Translating Under Pressure: Domain-Aware LLMs for Crisis Communication

Antonio Castaldo
University of Pisa
University of Naples "L'Orientale"
antonio.castaldo@phd.unipi.it

Maria Carmen Staiano
University of Macerata
m.staiano@unimc.it

Johanna Monti
University of Naples "L'Orientale"
jmonti@unior.it

Sheila Castilho
Dublin City University
sheila.castilho@dcu.ie

Francesca Chiusaroli
University of Macerata
f.chiusaroli@unimc.it

Abstract

Timely and reliable multilingual communication is critical during natural and human-induced disasters, but developing effective solutions for crisis communication is limited by the scarcity of curated parallel data. We propose a domain-adaptive pipeline that expands a small reference corpus, by retrieving and filtering data from general corpora. We use the resulting dataset to fine-tune a small language model for crisis-domain translation and then apply preference optimization to bias outputs toward CEFR A2-level English. Automatic and human evaluation shows that this approach improves readability, while maintaining strong adequacy. Our results indicate that simplified English, combined with domain adaptation, can function as a practical lingua franca for emergency communication when full multilingual coverage is not feasible.

1 Introduction

Global crises such as the COVID-19 pandemic, wildfires, or earthquakes require the rapid, trustworthy, and reliable dissemination of public information (Piller et al., 2020; Hajek et al., 2024). In linguistically diverse contexts such as Italy, where international residents and visitors may have limited proficiency in Italian, ensuring access to emergency communications becomes a matter of public safety. When full multilingual coverage across all community languages is not feasible, English plays a pivotal role as a contact language for cross-cultural communication (Jenkins and Mauranen, 2019; Seidlhofer, 2011) and is already attested in interactions between migrants, institutions and interpreters (Amato and Cirillo, 2024). We therefore argue that simplified English can serve as a practical lingua franca, enabling broader comprehension and supporting timely action among individuals with different linguistic backgrounds (Musacchio et al., 2017; O'Brien and Federici, 2020; Radicioni and Rosendo, 2022).

In this context, machine translation (MT) systems play a critical role in assisting emergency response teams and public authorities in disseminating information efficiently (Lewis, 2010). The emergence of Large Language Models (LLMs) has further expanded the potential of MT in crisis settings (Lankford and Way, 2024). Their adaptability to specific domains and communicative requirements makes them promising tools for producing emergency messages tailored to diverse audiences.

However, effective risk communication is not solely a matter of semantic accuracy. Emergency messages must be actionable, terminologically precise, and easily understandable, particularly for non-native readers. This requires careful handling of domain-specific terminology along-

side explicit attention to readability. Achieving this balance depends on carefully curated parallel corpora that support reliable domain adaptation (Moorkens et al., 2025).

Within the Italian context, existing Institution-to-User communication resources (Torresi, 2020) provide valuable resources, but remain limited in size. In this study, we adopt ITALERT as our reference corpus (Staiano et al., 2025), and use it to retrieve relevant parallel data from general corpora. ITALERT is an English-Italian corpus that encompasses major natural disasters in Italy, and includes high-quality parallel data, making it a particularly suitable resource for our study.

In this work, we propose a domain-adaptive pipeline for generating simplified English translations of emergency messages from the Italian Civil Protection Department. We introduce a two-step methodology to retrieve and classify relevant in-domain data in the Italian-English language pair from general corpora. Using the ITALERT corpus as a reference, we employ cluster-based similarity search to construct a crisis-relevant parallel dataset. We then fine-tune a small language model on this curated corpus and further apply preference optimization to bias its outputs toward CEFR A2-level English, aiming to produce translations that are both reliable and accessible for linguistically diverse populations in Italy.

As CEFR A2-level English is characterized by the use of high-frequency vocabulary and simple syntactic structures (North and Piccardo, 2020), we demonstrate that optimizing the model towards this proficiency level increases readability and maximizes the potential audience of the translated messages, while only incurring in a partial decrease of translation quality.

2 Related Work

Prior work has shown that the limited availability of appropriate in-domain datasets represents a major challenge when deploying MT in crisis settings (Cadwell et al., 2019; O'Brien, 2022; Moorkens et al., 2025), particularly with regard to context-sensitive translations, terminology consistency, and register accuracy. To address this problem, we elaborated a two-stage pipeline to augment crisis-domain data for a high resource language combination (Italian-English). In this context, datasets related to the crisis domain are limited, especially when taking into account information from public sources, such as government agencies. The existing parallel datasets available for our specific language combination are mostly related to public-health emergencies like COVID-19 (Anastasopoulos et al., 2020; Way et al., 2020).

A strategy to mitigate data scarcity in specialized domains is to apply data retrieval techniques to mine relevant parallel segments from large general-domain corpora. Previous work has shown that semantic vector similarity, embedding-based retrieval, and classifier-driven filtering can effectively extract in-domain data for MT adaptation (Espana-Bonet et al., 2017; Sugathadasa et al., 2017; Glavaš et al., 2018).

Research on crisis-oriented corpora has highlighted the need for structured annotation frameworks to distinguish disaster-related content from general communication (Alam et al., 2018). The taxonomy introduced by the United Nations Office for Disaster Risk Reduction and the International Science Council (2025) provides an authoritative reference for hazard classification, organizing 281 hazards into eight clusters: Meteorological and Hydrological, Extraterrestrial, Geological, Environmental, Chemical, Biological, Technological, and Societal.

