study should be interpreted within the context of several inherent limitations. First, the `UrgencyScenarios` dataset is composed of synthetic scenarios rather than authentic crisis communications. While this approach allows for a controlled exploration of LLM performance on a simplified classification task, it may not fully capture the complex nuances and linguistic unpredictability characteristic of real-world emergencies. We

anticipate that real-world data would present a significantly greater challenge to both translation and classification pipelines than the task addressed in this work.

Additionally, the use of the distilled version of the `NLLB-200` model likely constrained the translation quality across the evaluated languages. Because this smaller architecture was used, it may have contributed to the performance degradation observed during the urgency assessment phase.

Similarly, this study utilized Llama 3.2 as the primary LLM for classification. Future research should therefore evaluate a broader spectrum of state-of-the-art LLMs to more precisely isolate how different models and varying levels of translation accuracy influence the variance in urgency perception.

Furthermore, a comprehensive analysis of the correlation between translation scores and urgency shifts was not feasible within the current scope. Such an analysis would require a larger volume of parallel data for languages that also possess human annotations to serve as a reliable reference point.

The profiles of our annotators also represent a limitation, as they are not professional crisis responders or subject matter experts. Although our results suggest a robust common understanding of urgency among general users, their perceptions might differ from those of trained crisis actors.

Finally, we recognize that a more rigorous data creation process in the future could involve crisis responders to develop semi-synthetic datasets. Despite these efforts toward better simulation, the ideal foundation for research in this domain remains the collection of authentic crisis data accompanied by high-quality human translations to serve as a definitive gold standard for computational modeling.

CONCLUSION

This work examined the suitability of state-of-the-art machine translation systems and LLMs for crisis communication, specifically focusing on the preservation and assessment of urgency. Our results demonstrate that current open-source models, including those explicitly designed for multilingual translation, exhibit substantial limitations in crisis-domain settings, particularly for low- and medium-resource languages. More critically, we show that translation and prompting choices can systematically distort perceived urgency, and that LLM-based urgency classification remains highly unstable across different linguistic contexts.

As a first approach in this domain, we acknowledge that this research serves as a foundational work-in-progress. The study is currently limited by its reliance on a single LLM architecture (Llama 3.2), the use of a distilled translation model (NLLB-200), and a synthetic dataset that, while useful for initial testing, may not capture the full complexity of real-world crisis data. Furthermore, the absence of expert crisis responders in the annotation process highlights the need for more specialized data curation in future iterations.

These findings raise serious concerns about the deployment of general-purpose language technologies in high-stakes crisis response pipelines without targeted evaluation. For the ISCRAM community, our work underscores the necessity for crisis-aware benchmarks, evaluation criteria, and model design practices that prioritize communicative intent and reliability.

Ensuring that automated systems preserve urgency is essential for building resilient, trustworthy socio-technical systems capable of supporting effective crisis preparedness and response. Finally, we recommend close collaboration between the AI development research community and the crisis informatics community to ensure that future computational progress directly addresses the real-world challenges faced by crisis responders.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under the award CIRC 2346334. The authors thank anonymous reviewers, whose suggestions helped improve this article, and Will Lewis, Fei Xia, and Haotian Zhu for their feedback.

REFERENCES

Abdul-Mageed, M., Elmadany, A., Nagoudi, E. M. B., Pabbi, D., Verma, K., & Lin, R. (2021, April). Mega-COV: A billion-scale dataset of 100+ languages for COVID-19. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 3402–3420). Association for Computational Linguistics.

Al Amer, S., Lee, M., & Smith, P. (2023, September). Cross-lingual Classification of Crisis-related Tweets Using Machine Translation. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 22–31). INCOMA Ltd., Shoumen, Bulgaria.

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., & Tur, S. (2020, December). TICO-19: The Translation Initiative for COvid-19. In K. Verspoor, K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea, & B. Wallace (Eds.), *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics.

Carpuat, M., Asscher, O., Bali, K., Bentivogli, L., Blain, F., Bowker, L., Choudhury, M., Daumé III, H., Duh, K., Gao, G., Grissom II, A., Karpinska, M., Khoong, E. C., Lewis, W. D., Martins, A. F. T., Nurminen, M., Oard, D. W., Popovic, M., Simard, M., & Yvon, F. (2025, November). An interdisciplinary approach to human-centered machine translation. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 22859–22879). Association for Computational Linguistics.

Casacuberta, F., Ceausu, A., Choukri, K., Deligiannis, M., Domingo, M., García-Martínez, M., Herranz, M., Jacquet, G., Papavassiliou, V., Piperidis, S., Prokopidis, P., Roussis, D., & Hadj Salah, M. (2024). Findings of a machine translation shared task focused on covid-19 related documents. *Proceedings of the Poster Sessions of the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, *3846*, 26–35.

