ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO DI LAUREA MAGISTRALE IN SPECIALIZED TRANSLATION (CLASSE LM-94)

EVALUATING DOMAIN ADAPTATION IN NEURAL MACHINE TRANSLATION AND LARGE LANGUAGE MODELS: INSIGHTS FROM THE TICO-19 BENCHMARK

Tesi di laurea magistrale in Machine Translation

Relatore:

Prof. Federico Garcea

Co-relatrice:

Prof. Annalisa Crea

Co-relatore:

Prof. Adriano Ferraresi

Presentata da:

Lucia Galiero

Sessione di Laurea Marzo 2025

Anno Accademico 2023/2024

Acknowledgments

The road that got me here was no bed of roses. When I joined Dit.Lab in September 2022, I brought with me the background of a student with a Bachelor in Interlingual Mediation: a solid foundation in foreign languages and a strong interest in theoretical and applied linguistics, but little experience in academic writing, quantitative analysis, or coding—a gap that, to some extent, still remains. Looking back, however, I realize that stepping outside my comfort zone has been more rewarding than I could have anticipated. While continuing to work within the field of linguistics, I had to engage with new disciplines and methodologies. Despite the challenges, I have grown both academically and personally, gaining skills and resilience that I might not have developed otherwise.

First and foremost, I would like to express my sincere gratitude to Professor Garcea for introducing me to machine translation and for his unwavering support throughout the writing process. I am also deeply grateful to Professor Crea for providing invaluable feedback on the translations and to Prof. Ferraresi for his guidance throughout this adventure at DIT.

However, the most important part of this journey lies in the people that I encountered and made these two years better than I ever imagined. Therefore, thanks to Alice, for all the beers we had at Valverde and for the precious guidance in how to survive the amazing world of TraTec. Speaking about TraTec and SpecTra, I want to express my deepest affection to all my trusted travel companions during these two years: Marianna, Priya, Federico, Eleonora, Debora, Yasmin, Rossella, Domizia, Carmen, Alessandro, Federica, Chiara and Aurora. You are all some of the most kind, passionate and hard-working people I have ever met. I really could not ask for better colleagues and I am sure each of you will find their way into this world.

To Matteo, Marco and Enrico, for all the laughs, the cooking, the heated discussions on politics and the memes we shared that year we lived all in via Godoli. The atmosphere we created was something every fuori sede should have. Thanks for building a space where I was always happy to come back to. I will carry it forever in my fondest memories of Forlì.

To the extraordinary women working as volunteers at Associazione Paolo Babini, thank you for welcoming me as a (grand) daughter and for giving me a sense of purpose during one of my toughest times.

To all my beautiful friends in Genoa. Some of you have known me since high school, some arrived before my moving to Forlì, others later. Either way, you have become my chosen family, and I feel so blessed to have you in my life – I will always be.

Finally, thanks to my family, from those that share the same roof with, to my uncles, aunts and cousins (first, second and "adoptive") scattered around Italy. Some of you have always treated me like a second daughter, or the daughter you never had, teaching me that love really transcends time and space.

Once again, just thanks.

I love you all.

Abstract

English

This thesis examines the impact of domain adaptation on the performance of an adaptive Neural Machine Translation (NMT) system (ModernMT) and a Large Language Model (LLM - LLaMa 3.2 90B) in the English-Italian language pair. The analysis is based on an experiment using data from the TICO-19 benchmark, a multilingual dataset developed to support translation efforts during the COVID-19 pandemic. Since no Italian version of the benchmark is currently available, a preliminary phase involved the manual translation and alignment of two selected academic-scientific articles to create a high-quality reference set.

The research is framed within the broader context of Crisis Translation, a growing field in Translation Studies that investigates the role of linguistic mediation in emergency scenarios. Particular attention is given to Crisis Machine Translation (CMT), an emerging subfield exploring how MT systems can be optimized for use in crisis contexts, where the speed and accuracy of multilingual communication are paramount.

The evaluation includes both automatic metrics (BLEU, chrF3, COMET) and a human assessment phase to determine whether domain-adapted LLMs can achieve comparable or superior results to an adaptive NMT system. The findings reveal that, despite the increasing capabilities of LLMs, the domain-adapted NMT system consistently outperforms its counterpart, challenging the assumption that LLMs inherently excel in specialized translation tasks.

The structure of the thesis is divided into four main sections: an introduction to Crisis Translation and its applications, a discussion of the technologies employed, an overview of the experimental methodology, and a final analysis of the results. This research contributes to the growing discussion on the feasibility of integrating LLMs into domain-adaptive machine translation and provides insights into the practical implications of deploying such technologies in crisis scenarios.

Abstract

Italiano

Questa tesi analizza l'impatto dell'adattamento a dominio sulle prestazioni di un sistema di Traduzione Automatica Neurale adattiva (ModernMT) e di un Modello Linguistico di Grandi Dimensioni (LLM - LLaMa 3.2 90B) nella combinazione linguistica inglese-italiano. L'analisi si basa su un esperimento condotto utilizzando dati provenienti dal benchmark TICO-19, un dataset multilingue sviluppato per supportare gli sforzi di traduzione durante la pandemia di COVID-19. Poiché il benchmark non è attualmente disponibile in italiano, la prima fase del lavoro ha previsto la traduzione manuale e l'allineamento di due testi accademico-scientifici selezionati per la creazione di un set di riferimento di alta qualità.

Lo studio si colloca nel contesto più ampio della Crisis Translation, un ambito emergente all'interno dei Translation Studies che indaga il ruolo della mediazione linguistica in situazioni di emergenza. Particolare attenzione è rivolta alla Crisis Machine Translation (CMT), una branca che esplora come i sistemi di traduzione automatica possano essere ottimizzati negli scenari di crisi, dove la rapidità e la precisione della comunicazione multilingue sono essenziali.

La valutazione delle traduzioni include metriche automatiche (BLEU, chrF3, COMET) e un'analisi manuale per determinare se gli LLM adattati al dominio possano ottenere risultati comparabili o superiori rispetto a un sistema di traduzione automatica adattiva. I risultati mostrano che, nonostante le crescenti capacità degli LLM, il sistema di traduzione automatica adattiva mantiene prestazioni migliori, mettendo in discussione l'ipotesi che gli LLM eccellano automaticamente nei contesti di traduzione specialistica. La tesi è strutturata in quattro sezioni principali: un'introduzione alla Crisis Translation e alle sue applicazioni, un'analisi delle tecnologie impiegate, una descrizione della metodologia sperimentale e, infine, un'analisi dettagliata dei risultati. Questo studio contribuisce alla discussione sull'integrazione degli LLM nella traduzione automatica adattativa e fornisce spunti sulle implicazioni pratiche dell'uso di tali tecnologie nei contesti di crisi.

List of Figures

List of Tables

Contents

Introduction

Amid global pandemics, wars, and extreme weather events, the contemporary world has faced an increasing number of humanitarian crises. To prepare for future challenges, it remains crucial to ensure rapid access to accurate multilingual information. Machine Translation (MT) has already provided significant support in this area, as demonstrated by the Translation Initiative for COVID-19 (TICO-19; Anastasopoulos, 2020) and Mission 4636 during the 2010 Haiti earthquake (W. Lewis, 2010). However, recent advances in artificial intelligence, particularly in Large Language Models (LLMs), could make an even greater contribution for future development of language technologies in settings of crisis.

Literature on the LLM translation capabilities hints at mixed results: while excelling in general-purpose translation, sometimes surpassing commercial MT systems, LLMs which are available to the general public appear to fall behind when it comes to specialized translation, even after undergoing adaptation process (Wassie et al., 2024; Moslem, 2023a).

The present dissertation aims at testing the performance of domain adaptation in a Neural Machine Translation (NMT) system and a LLM for the language pair English-Italian. The testing material is represented by two PubMed articles coming from the TICO-19 dataset. As the TICO-19 benchmark has not been released to date in Italian[1], and considering the importance of high-quality human reference for MT evaluation (Freitag et al., 2020; Freitag et al., 2023), the articles have also been manually translated and reviewed prior to any experiment. The objectives of this dissertation are further outlined as follows:

> ***Hypothesis:*** *Domain adapted, LLM-driven translation can achieve equal or better translation results than a NMT system that has, in its turn, undergone a domain adaptation process.*

> ***Research Question n°1 (RQ1):*** *Do LLMs deliver better results both before and after domain adaptation?*

[1] 28th December 2024

Research Question n°2 (RQ2): *Does domain adaptation yield such an impact on the output to be necessary for future development of translation technologies in crisis scenarios?*

The present dissertation is organized in four chapters. To better understand how translation was assessed during the COVID-19 pandemic and contextualising TICO-19, Chapter 1 is devoted to translation in crisis scenarios. It first frames Crisis Translation as a research field, providing a definition for the term and illustrating its current research topics. The discussion then moves on to how automated translation has served in crisis settings, offering the Haiti Earthquake as the major example, and illustrates several ethical observations for developing MT engines in respect of all involved parties. The Chapter contains a brief outline of translation as a response tool during the COVID-19 pandemic, closing on a detailed description of TICO-19, its objectives, achievements and additional resources made available to foster multilingual communication in future crises.

Chapter 2 focuses on the technologies leveraged in this work: starting with a brief overview of neural machine translation and its state-of-the-art architecture, it delves into domain adaptation for NMT systems, explaining its mechanisms and illustrating its achievements from recent studies. Following this, the chapter will outline the history of language modelling, explore LLMs and their functioning, and conclude with recent findings on LLM-based translation and their application to domain adaptation.

Chapter 3 outlines the applied methodology, including the selection of texts, the translation workflow, and the construction of test and tuning sets. Additionally, it delves into the chosen NMT engine, the LLM, and the specifics of the domain adaptation process, as well as evaluation frameworks adopted.

Chapter 4 provides answers to the outlined hypothesis and research question. It will analyse results of the comparative quality evaluation between the generic and adapted systems for NMT and LLM-based translation, and will address limitations of our presented work.

1. Translation Technologies and Crisis Situations

1.1. Introduction

This first Chapter provides an overview of the role of translation technologies in crisis scenarios. It begins by defining crisis translation (1.2.1.) and outlining key research topics (1.2.2.), including multilingual information access, citizen translators, and technological challenges. The discussion then shifts to introducing Crisis Machine Translation (CMT; 1.2.3.) and major crisis translation initiatives, such as Mission 4636 (1.2.4.) during the Haiti earthquake. Ethical concerns regarding the use of MT in crisis scenarios are also explored (1.2.5.), emphasizing the need for responsible development, data privacy, and linguistic inclusivity. The Chapter closes in to the more recent COVID-19 pandemic, exploring how multilingual information access was provided during the global crisis (1.3.1). In this context, the Translation Initiative for COVID-19 (TICO-19) serves as a case study, illustrating how high-quality multilingual translation data was developed to address information gaps during the pandemic (1.3.2.). The initiative is also discussed in terms of benchmark composition (1.3.3.), its construction (1.3.4.), additional resources provided (1.3.5.) and major achievements (1.3.6.).

1.2. Research framework

1.2.1. Defining Crisis Translation

The European Convention on Human Rights integrates access to information inside Article 10, as a founding pillar of freedom of expression. Specifically, information should be imparted to everyone “without interference of authority and regardless of frontiers” (Council of Europe, 1950). When said frontier is represented by language barriers and is collocated in situations where accuracy, clarity and timeliness are key, this human right can be seriously compromised. In such scenarios, translation, whether handmade or automated, can serve as a vital tool for ensuring this right in diverse and multilingual populations.

The concept of “crisis translation” presents a significant terminological challenge, as the expression designates a relatively recent and currently evolving (O’Brien, 2022) subfield within Translation Studies (TS) with tangible implications for its research focus.

In her recent attempt at framing the current state of crisis translation and its potential research directions, O'Brien (ibid.,) notes that the concept of "crisis" is often used interchangeably with "disaster," although the latter lacks a universally agreed definition within disaster studies (Perry, 2007; Perry, 2018). According to the definition offered by Quantarelli (1998), a crisis is "an unexpected event, with sudden or rapid onset that can seriously disrupt the routines of an individual or a collective and that poses some level of risk or danger" (as cited in O'Brien, 2022, p. 86), that can be provoked by triggers ranging from natural hazards to technological failures. The term "crisis" can then refer to a variety of disruptive events, such as natural disasters, health emergencies, armed conflicts, cyber- and terrorist attacks (Alexander and Pescaroli, 2019; O'Brien and Federici, 2019) with cascading effects on the lives of those affected. It is also important not to conceive the crisis as the *response* phase alone, but rather as having a life cycle composed by four distinct phases: (1) mitigation, (2) preparedness, (3) response, and (4) recovery (NGA, 1979; Landahl et al., 2019; O'Brien, 2022).

Turning to the definition of crisis translation, Federici et al. (2019) define it as "any form of linguistic and cultural transmission of messages that enable access to information during an emergency, regardless of the medium" (p. 247), emphasizing the importance of communication that transcends linguistic barriers in urgent contexts. Additionally, O'Brien (2022) presents three takeaways to consider when navigating crisis translation as a discipline. Firstly, the local extent of a crisis does not necessarily imply that communications are only to be issued in one single language (Federici & O'Brien, 2019, p. 3; P. Wang, 2019).