In this study, we use this taxonomy as a reference framework to filter in-domain sentences from general corpora, as seen in Section 3. Domain relevance is determined not only by explicit references to hazards but also by the presence of terminology commonly used in crisis management.

Research in crisis communication has also emphasized the importance of message accessibility and readability when disseminating emergency information to multilingual populations. Prior studies have shown that comprehension plays a central role in shaping public trust and behavioral compliance during crises. In particular, Rossetti et al. (2020) demonstrate that clearer crisis messages contribute to higher levels of trust toward emergency authorities and increase the likelihood that individuals will follow preparedness and response instructions. This line of research motivates our adoption of simplification-informed model adaptation, aimed at producing translations that are not only accurate but also comprehensible for broad and linguistically diverse audiences.

Despite these advances, to the best of our knowledge, no existing work combines semantic retrieval, disaster-informed annotation, and

classifier-based filtering into a unified pipeline for crisis-domain MT.

Taken together, previous research highlights two main gaps:

(1) the lack of Italian–English crisis-domain parallel data, and (2) the absence of combined retrieval–annotation–classification approaches for corpus augmentation.

We address these limitations by proposing a structured end-to-end pipeline that expands crisis-domain data, fine-tunes a domain-specific MT model, and evaluates translation quality through both automatic and human-centered assessments, with the goal of improving MT reliability in crisis communication.

3 Data Collection

Obtaining sufficient in-domain parallel data for risk communication remains a challenge, where communications are generally produced into monolingual corpora and rarely compiled into parallel MT resources. We address data scarcity with a two-stage approach: first assembling and expanding a small reference corpus of authentic crisis communication, starting from ITALERT (Staiano et al., 2025), then using it to extract in-domain sentences from large general corpora, using embedding-based similarity. Our data selection pipeline is described in Figure 1.

3.1 Reference Corpus

Our reference corpus builds upon the dataset introduced by Staiano et al. (2025). To ensure high data quality, the corpus underwent a rigorous post-processing pipeline. We applied exact and MinHash deduplication to eliminate redundancy and filtered for a minimum length (> 8 words) to ensure sufficient context.

This filtering was necessary to minimize semantic noise and ambiguity in the vector space. By retaining only linguistically robust examples, we ensured that the reference data formed distinct, high-density clusters, which is a prerequisite for the data selection of the following stage.

3.2 General Corpora

To expand our training resources beyond the limited reference corpus, we utilized the OPUS collection (Tiedemann and Nygaard, 2004), specifically targeting the English-Italian language pair. We selected four sub-corpora, described in Table 1. These sources consist of parallel sentences with human or post-edited translations, and we selected them on the hypothesis that they contain high-quality crisis examples embedded within broader discourse. Although generic, these sources cover distinct facets of crisis communication, from natural disasters to legislative discussions on civil protection and disaster relief. This diversity ensures that our extraction pipeline can recover a wide spectrum of in-domain topics, from urgent alerts to procedural descriptions.

Before retrieval, the general corpora underwent rigorous preprocessing. We removed exact and fuzzy duplicates, using the MinHash deduplication algorithm. Sentences shorter than eight words were filtered out, and malformed or incomplete segments were excluded through dependency parsing with SpaCy (Honnibal et al., 2020). Finally, we made sure the target corpora only contained sentences in our language pair, using langdetect.

Source	Raw	Clean
EuroParl	2,128,356	1,890,114
Wikimedia	1,167,437	795,981
Wikipedia	1,000,951	758,228
GlobalVoices	147,829	116,360
ELRC-CORDIS	123,991	114,063
NewsCommentary	98,992	90,477
Reference	555	498

Table 1: Data statistics showing the reduction from raw data to the final dataset used for retrieval. The cleaning pipeline included length filtering, deduplication, and linguistic filtering.

3.3 Extraction Method

We extract domain-relevant sentences using embedding similarity with multiple centroids, following the intuition that crisis communication spans distinct registers that may occupy different regions of the embedding space.

Embedding and Clustering. We encode both the reference and general corpus using paraphrase-multilingual-MiniLM-L12-v2, a multilingual sentence encoder (Reimers and Gurevych, 2019). Since the reference corpus was designed to cover five main crisis scenarios, we cluster the 498 reference embeddings into k=5 clusters using k-means, generating five centroids that represent