Cavalin, P., Domingues, P. H., & Pinhanez, C. (2025). Sentence-level aggregation of lexical metrics correlates stronger with human judgements than corpus-level aggregation. *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, *39*, 23532–23540.

Coombs, W. T., & Holladay, S. J. (Eds.). (2010). *The handbook of crisis communication*. Blackwell Publishing Ltd.

D. Lewis, W., Zhu, H., Strawn, K., & Xia, F. (2025, July). Tapping into social media in crisis: A survey. In K. Atwell, L. Biester, A. Borah, D. Dementieva, O. Ignat, N. Kotonya, Z. Liu, R. Wan, S. Wilson, & J. Zhao (Eds.), *Proceedings of the fourth workshop on nlp for positive impact (nlp4pi)* (pp. 306–331). Association for Computational Linguistics.

De Gibert, O., Pugh, R., Marashian, A., Vazquez, R., Ebrahimi, A., Denisov, P., Rice, E., Gow-Smith, E., Prieto, J., Robles, M., Manrique, R., Moreno, O., Lino, A., Coto-Solano, R., Alvarez, A., Agüero-Torales, M., Ortega, J. E., Chiruzzo, L., Oncevay, A., … Mager, M. (2025, May). Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In M. Mager, A. Ebrahimi, R. Pugh, S. Rijhwani, K. Von Der Wense, L. Chiruzzo, R. Coto-Solano, & A. Oncevay (Eds.), *Proceedings of the fifth workshop on nlp for indigenous languages of the americas (americasnlp)* (pp. 134–152). Association for Computational Linguistics.

Forcada, M. L., & Tyers, F. M. (2016). Apertium: A free/open source platform for machine translation and basic language technology. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*.

Global CLEAR. (2016). Translators without Borders develops world's first crisis-specific machine translation for Kurdish.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation (B. Roark & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, *10*, 522–538.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024, November 23). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs].

Imran, M., Qazi, U., & Ofli, F. (2022). TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. *Data*, *7*(1), 8.

Khullar, M., Desai, U., Malviya, P., Dalmia, A., & Shi, Z. R. (2025). Script Gap: Evaluating LLM Triage on Indian Languages in Native vs Roman Scripts in a Real World Setting.

Kocmi, T., Artemova, E., Avramidis, E., Bawden, R., Bojar, O., Dranch, K., Dvorkovich, A., Dukanov, S., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Lakougna, H., Lundin, J., Monz, C., Murray, K., … Zouhar, V. (2025, November). Findings of the WMT25 general machine

translation shared task: Time to stop evaluating on easy test sets. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the tenth conference on machine translation* (pp. 355–413). Association for Computational Linguistics.

Lankford, S., & Way, A. (2024, September). Leveraging LLMs for MT in Crisis Scenarios: A blueprint for low-resource languages. In R. Knowles, A. Eriguchi, & S. Goel (Eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (pp. 4–13). Association for Machine Translation in the Americas.

Lankford, S., Afli, H., & Way, A. (2024). Machine Translation in the Covid domain: An English-Irish case study for LoResMT 2021.

Lee, S., Jung, S., Park, J.-H., Cho, H., Moon, S., & Ahn, S. (2025). Performance of ChatGPT, Gemini and DeepSeek for non-critical triage support using real-world conversations in emergency department. *BMC Emergency Medicine*, *25*(1), 176.

Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q. R., & Li, J. (2025, July). Harnessing large language models for disaster management: A survey. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: Acl 2025* (pp. 14528–14551). Association for Computational Linguistics.

Lerbinger, O. (2012). *The crisis manager: Facing disasters, conflicts, and failures* (2nd). Routledge.

Lewis, W. (2010, May). Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In F. Yvon & V. Hansen (Eds.), *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Li, S., Wang, J., Ali, F. D. M. A., Cherry, C., Deutsch, D., Briakou, E., Sousa-Silva, R., Lopes Cardoso, H., Stenetorp, P., & Adelani, D. I. (2025, November). SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 12979–12998). Association for Computational Linguistics.

Masanneck, L., Schmidt, L., Seifert, A., Kölsche, T., Huntemann, N., Jansen, R., Mehsin, M., Bernhard, M., Meuth, S. G., Böhm, L., & Pawlitzki, M. (2024). Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. *Journal of Medical Internet Research*, *26*, e53297.

Merx, R., Suominen, H., Cohn, T., & Vylomova, E. (2025, November). OpenWHO: A document-level parallel corpus for health translation in low-resource languages. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the tenth conference on machine translation* (pp. 142–160). Association for Computational Linguistics.

Odubola, O., Adeyemi, T. S., Olajuwon, O. O., Iduwe, N. P., Inyang, A. A., & Odubola, T. (2025). AI in Social Good: LLM powered Interventions in Crisis Management and Disaster Response. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *3*(1), 2353–2360.