Secondly, crisis translation shall not be considered entirely synonymous with conflict translation, nor with community translation or translation in development settings (O' Brien, 2022, p.90). In fact, not all crises result from conflict, although the latter can be characterized by an unexpected nature, rapid onset, and disruptive effects, and developmental settings instead often lack the urgency associated with unexpected crises. As for community translation, it is described as typically addressing the routine translation needs of minority and immigrant populations (ibid., p. 91), and stands in contrast with crisis translation because of the planned and structured nature of community communication (p. 92).

Thirdly, the overview conceives crisis translation as entailing both written translation and oral interpretation, since both are crucial in urgent scenarios (ibid., p. 89).

However, if we consider the entire life cycle, written translated content might have a longer lasting effect and impact. For this reason, O'Brien (2022) uses the notion of "crisis translation" as indicating the written modality of translation, and the same will be done for the scope of the present work as well.

1.2.2. Current research topics and future directions

The overview provided by O'Brien (ibid., p. 93), summarized topics dealt in the field of crisis translation as follows:

1. **Emergency Management Policy and Translation -** examines the role of translation in emergency policies by national and regional institutions. Ideally, effective emergency policies should cater to a global, multilingual society, ensuring accurate information reaches diverse language communities and vulnerable groups like the deaf, blind, and disabled. However, latest research (Civico, 2021; O'Brien et al., 2018; P. Wang, 2019) highlights that even industrialized nations struggle to integrate translation into their emergency frameworks.
2. **Citizens Translators, Training and Ethics –** explores how volunteer translators can support crisis response when professionals are unavailable. The most relevant topic in this sense is represented by the training of untrained personnel and volunteer to provide translation in emergencies (Federici and Cadwell, 2018; Federici et al., 2019).
3. **Technological Issues -** investigates how tools like translation memory (TM), machine translation (MT), and terminological databases contribute to crisis communication. MT is particularly valued for its speed, playing a crucial role in various crisis scenarios (Anastasopoulos et al., 2020; W. Lewis, 2011). However, challenges remain, such as ensuring IT infrastructure and power availability to use these tools effectively (O'Brien, 2019, as cited in O'Brien, 2022).

To date, the International Network in Crisis Translation (INTERACT[2]) project might be considered the most comprehensive project on crisis translation and all its research topics. Active between 2017 and 2020 and funded by the European Commission, it aimed

[2]Site at: https://sites.google.com/view/crisistranslation/home [Last accessed 16/12/2024]

at enhancing preparedness, response, and recovery efforts in crises by fostering collaboration and innovation in a cross-disciplinary approach and with ethics as a central pillar.

The focus on interdisciplinarity such as the one adopted within INTERACT is considered key for the advancement of crisis translation (O'Brien, 2022), as TS scholars have seldom been involved in disaster studies teams. The US Federal Emergency Management Agency (FEMA, 2018) outlines five key principles for disaster research: reviewing past work, fostering interdisciplinarity, ensuring studies are carried in respect of all involved parties, transferring knowledge from use cases, and maximizing impact of work carried out. These principles can also guide TS contributions and incorporated into disaster studies, so that researchers can ensure that linguistic and cultural diversity are prioritized in crisis response strategies. Ethical issues, such as the use of untrained citizen translators and the consequences of non-translation, also warrant closer scrutiny to balance professional standards with the urgency of emergencies. While disaster studies frameworks offer valuable insights into integrating translation into global crisis management, technology — particularly automated translation — transfers these theoretical insights into practice. By doing so, it ensures timely multilingual communication, provided that potential risks of misuse are addressed.

1.2.3. Crisis Machine Translation

It has been recently argued (Roussis, 2022) that machine translation in contexts of crisis, or "Crisis Machine Translation" (ibid.), should better be considered as a special branch of Natural Language Processing (NLP). While Machine Translation in general is one of the most important applications of NLP, its technicalities and recent advancements will be discussed more in depth in Chapter 2. Instead, Crisis Machine Translation will be hereafter discussed as part of crisis plans. Additionally, some key takeaways on how to develop MT engines in crisis settings will be presented, alongside ethical questions raised by TS scholars on the use and development of MT in contexts of crisis.

1.2.4. Mission 4636 and the Crisis MT Cookbook

As mentioned in 1.2.2., the use of MT in crisis scenarios represents a current object of interest within crisis translation, since it constitutes the swiftest way to date to deliver

multilingual information to groups affected by a crisis. The most prominent contributions in this regard are currently offered by W. Lewis (2010; W. Lewis et al., 2011) and were both developed in the context of Mission 4636[3]. Such initiative was named after the phone number used for emergency communications via phone and text messages in the immediate aftermath of the 2010 earthquake, where swift translation of content from and into Haitian Creole was urgently needed by international aid organizations. To further exacerbate difficulties in relief operations there was a significant lack of knowledge and resource coverage for Haitian Creole, which mostly represented the only language spoken by most Haitians in affected areas.

With the help of native speakers and the relief community, a team of experts from Microsoft Translator managed to build, in a matter of few days, a fully functioning machine translation engine for translation from and into Haitian Creole, which was also integrated in the Mission 4636 relief infrastructure. The system achieved commercial-grade quality on a test dataset comprised of actual SMS from Mission 4636 and a corpus provided by CMU (W. Lewis, 2010, p. 5). These commendable efforts were promptly recognized as a starting point for future development of MT engines in crisis scenarios (Callison-Burch, 2011; Munro, 2010).

Building on the lessons learned with Mission 4636, W. Lewis et al. (2011) proceeded to set up a "cookbook" in how to develop MT systems within a crisis response framework, particularly in the case of low resource languages (LRLs[4]) involved.

The first is represented by the data used for engine development, namely phrases, vocabulary, sentences with crisis-related content. Ideally, content should be related to the type of crisis at stake and available in a pivot language (e.g., English) to facilitate translation into the target language(s) and distribution to aid organizations (ibid.). However, as seen with Haitian Creole in the 2010 earthquake (W. Lewis, 2010), lack of parallel and domain-specific data[5] seriously challenges MT output quality.

In terms of data sources, useful content might come from NGOs and other international organizations (e.g., the Emergency Multilingual Phrasebook by the British Red Cross; see NHS Confederation, 2004), but also from crowdsourced translation. This practice, which generally engages the local community in the translation of content from

[3] Official page of the initiative: https://www.mission4636.org [last accessed 12/12/2024]
[4] We are here sticking to the definition by Magueresse et al. (2020), who define Low Resource Languages as "languages for which statistical methods cannot be directly applied because of data scarcity." (p.1)
[5] Such issue takes also the name of data sparsity

the field, can be used both for generating new data and applying corrections to existent MT output. Therefore, a sound response infrastructure that integrates crowdsource translation is regarded as the second crucial aspect of the Crisis MT cookbook (W. Lewis et al., 2011, p. 508). Three key components are listed by the W. Lewis et al. (ibid.) to ensure robustness of the response infrastructure:

1) A crowdsourced micro tasking system to translate and route field messages, as mentioned prior, like the one that was implemented by Mission 4636 and contributed to its ultimate success (Munro, 2010).
2) Fully integrated APIs for public MT engines (e.g., Microsoft Translator, Google Translate) into messaging systems, to allow immediate deployment of MT services.
3) A mobile application functioning as a crisis-specific translation memory could provide relief workers with up-to-date resources offline. In this way, reliance on paper materials is reduced while ensuring access in areas with limited connectivity.

The attention drawn to this latter aspect suggests that the MT community has already been discussing potential issues with MT engines and IT infrastructure capacity in crisis-ridden areas, anticipating concerns raised elsewhere by TS scholars (O'Brien, 2022, p. 97-99).

1.2.5. Ethical concerns about MT in crisis scenarios

As technologies develop, it is natural to raise issues on how it is best to use them and do so without damaging any of the parties, and Crisis MT is no different. A common general question is posited by the accuracy of raw MT output (Nurminen & Koponen, 2020). Neural Machine Translation (NMT) in particular offers highly fluent output, though such fluency does not prevent engines from generating inaccurate linguistic content (Castilho et al., 2017; Koehn & Knowels, 2017, Moorkens et al., 2018). Besides, output quality notably depends on language direction and domain (Koehn & Knowels, 2017).

At the time of writing, the extent to which MT can reduce or heighten the effects of crises remains relatively underexamined (Cadwell et al., 2019), but significant observations for an ethical implementation of Crisis MT have been advanced in recent times (Federici, 2023; Parra Escartín & Moniz, 2020). A significant point of discussion originates by the fact that, whether used for specialized or generalized purposes, NMT

engines tend to rely on large amounts of data. Thus, ethical concerns with MT use in crisis scenarios have shifted towards dealing with data quality, privacy and ownership.

According to Federici (2023), good quality data and domain-specific resources go hand in glove with anticipation. In other words, language resources for Crisis MT development should be created, curated and stored before the next crisis sets on. In this way, MT can be leveraged through the whole crisis life cycle as described above, i.e., also as a mean for crisis *preparedness* and not just for crisis *response*. Parra Escartín & Moniz (2020) had also stressed the importance of creating ad-hoc resources for future emergencies, and even proposed four alternative workflows to that described by W. Lewis et al. (2011). The workflows ranged from relying on human translation only to implementing MT with full Post-Editing (PE), aiming at accommodating crisis scenarios involving failure of online services and lack of online resources.

Most importantly though, Parra Escartín & Moniz (2020) illustrate several privacy issues on collected data. The first step to take, even before setting up any operational translation workflow, is to establish clearly

> "who has access to the data, who is the data curator and manager, how is the data processed and where and how it is stored are key prior to establishing any translation workflow to ensure that all parties are protected from potential data and privacy breaches, or even potential threats like cyberattacks" (ibid., p. 16).

As a general suggestion, Parra Escartín & Moniz (ibid.) indicate that data should be anonymized or encrypted whenever possible to safeguard privacy of stakeholders. On the other hand, anonymization can still pose significant problems, since anonymized datasets may risk re-identification when aggregated with diverse sources (ibid.). Caution has also been advised when using free online MT systems (Boulanger, 2024; Nurminen and Koponen, 2020), since potential data breach may be involved in their use. On the issue of ownership, Parra Escartín & Moniz (2020, p. 18) note the difficulty of addressing it in crowdsourced MT workflows, where contributions from different stakeholders might blur intellectual property rights. Shared ownership has been proposed as a solution (Moorkens and D. Lewis, 2019) but can soon become impracticable in crisis scenarios (Parra Escartín & Moniz, 2020, p. 18). Other general proposals include recruiting professional editors and translators for MT quality evaluation, given the centrality of information accuracy in crisis settings.

Finally, Parra Escartín & Moniz (2020) recommend that MT engines or

autonomous systems involved in crisis settings should be compliant with the four IEEE Global Initiative for Artificial Intelligence and Autonomous Systems (AI/AS; 2016). The first of the principles, Human Benefit, emphasizes respect for human rights, dignity, and cultural diversity, with systems designed to be secure and traceable. Secondly, Responsibility, calls for accountability through legislation, culturally appropriate implementation, and comprehensive documentation of workflows and system parameters. Transparency, the third principle, demands that AI/AS operations and decisions be explainable, fostering trust through clear quality thresholds and accessible processes. Finally, Education and Awareness underline the need to train stakeholders and raise public understanding of AI/AS to prevent misuse. In the context of Crisis MT, these principles may be implemented by developing systems that respect cultural nuances, documenting decisions and workflows for accountability, setting clear performance benchmarks, and ensuring all stakeholders demonstrate literacy in the ethical use of MT (Parra Escartín & Moniz, 2020).

The complete application of the four IEEE principles for Crisis MT cannot thus function without training programs inside and outside universities that foster MT and Post-Editing (PE) literacy (Federici, 2023; Bowker & Buitrago Ciro, 2019). Besides, Crisis MT cannot be integrated as tool for crisis management if there is no acknowledgment of multilingualism, translation and interpreting at the core of emergency plans. Unfortunately, as it will soon be discussed, translation and multilingualism seem to have been overlooked even in recent, larger crises.

1.3. The Translation Initiative for COVID-19 (TICO-19)

1.3.1. Multilingualism and translation during the COVID-19 pandemic

From the initial outbreak in Hubei, China, in December 2019 to its declared end in May 2023, the COVID-19 pandemic marked one of the most significant global health crises since 1918, underscoring challenges in accessing reliable information. While the Sars-Cov-2 genome was sequenced and made public in January 2020, key details about transmission, prevention, and treatment remained unclear. As research advanced, providing critical insights for vaccine development and containment, an "infodemic" of mixed information emerged, driven by low literacy and mistrust in news sources (Balakrishnan et al., 2022). The global nature of the pandemic amplified these challenges,

complicating efforts to deliver accurate information across diverse populations. Despite calls from the World Health Organization for coordinated responses, most countries managed the crisis independently, reacting primarily to local outbreaks.

In terms of multilingual information access, Mulloch (2020) promptly defined the COVID-19 pandemic as the biggest translation challenge in history, especially for minority languages and LRLs. The global health emergency posed by the Sars-CoV-2 thus accentuated once more the need for translation and interpreting to be recognized and included in crisis management and risk reduction (Civico, 2021; O'Brien, 2022; Jie Zhang & Yu-chen Wu, 2020). However, Civico (2021) points out that multilingualism, where addressed, was arguably approached poorly to the benefit of minorities marginalized communities. His analysis investigated how language barriers were tackled during the pandemic, and ultimately revealed that only a handful of countries with a tradition of multilingualism or migrant communities (e.g., Belgium and Portugal; ibid, p. 10) succeeded to leverage multilingualism to cover the needs of the general population and marginalized groups. As to other nations, Civico (ibid., pp. 8-9, 11-14) reports that most countries relied primarily on official languages, leaving linguistic minorities underserved and exacerbating inequalities. Moreover, reactive measures often involved using machine translation or untrained interpreters, resulting in inconsistent quality and limited accessibility. The conclusions advanced by Civico (ibid., p. 17) fall perfectly in line with calls made by other scholars in crisis translation on multilingual and cultural preparedness, integration of technology and human oversight of translation workflows (Civico, 2021; Federici, 2023, Parra Escartín & Moniz, 2020).