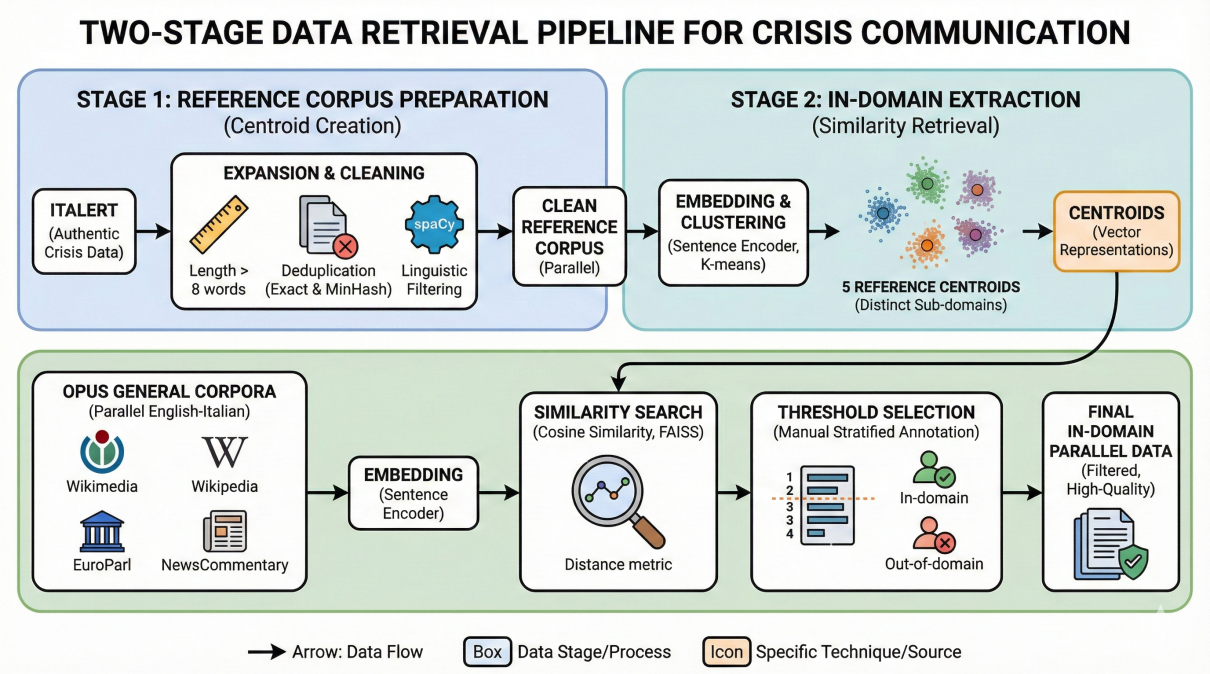

Figure 1: Overview of our two-stage data retrieval pipeline. **Stage 1** focuses on cleaning and clustering the reference corpus to generate distinct semantic centroids. **Stage 2** leverages these centroids to retrieve in-domain sentences from general corpora (OPUS) via embedding similarity, validated by a stratified manual annotation.

sub-domains within crisis communication.

Similarity Search. For each sentence in the general corpus, we compute its cosine similarity to all five reference centroids and retain the maximum, allowing candidate segments to be matched to the most relevant crisis profile, and improving coverage of the resulting corpus. We use FAISS (Douze et al., 2024) for efficient retrieval, extracting the top 50,000 most similar sentences from the general corpus, as candidates.

Stratified Annotation. To determine the optimal similarity cutoff, two annotators ranked the candidate sentences by their maximum centroid similarity and divided the list into six partitions. They then performed a stratified manual evaluation by randomly sampling and annotating 50 sentences from each partition as either in-domain or out-of-domain. The annotation was based on the hazard classification taxonomy established by the United Nations Office for Disaster Risk Reduction and the International Science Council (2025) and final decisions were based on consensus.

Threshold Selection. Finally, we analyzed the domain relevance distribution across these partitions and established the final filtering threshold at the point where the proportion of out-of-domain sentences first exceeded that of in-domain examples. We retain some out-of-domain sentences in the upper ranks, as we posit that these borderline examples are beneficial. While they may not strictly describe a crisis event, both the human inspection and their proximity in the embedding space indicate that these sentences share register with the target domain. For this reason, we consider a soft domain boundary where these sentences are included and considered valuable examples for model training. As a result of the process, we retained 36,000 segments that we use for SFT training, as described in Section 4.

4 Training

We perform Supervised Fine-tuning (SFT) on a **Tower-Plus-2B** model using parameter-efficient adaptation with QLoRA (Hu et al., 2021). We intentionally opt for a compact language model rather than a larger LLM, as prior work has shown that for high-resource language pairs, domain adaptation yields diminishing returns beyond moderate model sizes when high-quality in-domain data is available (Pang et al., 2024; Vieira et al., 2024).

In crisis-response settings, smaller models are also better aligned with practical deployment constraints, including limited computational resources and low-latency requirements faced by

public authorities and humanitarian organizations (Moorkens et al., 2025). Moreover, compact models tend to exhibit more controlled generation behavior (Sun et al., 2023), reducing verbosity and the risk of hallucinations (Zhou et al., 2024), which is particularly important in safety-critical communication where clarity and precision directly affect user comprehension and actionability.

4.1 Paragraph Construction

The model is fine-tuned on the crisis-domain corpus obtained through the retrieval pipeline described in Section 3. To better approximate the structure of authentic emergency communications, which typically consist of short informational blocks rather than isolated sentences, we adopt a paragraph-level training strategy.

Contextually similar sentence pairs are grouped into short paragraphs, generating approximately one paragraph every ten segments. Sentences within a paragraph vary in length and syntactic structure but are semantically coherent. The resulting training data therefore consists of a mixture of sentence-level and paragraph-level parallel examples, enabling the model to generalize across different granularity levels at inference time (Stap et al., 2024).

4.2 Readability-Oriented DPO

While translation accuracy is important, the actionability (Coche et al., 2021) of crisis communications heavily depends on how comprehensible they are, when the target audience consists of non-native English speakers. To meet this requirement, we further optimize the model using Direct Preference Optimization (Rafailov et al., 2024) on a corpus of translations simplified to an English A2 proficiency level, as per the CEFR.