Ojo, J., Ogundepo, O., Oladipo, A., Ogueji, K., Lin, J., Stenetorp, P., & Adelani, D. I. (2025). Afrobench: How good are large language models on african languages?

Pakray, P., Krishna, R., Pal, S., Vetagiri, A., Dash, S., Maji, A. K., Lyngdoh, S. A., Laitonjam, L., Jamatia, A., Sambyo, K., Das, A., & Manna, R. (2025, November). Findings of WMT 2025 shared task on low-resource Indic languages translation. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the tenth conference on machine translation* (pp. 532–553). Association for Computational Linguistics.

Paraskevas, A. (2013). Mitroff's five stages of crisis management. In K. B. Penuel, M. Statler, & R. Hagen (Eds.), *Encyclopedia of crisis management* (pp. 629–632, Vol. 2). SAGE Publications, Inc.

Popović, M. (2017, September). ChrF++: Words helping character n-grams. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the second conference on machine translation* (pp. 612–618). Association for Computational Linguistics.

Post, M. (2018, October). A call for clarity in reporting BLEU scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the third conference on machine translation: Research papers* (pp. 186–191). Association for Computational Linguistics.

Roussis, D. G. (n.d.). Building End-to-End Neural Machine Translation Systems for Crisis Scenarios: The Case of COVID-19.

Sánchez, C., Abeliuk, A., & Poblete, B. (2025). Large Language Models in Crisis Informatics for Zero and Few-Shot Classification. *ACM Trans. Web*, *19*(4), 45:1–45:25.

Strassel, S., & Tracey, J. (2016, May). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3273–3280). European Language Resources Association (ELRA).

Team, N. L. L. B., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., … Wang, J. (2022, August 25). *No Language Left Behind: Scaling Human-Centered Machine Translation*. arXiv: 2207.04672 [cs].

Tracey, J., Strassel, S., Bies, A., Song, Z., Arrigo, M., Griffitt, K., Delgado, D., Graff, D., Kulick, S., Mott, J., & Kuster, N. (2019, August). Corpus Building for Low Resource Languages in the DARPA LORELEI Program. In A. Karakanta, A. K. Ojha, C.-H. Liu, J. Washington, N. Oco, S. M. Lakew, V. Malykh, & X. Zhao (Eds.), *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages* (pp. 48–55). European Association for Machine Translation.

Üstün, A., Aryabumi, V., Yong, Z., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., & Hooker, S. (2024, August). Aya model: An instruction finetuned open-access multilingual language model. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 15894–15939). Association for Computational Linguistics.

Way, A., Haque, R., Xie, G., Gaspari, F., Popović, M., & Poncelas, A. (2020). Rapid Development of Competitive Translation Engines for Access to Multilingual COVID-19 Information. *Informatics*, *7*(2), 19.

Xu, F., Ma, J., Li, N., & Cheng, J. C. (2025). Large language model applications in disaster management: An interdisciplinary review. *International Journal of Disaster Risk Reduction*, *127*, 105642.

Xu, H., Murray, K., Koehn, P., Hoang, H., Eriguchi, A., & Khayrallah, H. (2025, March). X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale.

Zebaze, A. R., Sagot, B., & Bawden, R. (2025, April). In-context example selection via similarity search improves low-resource machine translation. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the association for computational linguistics: Naacl 2025* (pp. 1222–1252). Association for Computational Linguistics.

Zouhar, V., Chen, P., Lam, T. K., Moghe, N., & Haddow, B. (2024, November). Pitfalls and outlooks in using COMET. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the ninth conference on machine translation* (pp. 1272–1288). Association for Computational Linguistics.

ANNOTATION PROMPT

We explore different prompt designs to describe the task and each of the urgency categories, trying out different punctuation marks to signal the classes.

Listing 1. Urgency Classification Prompt

```
Classify the following sentence into one of the 6 levels of urgency in the
    context of a crisis: Critical, High, Medium, Low, Very Low, Not Urgent.

To help you understand the task, here are some examples for each category in the
     context of covid 19:

Critical: 'All non-essential movement must cease immediately, a sund all public
    gatherings are strictly prohibited due to a critical rise in cases
    overwhelming healthcare capacity.'
High: 'Due to a significant surge in cases, it is now mandatory to wear a
    medical-grade mask in all public indoor spaces, effective immediately.'
Medium: 'We urge citizens to maintain a physical distance of at least one meter
    from others in public spaces to mitigate ongoing transmission risks.'
Low: "We encourage all eligible individuals to receive their COVID-19 booster
    shots to maintain optimal protection against evolving variants."
Very Low: 'Consider wearing a face mask in crowded indoor settings to further
    reduce your risk of exposure to respiratory viruses.'
Not Urgent: 'Maintaining good hand hygiene, including frequent washing with soap
     and water, remains a cornerstone of public health, extending beyond the
    current pandemic.'

Now, classify this sentence in the context of a crisis:

[SENTENCE]

What is the level of urgency for this sentence?
```

CROSS-LINGUAL URGENCY DIVERGENCE EXAMPLES

Table 4 shows four crisis scenarios selected for their highest label entropy across six languages. Each scenario was translated using `NLLB-200` and then classified by `Llama` 3.2. The labels illustrate the breadth of divergence: a single scenario can span the full range from *Not Urgent* to *Critical* depending solely on the prompt language. *N/C* denotes cases where the model refused to classify or produced output outside the label schema.