1.3.2. Launch and objectives of TICO-19

The Translation Initiative for COVID-19 (TICO-19; Anastasopoulos et al., 2020) delivered a significant contribution in addressing both misinformation and swift multilingual communication during the pandemic. Coordinated by a group of experts from prominent universities and tech companies, and in collaboration with the NGO Translation Without Borders, the initiative set out in March 2020 and results of the original project were first published in the summer of the same year[6] (ibid.).

[6] Paper and resources also available at: https://tico-19.github.io/ [Last accessed on 17/11/2024]

The most important resource created within this collaboration is a collection of machine-readable translation data related to the Coronavirus pandemic. Said set is composed of 3,000 sentences on COVID-19-related content coming from a diversified range of sources, which were translated by professional translators in 38 languages to ensure the highest possible quality of parallel data[7]. The dataset is designed to be used as a benchmark for the swift development of translation technologies. Ultimately, the goal of the TICO-19 dataset is to disseminate accurate technical information on symptoms, testing and treatment of the virus in the widest range of languages possible, especially low-resourced ones, and in a timely manner. Moreover, authors of TICO-19 encourage public and private collaborators to share any linguistic resource related to COVID-19, and to expand the translation of the benchmark to other languages[8].

1.3.3. The benchmark

The 38 languages translated from English and included in the TICO-19 benchmark are organized by Anastasopoulos et. al (ibid.) in three categories, namely pivot (9), priority (21) and important (8). The first category refers languages that are commonly used in different parts of the globe as lingua franca, whereas the label "priority languages" identifies a series of African and Asian languages most heavily requested by the partners of Translation Without Borders for translation work during the pandemic. The remaining languages falling under the third category include other and South and South-Eastern Asian languages that were deemed relevant because of their large number of speakers and relevance as LRLs. Table 1 provides a summary with the complete listing of languages, organized per category.

[7]Anastasopoulos et al. (2020) also illustrate a rigorous Quality Assurance protocol in order to double-check translation quality; more on that in Section 1.2.1.1.

[8] Provided that collaborators follow the outlined QA procedures outlined in the introductive paper by Anastasopoulos et al. (2020).

Category	Total	Translation Languages
Pivot	9	Arabic (modern standard), Chinese (simplified) French, Brazilian Portuguese, Latin American Spanish, Hindi, Russian, Swahili, Indonesian
Priority	21	Dari, Nigerian Fulfulde, Hausa, Kanuri, Central Khmer, Kinyarwanda, Kurdish Kurmanji (Latin script), Lingala, Luganda, Nepali, Nuer, Oromo, Pashto, Somali, Kurdish Sorani (Arabic script), Congolese Swahili, Ethiopian Tigrinya, Zulu
Important	8	Bengali, Burmese (Myanmar), Farsi, Malay, Marathi, Tagalog, Tamil, Urdu

Table 1 - Translation languages covered by TICO-19 as classified by Anastasopoulos et al. (2020)

Texts selected for the creation of the TICO-19 benchmark are open-source documents that were first issued in English and then translated in the 38 languages of the project. All documents are related to some aspect of the COVID-19 pandemic but come from different and diverse sub-domains to deliver "diversity, relevance and quality" (ibid., p. 3) into the finalized resource. The total number of selected documents amounts to with the 30 and sub-domains are represented as follows:

- **Medical-scientific:** 6 COVID-19 related articles from PubMed
- **General:** 15 articles from the English Wikipedia about specific aspects of COVID-19
- **News:** 6 entries on COVID-19 from Wikinews
- **Travel:** 1 article on travel restrictions during the pandemic on Wikivoyage
- **Announcements**: 2 entries from Wikisource – a public executive order to the population of the State of California and a communiqué to the WikiMedia staff
- **Medical-conversational:** 141 phrases and sentences in the medical domain with explicit mentions to COVID-19 keywords[9], collected by Carnagie

[9] Said keywords are the ones contained in the COVID-19 terminologies provided by Facebook and Google to the collaborators of TICO-19 (see 1.3.5. for more)

Mellon-University (CMU)[10][11]

From the abovementioned documents, a total of 3,071 sentences was selected for building the dataset. In their turn, sentences were proportionally distributed into a testing set (2100 sentences) and a development set (971 sentences).

1.3.4. Translation Quality Assurance

A thorough Quality Assurance (QA) process for translations was implemented by Anastasopoulos et al. (ibid., p. 5) with the aim of obtaining parallel data of the highest possible quality for automated translation output. Adding such a step mainly stemmed from the need for professionally translated data enforced by robust quality checks, which is particularly crucial when translating from and into LRLs (Guzmán et al., 2019). The whole process adopted by TICO-19 contributors can be summarised and broken down in the following phases:

1. **Translation**: Each document was delivered to professional Language Service Providers (LSPs) and translated by human translators.
2. **First editing**: Sentences were reviewed by experts, particularly those with medical knowledge when available. Discrepancies between translators and editors were resolved through arbitration.

    2.1. **Second review (high priority data only):** A subset of the data was reviewed a second time to ensure precision. In this context, content from PubMed was prioritized due to its complexity.

3. **Final Quality Assurance**: Translations were iteratively refined until every language achieves a quality rating exceeding 95%, with several low-resource languages requiring multiple translation rounds to meet said rating. Any remaining errors were corrected before final release of the dataset.

    Once again, the PubMed subdomain was identified as particularly challenging due

[10] Original dataset available at: http://www.speech.cs.cmu.edu/haitian/text/ [Last accessed 03/12/2024]
[11] The CMU Haitian-Creole dataset was initially developed under the EU- and NSF-funded project NESPOLE!; Moreover, the dataset was used in the immediate aftermath of the Haiti earthquake to quickly build and deploy statistical MT systems. One of the related projects for MT development was the aforementioned Project 4636 (Lewis, 2010).

to the occasional lack of medical expertise by some translators, often leading to inaccurate translations.

4. **Error categorization and dataset release**: The validated dataset included detailed annotations for detected errors. The latter were categorized into Addition/Omission, Grammar, Punctuation, Spelling, Capitalization, Mistranslation, Unnatural Translation, and Untranslated Text. Each error is also classified by severity: minor, major, or critical.

   The annotated dataset is finally released to support research in automatic quality estimation and post-editing, especially for under-resourced languages.

Authors of the TICO-19 encourage all future collaborators to translate the dataset by strictly following this QA process.

1.3.5. Other resources provided

The resources made available through the TICO-19 website by their collaborators (Anastasopoulos et al. 2020) are not limited to the translated sentence dataset alone, but also extend to translation memories, two translation terminologies by Facebook and Google, and additional translation material offered by Translation Without Borders and other project partners.

A total of 102 translation memories was obtained by converting the content of the translation dataset into .TMX files. Currently, 36 are available for English-to-X translation directions, 66 others that do not include English to accommodate potential needs of local populations (e.g., Aramaic-Ohromo, Hindi-Urdu, Kurdish Kurmanji-Kurdish Sorani). All of them are free for use.

As regards the simple terminologies on Covid-19 provided by Facebook and Google, a breakdown of their content is summarised in Table 2

	Facebook Terminologies	Google Terminologies
Terms	364	300
Language combinations	92[12]	127[13]
File format	Simple txt, 1 term/line, parallel format[14]	Simple csv, 1 term/line, parallel format
License of use	Free for use	Free for use

Table 2 - Terminologies' content

Other resources that can be obtained through the website of the initiative include glossaries[15], multilingual[16] and monolingual in-domain corpora[17]. Among these other resources, particularly notable are the MT developer datasets[18] developed by the NeuLab group at the Carnagie Mellon University, consisting in a collection of monolingual, comparable, back-translated and parallel data scraped from Wikipedia entries and news agencies issuing multilingual information. Although available online, TICO-19 founders mark that some of these developer data are subject to more rigid copyright restrictions, making them not suitable for developing commercial MT systems (Anastasopoulos et al., 2020).

1.3.6. Baseline achievements and benchmark applications

Once all translation data were at hand, the dataset was used by the TICO-19 collaborators (ibid.:7) to obtain baseline machine translation results across some of the possible language directions, using both pre-trained and newly trained MT systems. Specifically, the major pre-trained MT systems were provided by the OPUS-MT (Tiedemann and Thottingal, 2020), which were mostly used for a big proportion of the English-to-X, X-to-English, and French-to-X language pairs. Additionally, performance was compared

[12] all English-to-X
[13]100 English-to-X, 27 X-to-English
[14] i.e., for the Facebook Terminologies the term in each line is the same in every file. For the Google Terminologies, the source and target languages are aligned and always in the same file.
[15] e.g., the TWB glossary https://translatorswithoutborders.org/resource/twb-covid-19-glossary/ [Last accessed 04/12/2024]
[16]e.g., the EMEA corpus, accessible on OPUS (Tiedemann, 2012)
[17]e.g., the COVID-19 corpus from Open Research Dataset (CORD-19), available on Sketch Engine: https://www.sketchengine.eu/covid-19-corpus/ [Last accessed 04/12/2024]
[18] Available at: [Last accessed 04/12/2024]

with Fairseq (Ng et al., 2019) and other models (Bojar et al., 2018; Ott et. al, 2018) for the directions English-Russian, English-French and English-Chinese, respectively[19]. On the other hand, newly trained systems were implemented for language directions involving English and several LRLs, with OPUS corpora (Tiedemann, 2012) and TED talks datasets (Qi et al., 2018) functioning as main training data.

A major limitation of the experiments was posed by the absence of results for some ten LRLs, further underscoring the divide between low- and high-resource languages (HRLs) in terms of resource coverage. For all languages that were tested, results were measured by means of the BLEU score (Papineni, 2002) both for the performance in the language direction overall and in each subdomain.

The baseline results of the benchmark made the disparity between HRLs and LRLs even more evident, while also pointing to the impact of the type of training data on final scores. Additionally, it was noted that translation pairs with HRLs achieved relatively competitive results, whereas those with LRLs, even when partially supported by parallel data, delivered extremely low performances (Anastasopoulos et al., 2020). In terms of domain-specific scores, datasets from Wikipedia and news sources often yielded the highest results, and surprisingly, PubMed texts delivered more than acceptable quality evaluation scores. However, the authors of the initiative emphasize the need to analyse results alongside statistical significance tests, as each subdomain comprises a smaller test set compared to the full dataset (ibid., p. 7).

Despite the limitations encountered during the benchmarking process, recent research has demonstrated the value of the TICO-19 benchmark as an essential tool for testing multilingual language models (Alves et al., 2024; Mohammadshahi et al., 2022), improving NMT performance for LRLs (Ko et al., 2021; Öktem et al., 2021), and, more recently, facilitating adaptive MT through Large Language Models (Moslem et al., 2023a; Soudi et al., 2024).

1.4 Summing up

This first Chapter has provided a thorough introduction to the research context, focusing

[19] Note how pre-trained systems alternative to OPUS-MT were used for high resource language pairs. Almost all of translation directions involving English (or French) and other high-resource languages (HRLs) were tested on pre-trained systems.

on Crisis Translation and its implications in emergency scenarios. Given its relatively recent emergence within Translation Studies, this field still lacks a fully standardized definition, yet its significance in global crisis management continues to grow (O'Brien, 2018; O'Brien, 2022). A central aspect of this discussion was the role of Machine Translation (MT) in crisis response, with past initiatives like Mission 4636 (W. Lewis, 2010; W. Lewis et al., 2011) and TICO-19 (Anastasopoulos et al., 2020) illustrating its potential and limitations. Ethical concerns were also explored, highlighting the necessity of balancing technological advancements with linguistic inclusivity and responsible data management (Parra Escartín & Moniz, 2019). More recently, the role of anticipating the creation of in-domain language resources has been advanced as a potential tool to foster crisis preparedness (Federici, 2023). The Chapter also examined how multilingual information was disseminated during the COVID-19 pandemic, revealing however that translation still plays a minor role in crisis management frameworks (Civico, 2021). One exception was provided by the TICO-19 project, which invested huge efforts create a dataset for MT development in various low-resourced languages to spread vital information to vulnerable communities. Having fully outlined the background of TICO-19, Crisis Translation and Crisis MT, the upcoming Chapter will delve deeper into the theoretical and technical aspects of these two technologies, with a special focus on their most recent performance in domain adaptation settings.

2. Neural Machine Translation, Large Language Models: Foundations and Applications

2.1. Introduction

This second Chapter delves into the theoretical foundations and applications of two key translation technologies: Neural Machine Translation (NMT) and Large Language Models (LLMs). It first examines NMT (2.2.), detailing its state-of-the-art architecture (2.2.1.), the role of domain adaptation (2.2.2.), and evaluation frameworks used to assess translation quality (2.2.3.). LLMs (2.3.) will then be covered by the remainder of the Chapter, exploring their historical evolution (2.3.1.), training methodologies (2.3.2.), and classification based on architecture, modality, and purpose (2.3.3.). The chapter concludes with an examination of LLMs in crisis translation (2.3.4) and specialized translation settings (2.3.5), evaluating their potential benefits and limitations in domain-adaptive machine translation.