Preference pairs are constructed using synthetic simplified translations generated by `gpt-5-mini`, using a text simplification prompt used in relevant prior work (Barbu et al., 2025) and that we report in Appendix A.

These simplifications aim to preserve semantic content while reducing lexical complexity and syntactic density, in line with established guidelines for emergency communication targeting linguistically diverse populations (Federici et al., 2019). To control for risks introduced by synthetic supervision, the simplified outputs undergo automatic quality checks using readability and adequacy metrics. In addition, a stratified sample is evaluated by human annotators to verify that simplification does not introduce semantic distortion or omit safety-critical information in the preference training data.

The resulting preference data biases the model toward translations that are both accurate and accessible, aligning optimization with the needs of L2 English readers in high-stakes emergency contexts.

Metric	Baseline	Fine-tuned
Flesch Reading Ease ↑	30.21	**45.04**
Flesch–Kincaid Grade ↓	15.56	**12.65**
SMOG Index ↓	16.14	**13.66**
Coleman–Liau Index ↓	13.32	**11.33**
ARI ↓	17.12	**13.94**
Dale–Chall ↓	11.31	**10.13**

Table 2: Automatic evaluation on 1,000 test sentences, measured by readability metrics. Arrows indicate when higher or lower values are better.

Automatic Evaluation. Table 2 reports readability evaluation results on 1,000 test examples. The DPO model consistently improves all readability metrics, with a +14.8 increase in Flesch Reading Ease and substantial reductions in other indicators (Flesch-Kincaid, SMOG, ARI). These results indicate that the model optimization successfully reduces lexical and syntactic complexity, producing translations that are much easier to read for L2 English speakers.

In terms of quality metrics, we find in Table 3 that the fine-tuned model optimized with DPO yields lower BLEU (Papineni et al., 2002) and chrF (Popović, 2015) scores than the baseline, reflecting reduced surface similarity with the reference translations due to lexical simplification and paraphrasing. This is expected as the model was optimized to generate translations close to an English A2 proficiency level. While BLEU and chrF exhibit substantial drops, the actual trade-off remains comparatively small, as evidenced by COMET, which reports only a modest decrease (-6.8%), and by the following human evaluation.

4.3 Ablation Study

To determine whether SFT is required to achieve both translation quality and readability improvements, we conduct an ablation study comparing DPO applied with and without prior domain adaptation.

We evaluate four configurations: (i) the base model without adaptation; (ii) the base model optimized directly with DPO; (iii) a model fine-tuned on the crisis-domain corpus using supervised learning only; and (iv) the full pipeline combining supervised fine-tuning and preference optimization.

Our findings, displayed in Table 3 demonstrate that applying DPO directly to the base model yields large gains in readability but is accompanied by substantial drops across BLEU, chrF, and COMET, whereas applying DPO after supervised fine-tuning results in comparable readability improvements with small reductions in COMET.

Training Configuration	BLEU	chrF	COMET	FRE
Base	0.41	66.62	0.88	29.44
SFT	0.39	64.35	0.87	32.49
Base + DPO	0.07	31.10	0.73	**74.40**
Optimal System				
SFT + DPO	**0.21**	**47.40**	**0.82**	**46.13**

Table 3: Overall performance and ablation study across four training configurations. The SFT+DPO model represents our final system. We report BLEU, chrF and COMET as quality metrics, and readability measured by Flesch Reading Ease (FRE).

5 Evaluation

To complement the quantitative results and better understand the practical implications of readability-oriented MT, we conduct a manual error analysis following the MQM framework (Lommel and Melby, 2018), combined with Direct Assessment (Graham et al., 2015), using the platform Pearmut[1] (Zouhar and Kocmi, 2026). The two annotators were native speakers of Italian, proficient in English, and with domain-specific expertise in crisis communication. The full annotators profile is available in Appendix C. They applied the core set of MQM categories: accuracy, fluency, style, locale conventions, and verity, along with their respective subcategories. Errors were rated using four severity levels: trivial, minor, major, and critical, corresponding to weights of 0, 1, 5, and 25, respectively. In total, 250 segments were evaluated with the SFT model and the readability-oriented model respectively, resulting in 500 overall segments.

We note that our use of MQM follows a relatively strict error taxonomy, which leads to a high number of annotated errors, especially for outputs that deviate from more literal translation behavior. As such, MQM should be interpreted here as a fine-grained linguistic error analysis rather than as a direct proxy for overall user acceptability.

After annotating an initial set of 100 segments, inter-annotator agreement (IAA) was calculated to ensure the reliability of the annotations. The initial agreement, measured with Cohen's Kappa, was equal to $K = 0.69$ in identifying the most common error category, which was accuracy. The mean absolute difference between the scores assigned by the two annotators was equal to 7.5. Most disagreements concerned the MQM categorization for simplification-related phenomena, particularly the distinction between Omission and Undertranslation. Following this preliminary phase, the annotators jointly reviewed the disputed cases, clarified the annotation guidelines, and reached consensus on the appropriate error labels. The remaining 400 translation segments were then annotated independently.