Table 4. Cross-lingual urgency divergence for four crisis scenarios ranked by label entropy. `Llama` 3.2 assigns labels ranging from *Not Urgent* to *Critical* for identical text depending solely on the prompt language. *N/C* = non-classified.

Lang.	Translation	LLM Label
"Strong winds could cause minor damage to property. We request assistance with securing loose objects and trimming trees."		
en	*Strong winds could cause minor damage to property. We request assistance with securing loose objects and tri…*	High
es-LA	*Los vientos fuertes pueden causar pequeños daños a la propiedad, pedimos ayuda para asegurar objetos suelt…*	Critical
id	*Angin kencang bisa menyebabkan kerusakan kecil pada properti kami meminta bantuan untuk mengamankan …*	Low
mr	*जोरदार वाऱ्यामुळे मालमत्तेचे किरकोळ नुकसान होऊ शकते.*	Medium
zh	*强风可能会造成轻微的财产损害. 我们要求帮助安装松散的物品和修剪树木.*	Not Urgent
am	*ታይለኛ ነፋሳት በንብረት ላይ ትንሽ ጉዳት ሊያስከትሉ ይችላሉ።*	N/C
"A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed instantly."		
en	*A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed inst…*	Critical
es-LA	*Se ha detectado una importante fuga de gas en la zona residencial; equipos de evacuación y reparación deben…*	High
id	*Kebocoran gas besar telah terdeteksi di daerah perumahan; tim evakuasi dan perbaikan harus segera dikerah…*	Critical
mr	*निवासी भागात मोठ्या प्रमाणात गॅस गळती झाल्याचे आढळून आले आहे; स्थलांतर आणि दुरुस्ती पथके तातडीने तैनात करणे आवश्यक…*	Low
zh	*在住宅区发现大气泄漏, 必须立即派出疏散和维修小组.*	Medium
am	*በመኖሪያ አካባቢ አንድ ትልቅ የጋዝ ፍሳሽ ተገኝቷል፤ የማስወገጃና የጥገና ቡድኖች ወዲያውኑ መሰማራት አ…*	Not Urgent
"Rivers are rising rapidly; open community shelters and coordinate transportation for seniors and people with disabilities."		
en	*Rivers are rising rapidly; open community shelters and coordinate transportation for seniors and people with …*	Medium
es-LA	*Los ríos están subiendo rápidamente; abriendo refugios comunitarios y coordinando el transporte para perso…*	High
id	*Sungai naik dengan cepat; buka tempat penampungan masyarakat dan koordinasi transportasi untuk lansia d…*	Low
mr	*नद्यांची पातळी झपाट्याने वाढत आहे; ज्येष्ठ नागरिक आणि अपंग लोकांसाठी सामुदायिक आश्रयस्थाने उघडा आणि वाहतूक समन्वित …*	Critical
zh	*河流正在迅速上升; 开放社区庇护所, 协调老人和残疾人的交通.*	High
am	*ወንዞች በፍጥነት እየጨመሩ ናቸው፤ የማህበረሰብ መጠለያዎችን ይክፈቱ እና ለአረጋውያን እና ለአካል ጉዳተ…*	Not Urgent
"A tornado has touched down in the city center, causing widespread destruction — search and rescue teams are critically needed."		
en	*A tornado has touched down in the city center, causing widespread destruction — search and rescue teams are…*	Critical
es-LA	*Un tornado ha tocado el centro de la ciudad, causando destrucción generalizada. Se necesitan equipos de bús…*	Low
id	*Sebuah tornado telah mendarat di pusat kota, menyebabkan kerusakan luas. Tim pencarian dan penyelamatan…*	Medium
mr	*शहराच्या मध्यभागी वादळाचा धक्का बसला आहे, ज्यामुळे मोठ्या प्रमाणात विध्वंस झाला आहे. शोध आणि बचाव पथकांची अत्यंत ग…*	Critical
zh	*一场龙卷风袭击了市中心,造成大规模破坏,急需搜救队伍.*	Medium
am	*አውሎ ነፋስ በከተማዋ መሃል ላይ ደርሶ ሰፊ ውድመት አስከትሏል፤ የፍለጋና የማዳን ቡድኖች በአስቸኳይ ያስ…*	Not Urgent