2.2. Neural Machine Translation

2.2.1. Fundamentals and State-of-the-Art architecture

Machine Translation (MT) is the task, usually carried out by computers and without human intervention, of translating automatically from one natural language to another. Since the first commercial MT systems were developed in the 1970s (Koehn, 2020), MT systems have advanced following different approaches. Neural Machine Translation (NMT) currently represents the state-of-the-art paradigm for MT engines (Tan et al., 2020) and falls into the category of data-driven (or corpus-based) approaches, meaning that engine architecture relies on large amounts of parallel corpora for its development and optimal performance.

At its core, an NMT system implements artificial neural networks, i.e., models inspired by the structure and functioning of the human brain. These networks consist of three types of layers: an input layer, one or more hidden layers and an output layer, each composed by basic units known as neurons or nodes. All nodes are interconnected across layers via weighted connections. As for the functioning of the neuron in general, it processes input values, applies an activation function, and produces an output. If the value

computed by a node exceeds the threshold defined by the activation function, the node activates and transmits the result to subsequent nodes. (Kotsiantis et al., 2007; Y. Wu & Feng, 2018). Architecture-wise, NMT engines are developed on the encoder-decoder paradigm (Cho et al., 2014), a sequence-to-sequence structure based on two sub-models. The first one, the encoder, implements a neural network to encode all tokens[20] from a given source sentence into numerical representations, or word embeddings (Mikolov et al., 2013). These embeddings, which capture semantic meaning in a multidimensional space, are then passed to the second component, the decoder. This one, in its turn, employs a neural network to decode the embeddings into the target sentence. Currently, the leading typology of neural network implemented for NMT is represented by Self-Attention Networks (SAN; Lin et al., 2017; Vaswani et al., 2017)

The first application of this particular paradigm was proposed by Bahdanau et al. (2016), who, among other things, also introduced the idea of attention mechanism, a technique through which algorithms place more focus on relevant information. In the context of NMT, the mechanism allows the model to dynamically compute a context vector for each decoding step. By focusing on specific parts of the sentence, the model can process source sentences of variable lengths without requiring a fixed-size representation. This ultimately improves its ability to handle long-range dependencies and produce context-aware translations.

Optimizations of the attention mechanism eventually led to the development of Transformer models (Vaswani et al., 2017). For Machine Translation, Transformer models architecture follow the encoder-decoder setup (Hapke et al., 2019, pp. 311–317) and are currently the leading architecture for developing NMT models (Tan et al., 2020; H. Xu, 2021). Figure 1 illustrates the typical setup for an encoder-decoder Transformer model.

[20]A token is defined as "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing" [taken from: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html - Last accessed 14/01/2025].

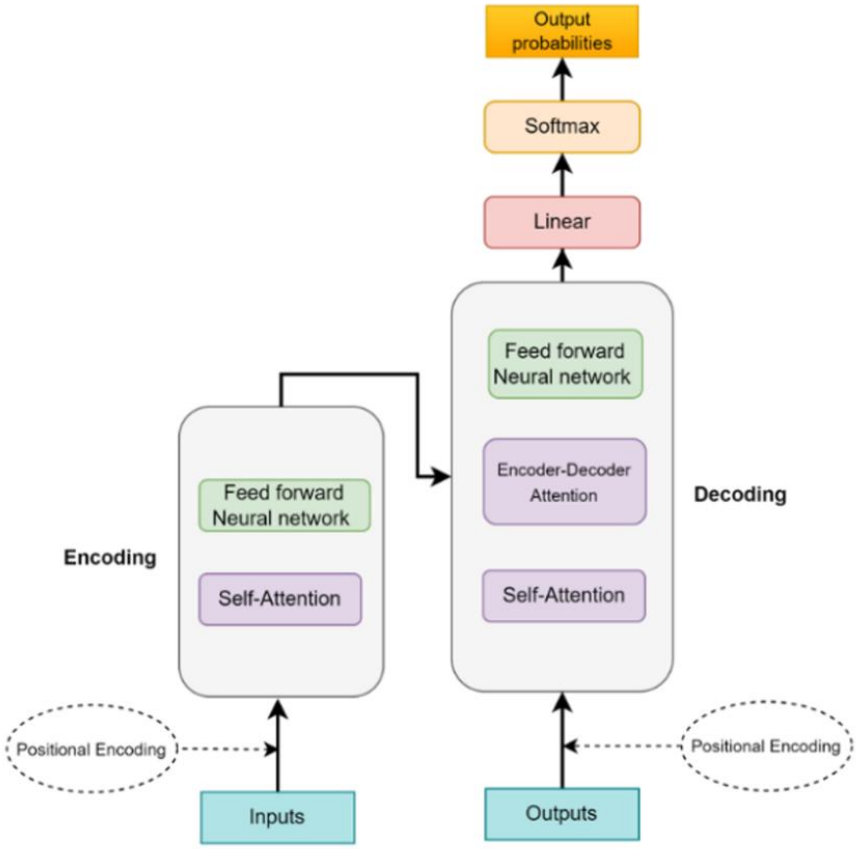

Figure 1 – The architecture of an Encoder-Decoder based Transformer (borrowed from Mubarak et al., 2023)

The most crucial innovation yielded by these models resided in integrating two additional sub-layers to both the encoder and the decoder: a self-attention layer responsible for running the attention mechanism multiple times in parallel, and a feed-forward neural network. In practical terms, the self-attention layer works so that:

> "Every time a new token is provided to the feedforward neural network, rather than building a single context vector out of the last encoder hidden state, this separate layer computes its relationship with all other tokens in the segment and selects which information is relevant for the current context vector." (Fernicola, 2022, p. 29)

To enhance learning, each sub-layer is surrounded by residual connections, which ensure that the original input to the given sub-layer remains accessible. These connections undergo a process called layer normalization, which helps stabilize the training process.

Additionally, to account for the word order in a sentence, the model uses positional encodings, i.e., patterns that are added to the embeddings. This design allows the Transformer to understand individual words in their positions within the sequence, and to handle long sentences better than older models.

Despite NMT having taken over as the leading approach for automated translation (Castilho et al., 2017) and achieving human-like quality in several language directions (Barrault et al., 2019; Bojar et al., 2018), the NMT paradigm still displays a significant margin for improvement on a general perspective. Most notably, output quality is contingent on the availability of large amounts of parallel data, as demonstrated with the cases involving low-resource languages (Anastasopoulos et al., 2020; Sennrich & B. Zhang, 2019; Jiajun Zhang & Zong, 2020). Additional critical requirements for NMT training data include the correct alignment of source and target sentences and removal of all noise to avoid performance deterioration (Koehn, 2020, pp. 298-307). Another significant problem concerns word semantics, as some words may have different translation equivalents depending on the domain (Koehn, 2020). In this regard, domain adaptation, which will be addressed briefly in the upcoming subsection, may be a method to overcoming the issue in specific cases.

2.2.2. Domain Adaptation strategies in MT

The notion of domain adaptation in the field of Machine Translation involves adapting an engine for automated translation to a task, usually to maximize system performance in a specialized domain (Koehn, 2020). In its turn, the concept of "domain" is here to be understood as:

> "a collection of text with similar topic, style, level of formality, etc. In practical terms, however, it typically means a corpus that comes from a specific source" (ibid., p. 239).

Another main objective of domain adaptation in MT is to avoid re-training the system from scratch, both for questions of computational power needed to train a system based on neural networks, and the fact that it is basically impossible to access the original MT training datasets. Generally, approaches for domain adaptive NMT fall into these two categories (Saunders, 2022):

- Data-centric – focus is placed on collecting or expanding in-domain parallel

datasets (either by selecting them manually or by generating them) for domain adaptation.

- Architecture-centric – which mainly involve adding trainable parameters to the NMT model, like a single new layer in the network, a new subnetwork or a domain discriminator.

At the time of writing, domain adaptive MT can be performed on the fly with several commercial MT systems. These include (but are not limited to) ModernMT[21], Unbabel[22], Language Weaver by RWS[23], AWS Amazon Translate[24] and Tilde MT[25].

Case studies on data-centric domain adapted MT (Contarino, 2021; Falsone, 2024; Rios et al, 2022) have thus far shown a remarkable increase in performance by NMT systems. To comprehensively evaluate general MT performance and the impact of domain adaptation strategies on translation quality, robust evaluation metrics are essential, as they enable consistent comparisons of system performance across various setups and domains.

2.2.3. MT Evaluation metrics and frameworks

According to Chatzikuomi (2020), MT Quality Evaluation tends to follow the same criteria for defining human translation quality, namely:

> "(i) fluency in the target language, which includes grammaticality and naturalness; (ii) adequacy as in semantic and pragmatic equivalence between the source and the target text; and (iii) compliance with possible requester specifications" (ibid., p. 138).

Evaluation of output quality constitutes one of the most widely discussed sub-branches in MT research (Koehn 2020; Kocmi et al., 2021; L. Han et al., 2021) and is typically carried out either via automated or manual metrics.

Automated metrics can be further broken down in other two groups: string-based metrics (also known as reference-based metrics) and pre-trained metrics. Reference-based frameworks, as the name suggests, rely on the comparison between a human reference translation and the MT output (also referred to as hypothesis). Always within this the

[21] https://www.modernmt.com/ [Last accessed 04/02/2025]
[22] https://unbabel.com/on-the-fly-machine-translation-domain-adaptation/ [Last accessed 04/02/2025]
[23] https://www.rws.com/language-weaver/ [Last accessed 04/02/2025]
[24] https://aws.amazon.com/translate/ [Lasr accessed 04/02/2025]
[25] https://tilde.ai/machine-translation/ [Last accessed 04/02/2025]

group of string-based metrics, another important distinction is drawn between metrics based on edit distance and those based on precision and recall.

In metrics set up on edit distance, the score is computed on the minimum number of operations (e.g., insertion, elimination and substitution at the word or character level) needed to convert the MT output to the reference translation. Word Error Rate (WER; Nießen et al., 2000) was one of the first implementations of this approach, and one of its most popular and widely used extensions is Translation Error Rate (TER; Snover et al. 2006), which also considers shifts of word sequences.

String-based metrics that adopt precision and recall are calculated instead on the ratio of matching n-grams between MT output and the reference translation(s). Specifically, precision measures the proportion of correct n-grams in the MT output that also appear in at least one reference translation. Recall, on the other hand, assesses how many of the expected n-grams from the reference(s) appear in the MT output. In other words, while precision focuses on the accuracy of the MT output relative to itself, recall evaluates how much of the reference translation has been captured.

The BiLingual Examination Understudy (BLEU; Papineni, 2002), is by far the most famous string-based metric modelled on precision and recall and is the most widely cited in MT research (Marie et al., 2021). BLEU scores are precision-based and computed on matching n-grams (ranging from 1 to 4 grams) and a sentence brevity penalty factor. However, BLEU is notorious for its limitations: it fails to account for different sentence structure, penalizing translations that use synonyms or paraphrasing, struggles with morphologically rich languages, and scores do not always align well with human judgment (Kocmi et al., 2021; Koehn, 2020; Mathur et al. 2020, Way, 2018). Moreover, variations in text preprocessing and tokenization techniques can lead to inconsistent BLEU scores, making it challenging to compare results across different studies or systems. (Post, 2018). This latter issue ultimately lead to the creation of SacreBLEU (ibid.), a standardized tool for computing BLEU scores that enforces standardized tokenization. Additionally, SacreBLEU enhances comparability by outputting a version string that records evaluation parameters, ensuring that reported scores can be reliably reproduced.

Since most modern NMT engines rely on segmentation at the sub-word level, character-based metrics started gaining relevance for automatic evaluation (Way et al., 2018) and eventually came to show higher correlation with human judgments than word-

based metrics (Lardilleux & Lepage, 2017). Character n-gram F score and its variants (chrF; Popović, 2015), which adopt a combination of precision and recall, and CharacTER (W. Wang et al., 2016) are two prominent examples of character-based metrics.

The second macro-category of metrics for automated evaluation includes frameworks that leverage pre-trained neural models. Some may employ embedding similarity to score sentence similarity between hypothesis and reference, like in the case of BERTscore (T. Zhang et al., 2020). Others, instead, estimate the quality of MT output given the source text, the reference text, or both[26].

One such example is provided by the Crosslingual Optimized Metric for Evaluation of Translation (COMET; Rei et al., 2020), a neural framework for training MT evaluation models. It utilizes cross-lingual pre-trained language models to predict MT quality based on both the source text and reference translation. Besides, it learns from human assessments to provide a more accurate quality estimate. Notably, COMET has thus far achieved high correlation with human judgements and has been indicated as one of the best performing automated metrics. (Freitag et al., 2021; Kocmi et al. 2021).

Despite limitations posed by BLEU, an impressive number of studies used the metric as the main tool for assessing system qualities (Marie et al., 2021). To ensure replicability and robustness of MT automatic evaluation, Kocmi et al (2021), and Marie et al. (2021), have proposed a series of recommendations, such as.

- Use a pre-trained metric as the main automatic metric, if applicable
- Use a string-based metric other than BLEU as a secondary metric, or as the main one for languages unsupported by pre-trained metrics.
- If BLEU scores are needed, compute them with the SacreBLEU tool
- Carry out a significance tests on automated metric scores to ensure that differences between two scores are not coincidental.