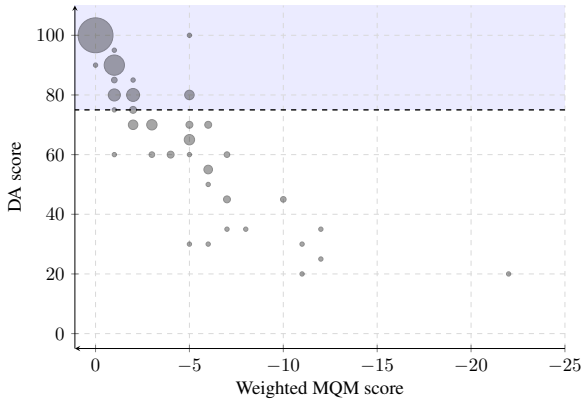

Figure 2: Relationship between weighted MQM score and DA score for the DPO model. Bubble size reflects the frequency of the segments. The shaded region (DA $\geq$ 75) highlights translations judged high quality by DA despite MQM penalties.

Results. Our results find that the fine-tuned model, optimized for readability (SFT+DPO), produces mostly accurate translations with a mean score of 83 points, compared to 95 for SFT. The translations produced by the DPO model were judged mostly acceptable, despite showing a higher number of minor and major errors compared to the SFT model. Particularly, 213 errors were annotated for the DPO model, and only 56 for the SFT one. However, 161 errors out of those 213 were considered minor in severity. This is substantiated by the relatively high DA scores assigned by

[1]https://github.com/zouharvi/pearmut

the annotators.

Figure 2 further illustrates the relationship between MQM and DA for the DPO model. Most segments are found in the upper-left region of the plot, indicating that many translations received high DA scores despite incurring minor MQM penalties. In particular, the shaded area (DA $\geq$ 70) contains 113 segments, showing that a large proportion of the DPO outputs (56.5%) were still judged high quality by human evaluators even when MQM identified minor issues. At the same time, the plot suggests a general correlation between the two measures: segments with heavily penalized MQM scores tend to receive lower DA scores. However, this tendency is not absolute, as several translations with moderately penalized MQM scores still achieve relatively high DA. This supports the view that MQM and DA capture overlapping but not identical aspects of translation quality. In particular, the minor MQM errors in these cases appear to be associated primarily with fluency, while DA traditionally evaluates how adequately the meaning the source sentence is expressed in the target (Bojar et al., 2017; Graham et al., 2015).

In the context of emergency communications, aimed at L2 English speakers, accuracy is arguably the most important quality dimension. The fact that many DPO translations remain highly rated in DA despite minor MQM errors therefore suggests that the model performs strongly for the communicative purpose of this task.

Furthermore, while SFT often outperforms DPO, over half of the evaluated segments received identical scores, indicating that DPO frequently produces translations of comparable quality despite its higher rate of errors, most of which are linked to simplification strategies. Although the gap in mean scores suggests systematic differences between the models, both systems consistently achieve scores in the upper half of the quality scale, indicating that overall translation quality remains high even when MQM annotations reveal a greater number of errors.

Consistently with the expected behavior of a model optimized for readability, most errors of the DPO model were found in the categories of Omission and Undertranslation, respectively with 43 and 74 annotated errors. For the SFT model, the most common error categories were Omission and Mistranslation. We include a detailed overview of the annotated errors for both models in Table 5 of the Appendix.

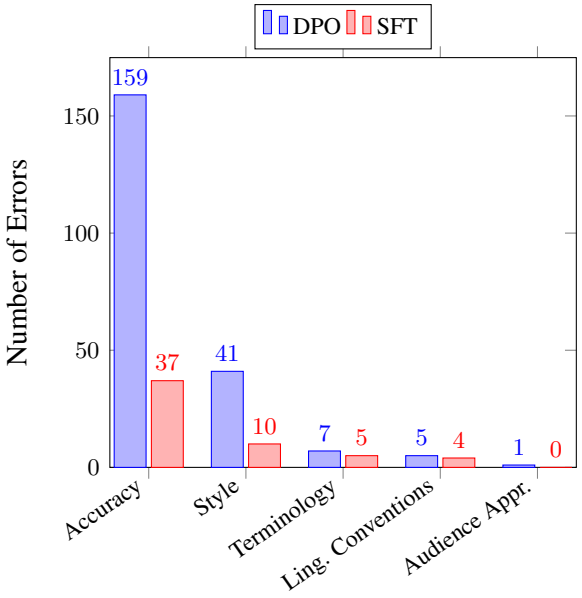

Figure 3: Distribution of dominant error categories per model. DPO shows substantially higher rates of errors related to its simplification behavior

6 Conclusions

In this study, we investigated the potential of using a hybrid data retrieval pipeline to collect relevant parallel data for domain adaptation, starting from a carefully curated small dataset. We used the retrieved data to fine-tune a small language model on the crisis domain, and then optimized via preferential learning to bias its outputs toward CEFR A2-level English.

Our results demonstrate that this two-stage optimization process enables translations that balance translation quality and accessibility, producing outputs that can be understood and acted upon by readers with diverse linguistic backgrounds.

Automatic metrics show that the readability-oriented learning substantially improves textual accessibility across all readability metrics, while incurring only moderate decreases in COMET. The ablation study further confirmed that applying preferential learning without prior domain adaptation leads to severe degradation in translation quality, whereas combining SFT with DPO yields improved readability with limited losses in adequacy. Human evaluation confirmed our results, showing that our model can generate acceptable translations, effectively conveying the meaning expressed by the source, while reducing unnecessary syntac-

tic and semantic complexity.