Despite automatic MT evaluation being able to deliver swift assessment of MT output, its major pitfall lies in its limited capability of capturing error granularity, semantic and syntactic equivalence (Castilho et al., 2018; L. Han et al., 2021). In this regard, manual evaluation can offer deeper insight on linguistic performance of the given system(s). Some of the most known frameworks for human assessment of MT output are

[26]For the sake of precision, the task of estimating MT output quality without a human reference is better referred to as Quality Estimation, rather than Quality Evaluation.

at sentence level and include evaluating output based on adequacy[27] and fluency[28], ranking, and direct assessment (DA). Adequacy and fluency are usually evaluated on a 5-point Likert scale (Chatzikuomi, 2020; Koehn, 2020). With ranking, output from different MT systems (generally two or three) is compared against each other on a sentence-by-sentence basis (Chatzikuomi, 2020; Koehn, 2020). Finally, direct assessment involves assigning a score to the MT output sentence on a continuous scale, usually in the range 0-100. Used in the past to evaluate both fluency and adequacy, it is now more focused on the latter (Bentivogli et al., 2018). Manual evaluation, however, does not come without its own share of drawbacks: it is a complex and slow procedure where annotator agreement can pose serious setbacks (Castilho et al. 2018; L. Han et al. 2021; Popović 2018).

While traditional MT evaluation methods provide valuable insights into system performance, the rise of LLMs has introduced new challenges and opportunities in assessing translation quality. Unlike conventional NMT systems, LLMs can rely on contextual learning (Brown et al., 2020) and can generate highly variable outputs, raising important questions about evaluation metrics and their effectiveness in this new paradigm (Briakou et al., 2024; Kocmi et al., 2024b). Before delving on their most recent performance in translation tasks, the upcoming Section of this Chapter will introduce their history, their major characteristics and observations on their implementation in translation (both specialized and in crisis scenarios) so far.

2.3. Large Language Models

2.3.1. Historical evolution of Language Modelling

What comes by the name of language model refers to a generative probabilistic model for natural language. Specifically, a probability model is designed based on large amounts of training data to predict the next word(s) inside a given sentence and to finally generate a string in a natural language. LLMs currently represent the state of the art for language models, but their advancement was built upon prior improvements in Deep Learning

[27] Also referred to as accuracy, adequacy may be defined as "the extent to which the translation transfers the meaning of the source-language unit into the target." (Castilho et al.2018, p. 9)
[28] Fluency may be defined as "the extent to which the translation follows the rules and norms of the target language (regardless of the source or input text)" (ibid.)

techniques and the increased availability of training data and computational resources.

The foundational work of Russian mathematician Andrej Markov in the early 1900s laid the groundwork for modern language modelling. In 1913, he applied his theories to assign probabilities to letter sequences inside novel in verse *Eugeny Onegin* by Aleksandr Puškin (Basharin et al., 2004). Decades later, significant advancements were delivered in the 1980s by a language research group at IBM, which applied probabilistic models to improve automatic speech recognition technology (Jelinek, 1990; Jelinek, 1998; Rosenfeld, 2000). These earliest studies drew on the mathematical framework introduced by Shannon (1949) for quantifying information and modelling probabilistic communication systems. This framework introduced key concepts like entropy, which measures the unpredictability of information, and information theory, which explains how data can be encoded and transmitted. It also offered the groundwork for n-gram models, which use probabilities to predict the next word in a sequence based on preceding words.

Building on these advancements, the evolution of language models can be organized into four key phases, as highlighted in several reviews (Hadi et al., 2023; Minaee et al., 2024; Zichong Wang et al., 2024). These phases correspond to specific families of language modelling approaches, namely: 1) Statistical Language Models (SLM), 2) Neural Language Models (NLM) 3), Pre-Trained Language Models (PTLM) and finally 4) Large Language Models.

Statistical Language Models marked the first milestone in the history of language modelling, leveraging probabilistic methods to predict and generate natural language. The core principle of SLMs is expressed mathematically as:

$$\Pr(s) \stackrel{\text{def}}{=} P(w_1 \ldots w_n) = \prod_{i=1}^{n} P(w_i|h_i)$$

Here, the probability *Pr(s)* of a sentence or sequence *s* is calculated as the joint probability of its words $P(w_1, w_2, \ldots, w_n)$ appearing in a specific sequence. Specifically, $P(w_i|h_i)$ denotes a conditional probability of a word w_i occurring, given its context or history h_i (which includes all the words preceding a given word w_i). $P(w_i|h_i)$ is iteratively calculated for each word in the sequence, starting from w_1 and ending at w_n. For instance, in the sequence "I like trains," the probability of "trains" depends on the words "I like".

This is represented by P("trains"| "I like") where "I like" forms the context h_3 for w_3. The equation should thus look like the following:

$$P(\text{"I like trains"}) = P(I)\ x\ P(like \mid I)\ x\ P(trains \mid I\ like)$$

In this way, the model determines the likelihood of a sequence by multiplying the conditional probabilities of each word, based on patterns learnt from the training data.

All throughout the SLM era (Bengio et al., 2003; J. Gao & Lin, 2004; Juang & Rabiner, 2005; Jelinek, 1998; Rosenfeld, 2000), practical implementation of this probabilistic framework often relied on n-grams, i.e., sequences of *n* elements, typically words in a sentence. N-gram models try to predict a word using a reduced history size, usually by examining groups of 2 (bigram models) or 3 words (trigram models), offering a computationally feasible solution for language modelling at that time. However, as research progressed into the new century, statistical approaches to language modelling faced increasingly critical questioning, with regards to their inability to capture semantic and syntactic similarities (Bengio et al., 2003; Juang & Rabiner, 2005).

Neural Language Models (NLMs) were therefore presented as a more robust alternative to SLMs, eventually establishing a new paradigm in language modelling (Bengio et al., 2003). Basically, NLMs combine neural networks and word embeddings to better compute the probability of word sequences. What eventually advanced the setbacks posed by SLMs, was the type of vector representation implemented, that is, a distributed representation. This means that each element in a word embedding represents a numerical value other than just 0 and 1, enabling a better conceptualization of words as related entities. Neural networks further enhance the capabilities of word embeddings by learning said representations and capturing their shared patterns. Ultimately, all the combined factors greatly facilitated semantic similarity comparisons and allowed for effective generalization of probabilities beyond word sequences that were absent during the training (ibid.), with the most implemented architecture within this approach being Recurrent Neural Networks (RNNs; Hadi et al., 2023; Kombrink et al., 2011; Mikolov et al., 2010). Nonetheless, the most significant drawback of neural networks in general is their need for substantial amounts of high-quality, labelled data, which posed a major hinderance in the early 2000s still due to limited computational resources and capabilities available.

As a result, the focus of the AI and NLP communities began intensifying research efforts towards "how to train effective deep neural models for specific tasks with limited human-annotated data" (X. Han et al., 2021, p. 2). There were several pivotal advancements that defined the turn to the era of Pre-Trained Language Models (PTLMs) in the mid-2010s. The first of these is the concept of transfer learning (Thurn and Pratt, 1998; Pan & Yang, 2010), which can be defined as the transfer of previous knowledge and experience acquired in older tasks to accomplish new ones. Most importantly, the shift to this third stage in language modelling research was prompted by the introduction of the attention mechanism (Bahdanau et al., 2016), and transformer models (Vaswani et al., 2017), bringing the possibility of enhanced performance to the table. PTLMs implement neural networks as well, but the word representations they are trained on also consider the context in which a given word appears (Peters et al., 2018). Their development begins with a pre-training phase, where models acquire knowledge representations through self-supervised learning (De Sa, 1993). Concretely, the model learns from unlabeled data by predicting parts of the input based on other parts, eliminating the need for manual labeling. For instance, a model might predict missing words in a sentence, uncovering patterns and structures within the data as pre-training progresses. Information gained during pre-training is later transferable to downstream tasks (Gui et al., 2024). In the case of NLP, these acquired patterns mostly involve representations related to syntax, semantics, and common facts, which can be leveraged for assignments such as classification, text summarization and information retrieval.

Optionally, models can then be optimized (fine-tuned) on smaller, task-specific datasets to enhance their general language understanding. PTLMs come in a variety of transformers-based architectures, such as encoder-only (one famous example being BERT; Devlin et al., 2017), encoder-decoder (e.g., XLM-R; Conneau et al, 2020) and decoder-only (like GPT 1 and 2; Radford et al., 2018; Radford et al., 2019). However, early PLM also presented a series of drawbacks; for instance, Bender and Koller (2020) argued and illustrated that PLMs, only learn a reflection of meaning and not meaning per se.

As it will be discussed hereafter, the shift to Large Language Models was driven by key factors such as the increase in hyper-parameters, the larger size of training datasets, and innovations in fine-tuning techniques

2.3.2. Large Language Models: design and development

Large Language Models (LLMs) are large-scale language models mainly built upon a Transformer architecture (Vaswani et al., 2017) which are generally trained for token prediction. As mentioned prior, one significant feature that sets LLMs apart from PTLMs is their scope: most importantly, the number of hyper-parameters in LLMs can reach up to hundreds of billions, requiring training datasets often as large as few TBs. The broader extent of parameters and training data volume allows LLMs to achieve optimal performance in a wide variety of complex tasks.

As for the training of LLMs, the pre-training process is not dissimilar from the one of PTLMs. Fine-tuning, where applied, is generally more complex depending on the desired target task, with differences arising mainly in learning objectives and optimization techniques (Sun & Drezde, 2025).

LLMs are pre-trained on large-scale text corpora from sources such as web pages, conversations, and books. Figure 2 illustrates and summarizes the typical data collection pre-processing pipeline. Typically, before training, data undergoes preprocessing to filter out noise, remove duplicates, and mitigate biases. Tokenization then converts text into a machine-readable format. The process also requires data scheduling (Zhao et al., 2024) to balance different sources and optimize learning. Finally, the model is trained using self-supervised learning, with its final performance depending on both architecture and training objectives (Lu et al., 2024; Zhang et al., 2022)

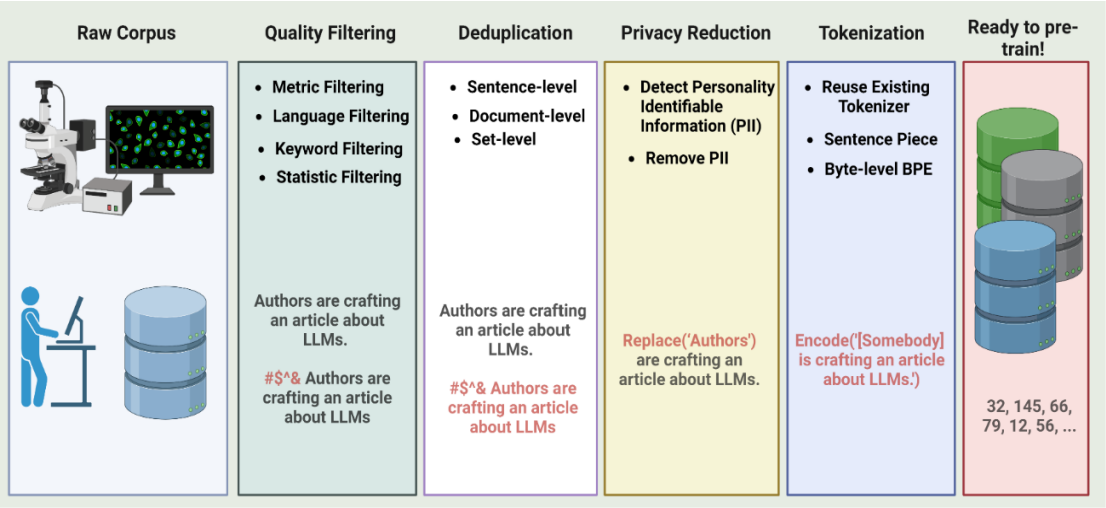

Figure 2 - The data pre-processing for LLMs (borrowed from Hadi et al., 2023)

As for fine tuning, beside the above-mentioned transfer learning, some of the most common include instruction tuning (Chung et al., 2024) and alignment tuning (Ouyang et al. 2022; Ziegler et al., 2020). For the first technique, the model is fine-tuned on instruction and input-output pair. The second technique involves asking the model to generate unexpected responses. Subsequently, based on the output, model parameters are adjusted to avoid harmful, biased or false output (Ziegler et al., 2020; Zhao et al., 2024). A particular widespread alignment technique is called Reinforcement Learning with Human Feedback (RLH; Ziegler et al., 2020), where human-annotated responses guide the model's training.

Once fully developed, LLMs are rigorously evaluated to measure their performance, verify alignment with user needs and ensure appropriate outputs. These evaluations focus on two broader main areas: Natural Language Understanding (NLU) and Natural Language Generation (NLG). Within this latter field, machine translation is a key area used to evaluate the language generation capabilities of LLMs (Naveed et al., 2024; Zhao et al, 2024). Specialized benchmarks are designed for testing the models on a given task; for automated translation, some of the most important benchmarks are represented by WMT (Bojar et al., 2016) and WMT20 (Barrault et al., 2020).

All in all, what may be regarded as a major novelty introduced by LLMs lies into the notion of in-context learning (ICL; Brown et al., 2020), i.e., the ability to learn unseen task by leveraging contextual information, even without undergoing further fine-tuning (ibid.). Moreover, once the LLM is ready for deployment, users can interact with it by *prompting* it, i.e., querying the model to carry out a specific task. Users might also adopt few-shot prompting, i.e., they may provide contextualized examples (or *shots*) for the desired assignment. The only limitation to the number of shots is bound the *context window* of each LLM (Agarwal et al., 2024), namely the number of tokens that can be processed in input by the model. In other words, LLMs can be prompted without prior examples (zero-shot prompting) or with a virtually large number of examples (many-shot prompting).

A more refined prompting strategy includes instructing the LLM as to how it should reason while solving the task. In one of these approaches, known as Chain-of-Thought (CoT), demonstration includes additional reasoning information (J. Huang & Chang., 2022; Kim et al., 2023; Wei et al., 2023) beside instructions on the task itself. Figure 3 compares examples of standard zero-shot prompting and CoT prompting.

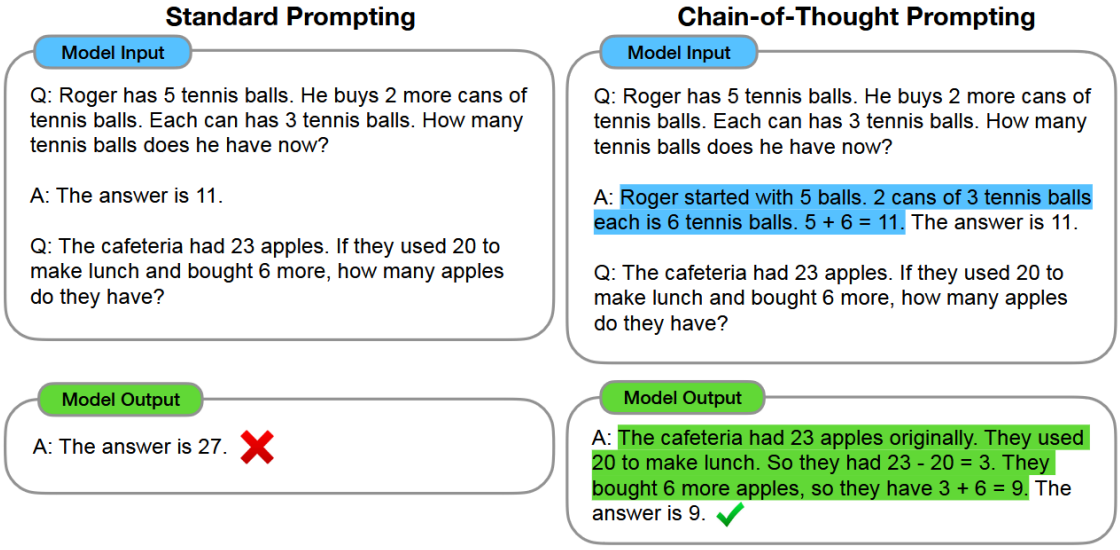

Figure 3 - Standard zero-shot and CoT prompting, examples for arithmetical operations (borrowed by Wei et al., 2022)

Generally speaking, it is essential to understand that the design and capabilities of LLMs can also vary significantly based on their architecture and other factors, which we will discuss below.

2.3.3. Taxonomy of LLMs – Architectures, modalities and applications

LLMs can be grouped based on architecture, modality, training languages, purpose, and overall availability. Architecturally, most LLMs are built on the Transformer model (Vaswani et al., 2017), with three main variants: causal decoders (e.g., GPT-3; Brown et al., 2020), prefix decoders (e.g., GLM-130B; Zeng et al., 2022), and Mixture of Experts (MoE),

Modality-wise, LLMs can be unimodal, focusing on text (e.g., GPT-3), or multimodal (MLLMs), meaning they capable of processing diverse formats like text, images, and audio (e.g., Kosmos-1; S. Huang et al., 2023).

Training language(s) further differentiate LLMs into monolingual, bilingual, and multilingual models, with BLOOMZ and mT0 (Muenninghof et al., 2022) representing prominent examples for multilingual LLMs. Purpose-wise, general-purpose models like T5 (Raffael et al., 2020) take on a broader range of everyday tasks, while domain-specific models such as Galactica (Taylor et al., 2022) and BloombergGPT (S. Wu et al., 2023) are tailored for specialized fields.

Availability is here to be intended as open or closed source. Typically, open source LLMs, such as LLM360 (Liu et al., 2023), are available to the public, allow for customization, come with no license or costs, rely on contributions of citizen developers and publicly disclose their pretraining dataset (Kukreja et al., 2024). These features do not apply instead to proprietary models like OpenAI's GPT series, or apply only partially (Tarkowski, 2023b).

Finally, as briefly mentioned prior, some LLMs perform well without fine-tuning (Brown, 2020), but pre-trained models often require optimization to enhance their usability in specific tasks and follow user intent more consistently (Ouyang et al., 2022). Examples include BLOOMZ and mT0, which were fine-tuned from BLOOM (Scao et al., 2022) and T0 (Sanh et al., 2021), respectively.

On the ethical side of things, a heated discussion is currently taking place on the correct development and implementation of modern-day language models (Bender & Koller, 2020; Weidinger et al., 2021). The responsible use and development of LLM should also address its soaring financial costs. It has been recently estimated that development can exceed hundreds of millions of US dollars (Cottier et al., 2024). With the next generation of LLMs, expenses could potentially rise to billions of dollars (Smith, 2024).

The expensiveness of LLM development also mirrors significant environmental impacts. In this sense, all parts of the supply chains are affected. To name a few issues, required hardware is energetically demanding (Crawford, 2021), fresh water is constantly needed to cool data centres hosting computations (Mytton, 2021; Patterson et al., 2021), and training is associated with high deployment of resources and carbon emissions (Bender et al., 2021; Patterson et al., 2021; Schwartz et al., 2020).

Besides, poorly curated training datasets often carry the most significant consequences on generated output. When trained on poorly curated datasets, LLMs can cause private data leaks, be leveraged to spread misinformation, and produce harmful or biased output (Bender et al., 2021; Ouyang et al., 2022; Yao et al., 2024).

In context where technology is adopted in aid of marginalized groups, such as Crisis Translation, the impact of LLM output toxicity is more relevant than ever. The next subsection will briefly discuss how and to what extent LLMs have been implemented for translation in crisis scenarios.

2.3.4. An overview of LLMs in crisis management and Crisis Translation

As for their implementation in crisis management frameworks, and particularly as a translation tool in crisis-ridden areas, potential and limitations of LLMs remain widely understudied at the time of writing. In fact, the topic appears to constitute a recent, niche branch of LLM research, with contributions focusing mostly on how to integrate LLMs or AI in general in emergency management frameworks (European Commission. Joint Research Centre, 2024; Otal et al., 2024). The most relevant work assessing translation capacities and, consequently, discussing suitability of LLMs for multilingual information access, has been provided by Lankford et al (2024). By using the datasets from the LoResMT2021 shared tasks (Ojha et al., 2021), their study compared performances of GPT-3.5, GPT-4 (OpenAI, 2023), NLLB-200 (Costa-jussà et al., 2022) and a custom GPT in both their pre-trained and fine-tuned instances. Results suggested that while the custom GPT swiftly delivers a functioning MT system, the fine-tuned multilingual LLM has consistently demonstrated SOTA performance, and could therefore turn out more useful in the long run for crisis management. A final important consideration is also dedicated to the limitations posed by proprietary models such as GPT-4:

> "when fine-tuning these models for specific tasks, there is a risk of overlapping data that cannot be easily identified or removed. This limitation underscores a broader issue within the field of NLP and MT research, where the exact composition of training data in SOTA models often remains opaque." (Lankford et al., 2024, p. 9).

These findings could suggest open-source LLMs as a new frontier for better multilingual information access in crisis scenarios, particularly due to transparency in dataset composition and parameter disclosure. However, while research on LLM-based MT in emergency contexts is still in its early stages, studies on general translation capabilities of LLMs have already provided valuable insights. The upcoming subsection will examine these aspects in greater detail.

2.3.5. LLMs in Specialized and Domain Adaptative Translation

The last two years have witnessed a sharp increase in studies regarding the translation capabilities of LLMs, both in generalized (Hendy et al., 2023; Jiao et al., 2023; W. Zhu et

al, 2024) and specialized settings (Eschbach-Dymanus et al., 2024; D. Gao et al., 2024; Wassie et al., 2024). Overall, LLMs tend to perform on-par with commercial MT systems based on the results of automated metrics (Jiao et al., 2023, Son & Kim., 2023). On the other hand, findings of the WMT24 General MT Task (Kocmi et al., 2024b) illustrate that some LLMs ranked first on automated metrics, but were not the outright winners among human evaluators. Consequently, contributors of the findings resorted to manual evaluation as the ultimate judge for translation quality assessment (ibid., p. 20).

More focused studies on specialized translation have highlighted the need for model fine-tuning to achieve state-of-the-art performance (Eschbach-Dymanus et al., 2024; Lankford et al., 2024; Wassie et al., 2024). In fact, several studies have demonstrated the enhanced performance of fine-tuned models compared to base models and in-context learning (Alves et al., 2023; Eschbach-Dymanus; Wassie et al., 2024). However, model optimization is quite costly and may not be available to small and medium enterprises or for academic use.

In the eventuality of cost constraints, domain adaptive, LLM-based MT may still deliver acceptable results through in-context learning (Moslem et al, 2023a, L. Wang et al., 2023). Suggestions to improve translation output through in-context learning mainly involve integrating terminology (Moslem, 2023b), dictionary words (Ghazvininejad et al., 2023) and translation pairs similar to the new source text (Agrawal et al., 2023; Moslem et al., 2023a; Vilar et al., 2023).

Open source and open weight models could yield another valuable contribution in this sense, since the deployment of open-source LLMs is generally less expensive. Interestingly enough, LLaMa 2 by MetaAI (Touvron et al., 2023)[29] has emerged in several studies for its performance improvement in both ICL and fine-tuning settings (Aycock et al., 2023; Eschbach-Dymanus, 2024; D. Zhu, 2024). Further studies have shed light on versatility of this model: while its baseline results are overshadowed by larger models such as GPT 3.5. (Hendy et al., 2023; H. Xu et al., 2024), LLaMa-2 has outperformed similarly sized, explicitly multilingual LLMs (B. Zhang et al., 2023). Additionally, since it comes under a less restrictive permissive license of use, Llama-2 has been deployed to

[29]We advise not to use the term “open source” when referring to LLaMa 2 and instances of LLaMa 3 (Grattafiori et al., 2024), as their training datasets have not been disclosed at the time of writing (07/02/2025). This stands in contrast with the definition of Open Source AI. (Open Source Initiative, n.d.; Maffulli, 2023; Tarkowski, 2023a; Tarkowski, 2023b). In this case, the expression “open weight models” is preferable, since parameters (weights) and inference code are available at: https://github.com/meta-llama/llama [last accessed 13/02/2025].

develop more sophisticated tools, such as the translation model ALMA (H. Xu et al., 2024).

Nonetheless, LLMs appear to share limitations with NMT when it comes to the quality of translation output. Modern-day language models generally rely on high amounts of data for training, which impacts performance dramatically in low-resource settings (Robinson et al., 2023; Shu et al., 2024). Bawden & Yvon (2023) also illustrate that LLM performance in translation may also depend on the similarity with seen languages during pre-training (ibid., p. 7). Similar conclusions were reached by Diandaru et al. (2024).

Finally, a recent issue pointed out at WMT24 (Kocmi et al., 2024b) concerns the production of verbose output. Upon prompting, it has been observed that some models offer reasoning insights, provide multiple translations or come as far as refusing to translate altogether (Briakou, 2024). This tendency poses challenges for both automatic and human evaluation. Including verbose results in automated scoring could distort model ranking, with otherwise well-performing models being penalized (ibid.). One practical solution, adopted at WMT24 (Kcomi et al., 2024b) involved excluding segments triggering verbose output for any LLM from the automatic evaluation (ibid., p. 14). In human assessment, verbosity can lead to inconsistencies, with added explanations being misclassified as errors (ibid.). Building on this, Briakou et al. (ibid.) suggest refining LLM prompting strategies to reduce verbosity and adapting evaluation metrics to account for more context-aware outputs.

2.4. Summing up

The present Chapter has outlined the foundational principles and applications of Neural Machine Translation (NMT) and Large Language Models (LLMs), offering a comparative analysis of their respective strengths and limitations. Beginning with NMT, we explored its state-of-the-art architecture, the role of domain adaptation, and the various evaluation frameworks used to assess translation quality. The discussion then shifted to LLMs, tracing their historical evolution, training methodologies, and emerging applications in specialized translation contexts, including crisis translation. Overall, a key takeaway from this Chapter is the ongoing debate surrounding the effectiveness of LLMs in domain-specific scenarios. Specifically, most recent discussions have assessed their performance compared to adapted NMT systems (Moslem et al., 2023a), but others have

express concerns on the high financial and energetic costs of LLM development (Bender et al., 2021; Cottier et al., 2024; Schwartz et al., 2020), as well as to the tendency to verbosity (Briakou et al., 2024; Kocmi et al., 2024b) and produce harmful output (Ouyang et al., 2022; Yao et al., 2024). All the above considerations set the stage for the experimental framework presented in the following Chapter, where we will describe how the knowledge acquired so far has been put into practice in our experimental framework.