In conclusion, our findings suggest that simplified English can function as a viable lingua franca for emergency communication in linguistically diverse contexts such as Italy. When fully multilingual dissemination is not feasible, domain-adapted MT systems optimized for readability can support the accessibility and actionability of emergency messages for speakers with varying levels of language proficiency.

7 Future Work

One of the main limitations of our study is the absence of a human evaluation explicitly targeting the acceptability of the generated translations. While we assessed translation quality and readability through automatic metrics and human evaluation, we did not directly measure whether the generated messages effective enable readers to act upon the information conveyed in the emergency warning messages.

Future work should therefore include qualitative human assessments involving L2 English speakers representative of the international communities in Italy. Such studies would assess not only perceived readability, but also clarity of instructions, accuracy in conveying key information, and the ability to identify recommended courses of action. Task-based evaluation, where participants are asked to interpret emergency messages and report their intended behaviors, could provide deeper insights into the communicative effectiveness of these translations.

In addition, future research should investigate whether simplification strategies can be further refined to minimize omission and undertranslation errors while preserving accessibility gains. Finally, while simplified English may function as a practical shared medium, investigating simplification across multiple languages could further enhance clarity in risk communication.

8 CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.352 kgCO_2eq/kWh. A cumulative of 4 hours of computation was performed on hardware of type RTX 4090 (TDP of 300W). Total emissions are estimated to be 0.42 kgCO_2eq of which 0 percents were directly offset.

Estimations were conducted using the Machine Learning Impact calculator presented in Lacoste et al. (2019).

Acknowledgments

We thank Maria Carmen Staiano for granting access to the ITALERT corpus and for their helpful support regarding its use.

This work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples "L'Orientale", through a doctoral grant (ID 39-411-24-DOT23A27WJ-6603) to the first author, established by Ex DM 318, of type 4.1, cofinanced by the National Recovery and Resilience Plan. This work was also partially supported by the PhD Programme in Humanities and Technologies funded by the University of Macerata under D.R. No 253/2023.

The fourth author benefits from being member of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Alam, Firoj, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Amato, Amalia and Letizia Cirillo. 2024. Mediating English as a Lingua Franca for Minority and Vulnerable Groups. *MediAzioni*, 41, June.

Anastasopoulos, Antonios, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, et al. 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*.

Barbu, Paul-Gerhard, Adrianna Lipska-Dieck, and Lena Lindner. 2025. EasyJon at TSAR 2025 Shared Task Evaluation of Automated Text Simplification with LLM-as-a-Judge. In Shardlow, Matthew, Fernando Alva-Manchego, Kai North, Regina Stodden, Horacio Saggion, Nouran Khallaf, and Akio Hayakawa, editors, *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 173–182, Suzhou, China, November. Association for Computational Linguistics.

Bojar, Ondrej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post,

Raphaël Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Conference on Machine Translation*.

Cadwell, Patrick, 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.

Coche, Julien, Jess Kropczynski, Aurélie Montarnal, Andrea Tapia, and Frederick Benaben. 2021. Actionability in a Situation Awareness world: Implications for social media processing system design. In *ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management (ISBN 978-1-949373-61-5)*, number 2391, pages p.994–1001, Balcksburg (online), United States, May.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library, January. arXiv:2401.08281 [cs].

Espana-Bonet, Cristina, Adám Csaba Varga, Alberto Barrón-Cedeno, and Josef Van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Federici, Federico M, Sharon O'Brien, Patrick Cadwell, Jay Marlowe, Brian Gerber, and Olga Davis. 2019. International network in crisis translation-recommendations on policies.

Glavaš, Goran, Marc Franco-Salvador, Simone P Ponzetto, and Paolo Rosso. 2018. A resource-light method for cross-lingual semantic textual similarity. *Knowledge-based systems*, 143:1–9.

Graham, Yvette, Timothy Baldwin, and Nitika Mathur. 2015. Accurate Evaluation of Segment-level Machine Translation Metrics. In Mihalcea, Rada, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado, May. Association for Computational Linguistics.

Hajek, John, Yu Hao, Ambrin Hasnain, Anila Hasnain, Ke Hu, Maria Karidakis, Rachel Macreadie, Anthony Pym, and Juerong Qiu. 2024. Understanding and improving machine translations for emergency communications.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models, October. arXiv:2106.09685 [cs].

Jenkins, Jennifer and Anna Mauranen. 2019. Researching linguistic diversity on English-medium campuses. In *Linguistic Diversity on the EMI Campus*. Routledge. Num Pages: 18.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lankford, Séamus and Andy Way. 2024. Leveraging llms for mt in crisis scenarios: a blueprint for low-resource languages. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 4–13.

Lewis, William. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*.

Lommel, Arle and Alan Melby. 2018. Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). In Campbell, Janice, Alex Yanishevsky, Jennifer Doyon, and Doug Jones, editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Boston, MA, March. Association for Machine Translation in the Americas.

Moorkens, Joss, Andy Way, and Séamus Lankford. 2025. Sociotechnical effects of machine translation. *arXiv preprint arXiv:2503.20959*.

Musacchio, Maria Teresa, Raffaella Panizzon, et al. 2017. Localising or globalising? multilingualism and lingua franca in the management of emergencies from natural disasters. *Cultus*, 10:92–107.

North, Brian and Enrica Piccardo. 2020. *Companion volume COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT Companion volume Language Policy Programme Education Policy Division Education Department Council of Europe*. July.