3. Experimental Framework

3.1. Introduction

This chapter outlines the experimental framework implemented to assess the effectiveness of domain-adaptive neural machine translation (NMT) and large language models (LLMs) in crisis translation. The study investigates whether LLM-driven translation can match or outperform a domain-adapted NMT system, with a focus on the English-Italian translation direction. After defining the research hypothesis and methodology, the chapter details the dataset selection process, preprocessing steps, and translation workflow. It also introduces the domain adaptation strategies used for both ModernMT and LLaMa 3.2, followed by an explanation of the evaluation metrics applied.[30]

3.2. Experimental setting

So far, Chapter 1 has discussed both potential and possible drawbacks of Machine Translation use in crisis management. With proper dataset curation and a strategic approach to managing language resources for Crisis MT, automated translation can play an active role in enabling swift multilingual communication, particularly for low-resource languages. Specifically, both Crisis MT development and effective crisis management would greatly benefit from treating resource management primarily as a means of crisis preparedness.

As discussed in Chapter 2, the rise of LLMs across various industry and academic fields, including Crisis Translation, has introduced new insights and approaches to technology development and artificial intelligence. A logical starting point for assessing their role in multilingual information access during crises is to compare their performance with that of traditional NMT systems on crisis-related datasets. In resource-constrained settings, adaptive MT may offer a time- and cost-effective solution for improving baseline system performance without further tuning. Therefore, also in consideration of recent results obtained by LLMs in specialized translation and on crisis translation datasets, we

[30] All project files and resources mentioned in this Chapter are available at the following GitHub repository: https://github.com/lucia-galiero/TICO-19_NMT_LLM [Last accessed 19/02/2025]

formulate the following hypothesis:

> ***Hypothesis:*** *Domain adapted, LLM-driven translation can achieve equal or better translation results than a NMT system that has, in its turn, undergone a domain adaptation process.*

To prove this hypothesis, we propose to adapt and evaluate a neural adaptive MT system and a Large Language Model. Our dataset will be represented by a couple of text from the academic-scientific subdomain in the TICO-19 benchmark. The examined translation pair will be English-Italian, for which the benchmark has not been released at the time of this thesis. More specifically, the work aims at answering the following research questions:

> ***Research Question n°1 (RQ1):*** *Do LLMs deliver better results both before and after domain adaptation?*

> ***Research Question n°2 (RQ2):*** *Does domain adaptation yield such an impact on the output to be necessary for future development of translation technologies in crisis scenarios?*

The following experimental setting will be implemented to address these research questions:

1. Select texts for human translation
2. Translate the documents and manually review them
3. Removing all possible sources of noise from the translated texts and splitting sentences into training, tuning and test set.
4. Carry out the domain adaptation for our chosen NMT engine and LLM
5. Perform both automated and manual evaluation for the output of each adapted system.

In this way, our work will provide not just a comparison between automated system, but it will also mark a first step to expand the TICO-19 benchmark for the pair English-Italian. Every phase of our proposed setting will be examined and broken down in the remainder of this Chapter, starting with the document selection and their translation.

3.3. Dataset selection and preprocessing

3.3.1. Text selection

In the proposed framework, our working material is represented by two among the 30 documents used for the TICO-19 dataset (Anastasopoulos et al., 2020). As mentioned in subsection 1.3.1., the benchmark has been translated into 38 languages, most of which fall in the category of low-resource languages. At the time of this thesis, the dataset has been not released in Italian yet, likely due to the prioritization of low-resource languages, which serve a larger number of vulnerable communities worldwide. Nevertheless, a complete evaluation of NMT and LLM capabilities would still require metrics based on human references for their scores, such as BLEU and chrF3. Therefore, the first fundamental step for our work is the manual translation of the documents of choice.

Additionally, as described in 1.3.3., the TICO-19 benchmark covers a wide variety of domains and further divides the dataset into a test and a development subset. In the light of the definition of domain adaptation illustrated in subsection 2.2.2., it was essential to choose two text belonging to the same sub-domain. Eventually, the choice fell upon two medical-scientific texts: one selected from the test subset, the other from the development set. Both documents are biomedical research papers indexed in PubMed and were chosen as they offered the largest number of segments inside each subset and for their good degree of register specialization. The first document (CDC COVID 19 Response Team, 2020) presents preliminary estimates on the prevalence of underlying health conditions among COVID-19 patients in the United States, analyzing how comorbidities influence the severity of the disease. The second article (Yi et al., 2020) provides a comprehensive review of COVID-19, including its epidemiology, virology, diagnosis, treatment, and prevention. To achieve optimal performance by the MT system and LLM of choice, it was essential to produce consistent translations, especially with regards to domain terminology. All these aspects will be discussed in the upcoming subsection, which will also be supported by some case examples.

3.3.2. Document translation – A few observations

Both documents were translated integrally on Trados Studio in their native .pdf format to

retain the largest amount of content possible. It is important to note that .pdf files are typically processed in Trados Studio through optical character recognition. The content of source texts is then transposed into .docx format, which also constitutes the standard for target files originating from .pdfs. Fortunately, no loss of content was registered at any stage.

Both documents targeted an expert audience, as the CDC report (2020) was published in the Morbidity and Mortality Weekly Report, while the COVID-19 review (Yi et al., 2020) appeared in the International Journal of Biological Sciences. However, the two documents were not comparable in terms of style and terminology, as each article presented a different type of analysis and was authored by distinct research teams. The first article (CDC COVID 19 Response Team, 2020) lists the results of a statistical analysis on underlying health conditions in COVID-19 patients and their impact on disease outcome. Additionally, it displayed a significant level of redundancy in syntactical sentence structure. This was not the case for the second article (Yi et al., 2020), which aimed at offering a systematic review of the disease, its symptoms, prevention and available treatment at the time.

Similar conclusions can be drawn from a terminological point of view. Most terms in the first text designated diseases, disorders or were associated with hospitalization. A different range of terms was observed in the second article instead, where denominations and acronyms for viruses, diseases, treatments, antibiotics, and proteins were far more widespread. To further analyze the terminological differences between the texts, keyword lists were later generated for each document using the corpus consultation platform SketchEngine (Kilgarriff et al., 2014)[31] covering both single-word and multi-word terms. EnTenTen21 (Lexical Computing CZ s.r.o., 2024)[32], an English corpus of texts collected from Internet containing 52 billion words, was used as a reference corpus. Table 3 and Table 4 display keywords list comparison for single word and multi word terms, respectively, and as reported on SketchEngine.

[31] https://www.sketchengine.eu/ [Last accessed 19/02/2025] – Please note that full access on SketchEngine requires a subscription is granted for students at the Department of Interpreting and Translation at the University of Bologna.
[32] More at: https://www.sketchengine.eu/ententen-english-corpus/ [Last accessed 21/02/2025]

CDC COVID 19 Response Team, 2020	Yi et al., 2020
mmwr	sars-cov
icu	sars
cdc	sars-cov-2
neurologic	mers-cov
non-icu	mers
hospitalize	cov
neurodevelopmental	wuhan
mellitus	ace2
hospitalization	cytokine
obstructive	tcm

Table 3 - Top 10 keyword list (single word terms) for each document, as reported by SketchEngine with EnTenTen21 as reference corpus.

CDC COVID 19 Response Team, 2020	Yi et al., 2020
underlying health condition	respiratory syndrome
underlying health	acute respiratory syndrome
underlying condition	cytokine storm
health condition	novel coronavirus
icu admission	covid-19 patient
severe outcome	severe acute respiratory syndrome
chronic lung disease	recovered patient
chronic lung	engl j med
case report form	incubation period
neurologic disorder	health-care provider

Table 4 - Top 10 keyword list (multi-word terms) for each document, as reported by SketchEngine with EnTenTen21 as reference corpus

However, as mentioned in 1.3.5., authors of the TICO-19 benchmark (Anastasopoulos et al., 2020) also include terminologies by certified partners (Facebook and Google) among resources for translators. To make the best out of the terminologies, a single bilingual glossary was obtained by using terms from both terminologies and

discarding duplicates, i.e., duplicate terms with the same translation in English and Italian. The resulting Excel file (.xlsx) with 607 terms was converted to a MultiTerm termbase for better interoperability with Trados. The termbase, however, did not prove particularly useful, since the original terminologies by Facebook and Google contained terms relative to general vocabulary about the disease, the epidemic, safety measures.

Regarding the translation per se, the task proved particularly challenging from a beginner's perspective, as the degree of specialization required extensive research and knowledge of the biomedical domain. Advanced translators with external support from experts in the biomedical field would be far more suitable for this kind of work. One notable translation challenge, however, involved the grammatical gender used in Italian to refer to COVID-19 as a disease. Since the word "disease" usually has an Italian equivalent in the feminine word "*malattia*", COVID-19 should be typically referred to with the feminine gender, i.e., "*la* COVID-19" and resulting declination of prepositions "*della* COVID-19", "*sulla* COVID-19" etc. However, as explained by the Academia Della Crusca (Giovine, 2020), while the feminine form was dominant in academic-scientific texts[33], the disease has been widely addressed in masculine form in informal and official contexts, such as government press releases and draft laws. In this sense, the use of the masculine is not regarded as incorrect, as long as its use is consistent in a given text. Similar observations were made by Treccani (Sgroi, 2020). All things considered, if the whole TICO-19 dataset were to be translated in Italian in the future, it would be advisable to maintain strict consistency across the whole dataset, which is the key for high-level translation output by automated systems. Therefore, bearing in mind the multi-domain nature of TICO-19 benchmark, which also includes texts featuring a less formal register (see 1.3.3), it was decided to adopt the masculine form for COVID-19.

Another significant issue, although concerning the CDC report alone (2020), was the translation of the word "condition" and especially its associated terms such as "underlying health condition". While the English word can refer to a state of health (Merriam-Webster Dictionary, 2025) or to a type of disease (Cambridge Dictionary, 2025), the first instances of the word "condition" seem indeed to indicate a set of diseases, as it can be inferred from the title of the report and some of the first introductory lines:

[33] We are here referring to the time of writing of the article by Giovine (2020), which first appeared in July 2020. At the time of this dissertation, no known comprehensive corpus-based review on the grammatical gender of COVID-19 in Italian has been reported.

> "Preliminary Estimates of the Prevalence of Selected *Underlying Health Conditions* Among Patients with Coronavirus Disease 2019 — United States, February 12–March 28, 2020" (ibid., p. 1)

> "U.S. older adults, including those aged ≥65 years and particularly those aged ≥85 years, also appear to be at higher risk for severe COVID-19–associated outcomes; however, data describing *underlying health conditions* among U.S. COVID-19 patients have not yet been reported." (ibid., p.1)

However, as 'risk factors' are mentioned shortly afterward, doubts arise about whether the word 'condition' also includes these factors when used on its own. Generally, having an underlying health condition when catching COVID-19 does pose a risk, but does the word 'condition' encompass other factors as well? The following example illustrates one such case:

> "As of March 28, 2020, U.S. states and territories have reported 122,653 U.S. COVID-19 cases to CDC, including 7,162 (5.8%) for whom data on *underlying health conditions and other known risk factors* for severe outcomes from respiratory infections were reported." (ibid., p.1)

Such distinction becomes even blurrier when the results of the analysis are broken down and described in a dedicated table, which is reported below in Figure 4.

TABLE 1. Reported outcomes among COVID-19 patients of all ages, by hospitalization status, underlying health condition, and risk factor for severe outcome from respiratory infection — United States, February 12–March 28, 2020

Underlying health condition/Risk factor for severe outcomes from respiratory infection (no., % with condition)	No. (%)		
	Not hospitalized	Hospitalized, non-ICU	ICU admission
Total with case report form (N = 74,439)	**12,217**	**5,285**	**1,069**
Missing or unknown status for all conditions (67,277)	**7,074**	**4,248**	**612**
Total with completed information (7,162)	**5,143**	**1,037**	**457**
One or more conditions (2,692, 37.6%)	1,388 (27)	732 (71)	358 (78)
Diabetes mellitus (784, 10.9%)	331 (6)	251 (24)	148 (32)
Chronic lung disease* (656, 9.2%)	363 (7)	152 (15)	94 (21)
Cardiovascular disease (647, 9.0%)	239 (5)	242 (23)	132 (29)
Immunocompromised condition (264, 3.7%)	141 (3)	63 (6)	41 (9)
Chronic renal disease (213, 3.0%)	51 (1)	95 (9)	56 (12)
Pregnancy (143, 2.0%)	72 (1)	31 (3)	4 (1)
Neurologic disorder, neurodevelopmental, intellectual disability (52, 0.7%)†	17 (0.3)	25 (2)	7 (2)
Chronic liver disease (41, 0.6%)	24 (1)	9 (1)	7 (2)
Other chronic disease (1,182, 16.5%)§	583 (11)	359 (35)	170 (37)
Former smoker (165, 2.3%)	80 (2)	45 (4)	33 (7)
Current smoker (96, 1.3%)	61 (1)	22 (2)	5 (1)
None of the above conditions¶ (4,470, 62.4%)	3,755 (73)	305 (29)	99 (22)

Abbreviation: ICU = intensive care unit.

* Includes any of the following: asthma, chronic obstructive pulmonary disease, and emphysema.

† For neurologic disorder, neurodevelopmental, and intellectual disability, the following information was specified: dementia, memory loss, or Alzheimer's disease (17); seizure disorder (5); Parkinson's disease (4); migraine/headache (4); stroke (3); autism (2); aneurysm (2); multiple sclerosis (2); neuropathy (2); hereditary spastic paraplegia (1); myasthenia gravis (1); intracranial hemorrhage (1); and altered mental status (1).