O'Brien, Sharon. 2022. Crisis translation: A snapshot in time. *INContext: Studies in Translation and Interculturalism*, 2(1):84–108.

O'Brien, Sharon and Federico Marco Federici. 2020. Crisis translation: Considering language needs in multilingual disaster settings. *Disaster Prevention and Management: An International Journal*, 29(2):129–143.

Pang, Jianhui, Fanghua Ye, Longyue Wang, Dian Yu, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2024. Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models, January. Issue: arXiv:2401.08350 arXiv:2401.08350 [cs].

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Piller, Ingrid, Jie Zhang, and Jia Li. 2020. Linguistic diversity in a time of crisis: Language challenges of the covid-19 pandemic. *Multilingua*, 39(5):503–515.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Radicioni, Maura and Lucía Ruiz Rosendo. 2022. Cultural mediation as a means of effective multilingual communication. *Translating Crises, London: Bloomsbury Publishing*, pages 237–251.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July. arXiv:2305.18290 [cs].

Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Rossetti, Alessandra, 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.

Seidlhofer, Barbara. 2011. *Understanding English as a Lingua Franca: A complete introduction to the theoretical nature and practical implications of English used as a lingua franca.* OUP Oxford, September. Google-Books-ID: Mcd6PgAACAAJ.

Staiano, Maria Carmen, 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.

Stap, David, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206, Bangkok, Thailand, August. Association for Computational Linguistics.

Sugathadasa, Keet, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2017. Synergistic union of word2vec and lexicon for domain specific semantic similarity. In *2017 IEEE international conference on industrial and information systems (ICIIS)*, pages 1–6. IEEE.

Sun, Jiao, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore, December. Association for Computational Linguistics.

Tiedemann, Jörg and Lars Nygaard. 2004. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus. In Lino, Maria Teresa, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Torresi, Ira. 2020. *Translating promotional and advertising texts*. Routledge.

United Nations Office for Disaster Risk Reduction (UNDRR), International Science Council (ISC). 2025. *UNDRR-ISC Hazard Information Profiles – 2025 Update*. United Nations Office for Disaster Risk Reduction; International Science Council.

Vieira, Inacio, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes, September. arXiv:2409.03454 [cs].

Way, Andy, Rejwanul Haque, Guodong Xie, Federico Gaspari, Maja Popović, and Alberto Poncelas. 2020. Rapid development of competitive translation engines for access to multilingual covid-19 information. In *Informatics*, volume 7, page 19. MDPI.

Zhou, Lexin, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, October.

Zouhar, Vilém and Tom Kocmi. 2026. Pearmut: Human evaluation of translation made trivial.

A Simplification Prompt

We report the prompt used in Section 4.2, to ensure the replicability of the study. The prompt was used with the model gpt-5-mini with low reasoning effort, to generate the simplification corpus. The prompt was first attested in Barbu et al. (2025).

Prompt for Text Simplification

You are a text simplification AI. Your task is to simplify the following input to A2 CEFR level. Use only common, everyday words that are appropriate for the context. Choose words that native speakers would naturally use. Explain essential terms if those can't be simplified and maintain the content as in the original.

Input: {text}

Answer just with the simplification and nothing else. Keep the original tone.

B Annotated Examples

Table 5 shows representative annotated translation examples drawn from the human evaluation dataset. The tables report outputs generated by the Supervised Fine-Tuned model and the readability-oriented DPO model.

For each segment, we provide the source text, the system translation, the identified MQM error category, the assigned severity level, and the final MQM score assigned by the annotators. These examples illustrate the most recurrent error patterns discussed in Section 5, with particular emphasis on omission, undertranslation, and stylistic simplification phenomena. Table 4 reports the distribution of MQM error categories, types, and severity levels for both the SFT and DPO systems.

C Annotators Profile

The two annotators described in Section 5 were involved in multiple stages of the study, including stratified annotation (§3.3), evaluation of the simplification dataset (§4.2), and the human evaluation. Both annotators hold a Master's degree in Translation Studies. They are native speakers of Italian, fluent in English, and have domain-specific expertise in crisis communication.

		DPO				SFT		
Category	**Type**	MAJ	MIN	NEU	**Tot**	MAJ	MIN	**Tot**
Accuracy	Addition	3	2	–	5	1	1	2
	Mistranslation	15	18	–	33	6	7	13
	Omission	17	26	–	43	4	5	9
	Overtranslation	–	4	–	4	–	2	2
	Undertranslation	10	64	–	74	3	8	11
Audience Appropriateness	Offensive	–	1	–	1	–	–	0
Linguistic Conventions	Grammar	1	3	–	4	–	4	4
	Punctuation	–	–	1	1	–	–	0
Style	Awkward style	2	14	–	16	–	4	4
	Language register	1	24	–	25	–	6	6
Terminology	Wrong term	2	5	–	7	2	3	5
Total		51	161	1	213	16	40	56

Table 4: MQM error counts by category, type, and severity for DPO and SFT systems. MAJ = major, MIN = minor, NEU = neutral. Note that SFT has no neutral errors.

SFT Model				
Source (IT)	**Output (EN)**	**Error Type**	**Severity**	**Score**
Negli Stati Uniti, ad esempio, il 72% delle donne con disabilità vive al di sotto della soglia di povertà. La discriminazione intensificata che le persone con disabilità possono subire a causa del loro sesso è particolarmente importante da considerare quando si tiene conto del fatto che, secondo l'Organizzazione per la cooperazione e lo sviluppo economico, le donne segnalano un'incidenza maggiore di disabilità rispetto agli uomini.	In the United States, for example, 72% of women with disabilities live below the poverty line. The intensified discrimination that people with disabilities may face because of their gender is particularly important to consider when taking into account the fact that, according to the United Nations, women report a higher incidence of disabilities than men.	Accuracy/Mistranslation	Major	65
Negli ultimi anni abbiamo avuto modo di constatare il disastroso impatto che possono avere le alluvioni e anche il mio paese, che non per nulla si chiama "Paesi Bassi", ha subito parecchie inondazioni.	In recent years we have seen the disastrous impact that floods can have, and even my country, which is not called the Netherlands for nothing, has had several floods.	Style/Language Resister Accuracy/Mistranslation Accuracy/Mistranslation	Minor Minor Minor	70
Mettere sullo stesso piano "l'inquinamento di Internet", le emissioni di diossido di carbonio e la contaminazione dell'acqua aumenterebbe di gran lunga la possibilità di controllo delle operazioni online da parte del governo.	Equating "internet pollution" with carbon emissions and water contamination would greatly increase the possibility of government control over online operations.	Style/Language register	Minor	90
AC: Abbiamo previsto 8 priorità, che concernono ognuno di noi: la qualità dell'acqua, l'inquinamento della plastica, il trasporto marittimo, il turismo sostenibile, le piattaforme d'alto mare, le energie marine rinnovabili, il cambiamento climatico e la condivisione delle conoscenze sull'oceano: ocean literacy (alfabetizzazione sull'oceano).	AC: We have identified eight priorities, which concern each of us: water quality, plastic pollution, shipping, sustainable tourism, offshore platforms, renewable marine energy, climate change and ocean literacy.	Accuracy/Omission	Major	65
In conclusione, credo sia necessario promuovere, e sostenere con un'organica attività di carattere formativo e con attrezzature e mezzi, l'azione del volontariato di protezione civile, che diventa una risorsa indispensabile per fronteggiare stati di emergenza e catastrofi naturali e di altri tipo.	In conclusion, I believe that it is necessary to promote and support, through training and equipment, the action of civil protection volunteers, who are an indispensable resource for dealing with emergencies and natural and other disasters.	Accuracy/Undertranslation	Major	80
DPO Model				
Source (IT)	**Output (EN)**	**Error Type**	**Severity**	**Score**
Negli Stati Uniti, ad esempio, il 72% delle donne con disabilità vive al di sotto della soglia di povertà. La discriminazione intensificata che le persone con disabilità possono subire a causa del loro sesso è particolarmente importante da considerare quando si tiene conto del fatto che, secondo l'Organizzazione per la cooperazione e lo sviluppo economico, le donne segnalano un'incidenza maggiore di disabilità rispetto agli uomini.	In the United States, for example, 72% of women with disabilities live below the poverty line. The increased discrimination that people with disabilities may face because of their gender is especially important to consider when you consider that, according to the United Nations, women are more likely to have disabilities than men.	Style/Awkward style Accuracy/Mistranslation	Minor Major	55
Negli ultimi anni abbiamo avuto modo di constatare il disastroso impatto che possono avere le alluvioni e anche il mio paese, che non per nulla si chiama "Paesi Bassi", ha subito parecchie inondazioni.	We have seen the impact of floods in recent years, and my country, which is called the Netherlands, has had a lot of flooding.	Accuracy/Omission Style/Language register Accuracy/Undertranslation Accuracy/Mistranslation	Minor Minor Minor Minor	60
Mettere sullo stesso piano "l'inquinamento di Internet", le emissioni di diossido di carbonio e la contaminazione dell'acqua aumenterebbe di gran lunga la possibilità di controllo delle operazioni online da parte del governo.	Comparing Internet pollution to carbon emissions and water pollution would make it easier for the government to control the internet.	Linguistic conventions/Punctuation Accuracy/Undertranslation Accuracy/Undertranslation	Neutral Minor Minor	75
AC: Abbiamo previsto 8 priorità, che concernono ognuno di noi: la qualità dell'acqua, l'inquinamento della plastica, il trasporto marittimo, il turismo sostenibile, le piattaforme d'alto mare, le energie marine rinnovabili, il cambiamento climatico e la condivisione delle conoscenze sull'oceano: ocean literacy (alfabetizzazione sull'oceano).	AC: We have eight priorities, which are about each of us: water quality, plastic pollution, shipping, sustainable tourism, offshore platforms, renewable energy, climate change and ocean literacy.	Style/Awkward style Accuracy/Omission Accuracy/Omission Accuracy/Undertranslation	Minor Major Major Minor	35
In conclusione, credo sia necessario promuovere, e sostenere con un'organica attività di carattere formativo e con attrezzature e mezzi, l'azione del volontariato di protezione civile, che diventa una risorsa indispensabile per fronteggiare stati di emergenza e catastrofi naturali e di altri tipo.	In conclusion, I believe that we must promote and support the work of civil protection volunteers, by providing them with the necessary training and equipment, so that they can be a valuable resource in the event of emergencies and natural or man-made disasters.	Style/Language register Accuracy/Omission Accuracy/Addition	Minor Minor Minor	60

Table 5: Representative MQM error examples identified in the SFT and DPO model outputs.