§ For other chronic disease, the following information was specified: hypertension (113); thyroid disease (37); gastrointestinal disorder (32); hyperlipidemia (29); cancer or history of cancer (29); rheumatologic disorder (19); hematologic disorder (17); obesity (17); arthritis, nonrheumatoid, including not otherwise specified (16); musculoskeletal disorder other than arthritis (10); mental health condition (9); urologic disorder (7); cerebrovascular disease (7); obstructive sleep apnea (7); fibromyalgia (7); gynecologic disorder (6); embolism, pulmonary or venous (5); ophthalmic disorder (2); hypertriglyceridemia (1); endocrine (1); substance abuse disorder (1); dermatologic disorder (1); genetic disorder (1).

¶ All listed chronic conditions, including other chronic disease, were marked as not present.

Figure 4 - The table with results in the CDC report (borrowed from CDC Covid 19 Response Team, 2020)

As it can be observed, all factors are displayed in the same list, without marking explicitly what is categorized as an "underlying health condition" or a "risk factor".

For the present case, using the Italian word "condizione" in translation would result in a calque rather than an accurate rendition of the source meaning. According to the definition by Vocabolario Treccani (n.d.), the word "condizione" can designate a state of physical being, but there is no mention to a specific disease or disorder. Therefore, to provide a clearer distinction between the concepts of "risk factors" and "condition", the latter has been translated with the word "patologia" in the target text. Consequently, "underlying health condition" has been translated as "patologia pregressa" in all its instances.

The translated documents in .docx format were outsourced for review by a

translator specialized in the academic-scientific field, and all corrections were promptly applied to the target texts. With the translation phase complete, the content of the documents was copied to two separate .txt files (one for language) in preparation for dataset cleaning. As a matter of fact, it is essential to ensure the highest level of quality in experimental datasets, since NMT models tend to be extremely sensitive to noise in training data (Chen et al. 2016; Koehn & Knowles 2017). The following subsection will briefly address how noise filtering was carried out, alongside dataset division for the experiments of this dissertation.

3.3.3. Dataset cleaning and splitting

Once both texts were reviewed, it was essential to remove all sources of noise. Based on different classifications (Contarino, 2021; Gupta et al., 2019; Khayrallah & Koehn, 2018), the most widespread types of noise observed in the documents were short sentences (mostly section titles and text in tables), non-alphabetical segments[34] (numbers reported in tables) and "do-not-translate terms" (DNTs; Gupta et al., 2019, p. 145)[35]. Upon closer observation of the TICO-19 benchmark, it was soon discovered that, for text of the scientific-academic subdomain, no segment was shorter than 5 words (mostly titles for each section of the text). Besides, elements such as bibliographic references, acknowledgements, headers, in text-citations and everything concerning tables were consistently omitted. All these noisy elements were removed manually since they were easily identifiable. Hyphenated words, when signaling word splitting and not an acronym, were reverted to their full form. Out of the original 1377 sentence pairs, the cleaned dataset contained 339 segments for each language.

After that, the data set has to be split into test, tuning and training sets, as is typical in machine learning experimental frameworks (Zafar et al., 2018). All sentences were aligned into a .tsv file, following the data format implemented for the TICO-19 benchmark. The subset building was achieved via a Python script: as a first step, sentences

[34] The expression refers to text fragments composed primarily of punctuation, digits, or whitespaces, which may be unhelpful or even detrimental to system performance (Gupta et al., 2019). Such segments are typically filtered out if they contain only non-alphabetical elements, exceed a predefined ratio of non-alphabetic to alphabetic characters, or display a significant imbalance between source and target sides (Riekters, 2018)

[35] Among other things, DNTs mostly involve "email addresses, URLs, numbers with two or more digits (without comma and dot), and any combination of number (at least two digits) and English characters" (Gupta et al., 2019)

shorter than 5 words and those longer than 80[36] would be discarded entirely from the selection. From the 338 sentences that were processed, a test set of 100 segments was built by selecting one sentence every three for the test set, until reaching the desired number. With the remaining sentences, a tuning set of 50 sentences was obtained by picking one segment every four. The other 188 translation pairs that did not end up in either of the former sets would be used for training the systems. For the adaptation with the MT system, the training set was used to create a small translation memory via LF Aligner[37]. In the case of the chosen LLM, sentences of the training set would be used to craft examples for prompting. With the resources ready for use, it became crucial to understand how to make the best out of them. The following Section will discuss the systems and methods used for adaptation to the target domain.

3.4. Domain Adaptation

3.4.1. Neural Machine Translation - Modern MT

ModernMT (Bertoldi et al., 2018) was first launched in 2017 as an open-source project through the collaboration of four prominent European institutions: Translated.net, the University of Edinburgh (UEDIN), Fondazione Bruno Kessler (FBK) and TAUS. The system currently supports more than 200 languages, is accessible via API and can be integrated in other systems for Computer Assisted Translation (CAT) tools.

The reason why ModernMT stands out from most MT systems lies in the approach it was developed on, namely *instance-based adaptation* (Farajian et al., 2017). Basically, from a pool of parallel data, the NMT model retrieves a set of translation segments similar to a given untranslated sentence. Model parameters are then fine-tuned locally by using the recalled translation pairs. Upon translating a segment, parameters are reset to their original values. From the CAT-tool perspective, the sentence retrieval operates within the translation memory provided by the translator. The model also keeps track of human post-editing in real-time, adding the corrected sentences to the pool of parallel data for future translations. The same process can be repeated for document-level adaptation, where the model generates a translation based on the content on the whole text. Overall, these

[36] If either segment in a translation pair did not meet the required length criteria, the entire pair was excluded from the selection.
[37] Software available for download at: https://sourceforge.net/projects/aligner/ [Last accessed 14/02/2025]

integrations allow for adapting the MT output to the terminology and style of the user. Several studies have illustrated promising improvements of output quality through instance-based adaptation, especially with regards to terminology translation (Farajian et al., 2017; Nayak et al., 2023) and even speech-to-text translation (Di Gangi, 2022). These recent results, combined with ModernMT's ability to fine-tune the model on the fly without additional costs, eventually lead to the choice of ModernMT for the work described in this dissertation.

Domain adaptation with ModernMT is relatively straight-forward and can be carried out on MateCat (Federico et al., 2014), a CAT tool which incorporates ModernMT as its main MT engine. The platform was also used for the experiments of this thesis for NMT domain adaptation. To ensure an optimal workflow, both documents with the tuning and test set sentences are uploaded into a dedicated project, thereby ensuring that the tuning set is the first to be translated. It is also crucial to add the translation memory obtained from the training set to the project, for better sentence retrieval. After the automated translation takes place, the segments in the tuning set are post-edited (where needed) and marked as accepted translations. At his point, thanks to the TM previously uploaded and the PE of sentences in the tuning set, the MT engine has been adapted to the style and terminology of our documents. We then proceed to submit sentences in the test set, which are only processed by the MT engine without further post editing nor segment status confirmation.

Before delving on the adopted methodology for replicating domain adaptive MT with language models, the upcoming subsection will introduce the LLM chosen for this experimental framework.

3.4.2. Large Language Model – LLaMa 3.2.

LLaMa 3.2.[38], developed by Meta AI and released in late September 2024, introduced multimodal capabilities in its 11B and 90B parameter versions, enabling both text and image processing. This enhancement improves document-level understanding and visual grounding tasks, positioning these models as vision-language systems. To integrate image reasoning, LLaMa 3.2. vision models incorporate weights that connect a pre-trained

[38] Official release announcement available at: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ [Last accessed 13/02/2025]

image encoder with the language model. Notably, only the image encoder parameters were updated, preserving the model's text-based capabilities.

The training process for these multimodal LLMs followed a multi-stage approach: pre-trained LLaMa 3.1 text models were used as baseline, with the addition of image adapters. Pre-training is carried out on large-scale image-text datasets. Beside fine-tuning, post-training involved rejection sampling (H. Xu et al., 2024), in which the LLM is asked to generate a large number of outputs. Responses that do not align with desired standards or safety guidelines are later filtered out. Finally, synthetic data generation was implemented to enhance alignment and safety of the final vision models.

Smaller, text-only instances of LLaMa 3.2 (1B and 3B parameters) are also available, featuring a 128K-token context window and strong multilingual text generation and tool-calling capabilities. To make these models more efficient, Meta AI employed pruning (Ziheng Wang et al., 2020) and knowledge distillation (Hinton et al., 2015). Pruning aims at reducing the number of model parameters (or its components, such as neurons in the neural network) while preserving as much knowledge and performance as possible. In the case of the lightweight LLaMa 3.2. models, the same pruning process for LLaMa 3.1 8B (Grattafiori et al., 2024) was adopted.

As for knowledge distillation, smaller models are trained on outputs from larger ones, so that the lightweight model captures knowledge and retains performance despite reduced size and computational resources. In this sense, the text-based LLaMa 3.2. models were refined by leveraging the output values from the small- and medium sized instances of LLaMa 3.1 (that is, 8B and 70B).

Post-training followed a similar approach to LLaMa 3.1, incorporating supervised fine-tuning, rejection sampling, and direct preference optimization (Rafailov et al., 2024), where the model is fine-tuned to align output with human preference data. Synthetic data generation was also employed for better optimization. Moreover, Meta AI collaborated with prominent tech companies such as Qualcomm, Mediatek, and Arm to increase on-device performance for its new lightweight models.

Currently, LLaMa models are downloadable via HuggingFace[39] and LLaMa.com[40][41]. However, they can also be accessed via a special platform provided by

[39] https://www.llama.com/ [Last accessed 13/02/2025]
[40] https://huggingface.co/meta-llama [Last accessed 13/02/2025]
[41] Inference code is also available at: https://github.com/meta-llama/llama [last accessed 13/02/2025].

GroQ.com[42], which allows fast inference of the latest AI models upon a free subscription.

Based on promising results obtained by its predecessor LLaMa 2 (Aycock et al., 2023; Eschbach-Dymanus, 2024; D. Zhu, 2024), and in the hope that a larger model size would deliver better translation output, LLaMa 3.2. 90B was chosen as the LLM for our experiments. Details on how the LLM was leveraged for domain adaptive translation will be addressed in the upcoming subsection.

3.4.3. Implementing Adaptive MT with LLMs

Moslem (2023a) provided the most recent and complete example of applying real-time, domain adaptive MT to LLMs. Leveraging in-context learning capabilities of LLMs, the study explores the performance of different language models in zero-shot MT and MT with fuzzy matches[43]. Figure 5 and Figure 6 provide a simplified representation for each of the two tasks.

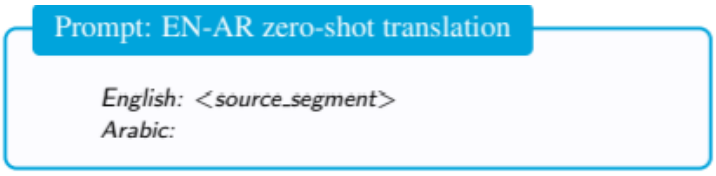

Figure 5 - Example of zero-shot translation (borrowed from Moslem et al., 2023a)

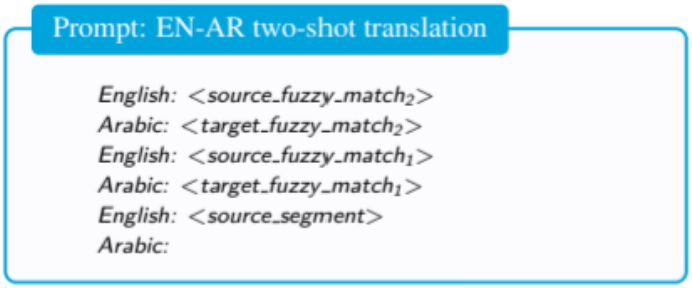

Figure 6 - Example of adaptive MT with fuzzy matches (borrowed from Moslem et al., 2023a)

[42] https://groq.com/ [Last accessed 17/02/2025]
[43] i.e., previously translated sentence pairs similar to a given untranslated source sentence

Among other things, the study examines also MT post-editing and MT-constrained terminology. Results suggest that LLM adaptive MT with fuzzy matches outperforms traditional MT models in high-resource languages. As regards fuzzy matches, these can be easily implemented as examples for few-shot prompting, and the study has reported that this method can yield promising results even for low-resourced languages.

For the scope of this dissertation, the workflow for LLM-based adaptive MT takes inspiration from the method adopted by Moslem et al. (ibid.) The baseline translation is obtained by prompting LLaMa 3.2. with the zero-shot setting illustrated in Figure 5. As regards the adaptation process, we implement the two-shot setting with fuzzy matches shown in Figure 6. We wrote a Python script to retrieve the top two matches from the training and tuning set for each English segment from the test set. Sentences are Specifically, the code compares sentences based on semantic similarity scores, which were computed with All-Mini-LM-L6-v2, a pre-trained SentenceTransformer model (Reimers & Gurevych, 2019). Given the limited size of our dataset, no minimum similarity score threshold was applied. Although Moslem et al. (ibid.) state that the order of fuzzy matches in the prompt does not impact translation quality, the matches in this framework were arranged with the highest similarity pair listed first.

Upon inference, some preliminary attempts revealed that LLaMa 3.2. tends to produce verbose outputs if not given necessary specifications. A viable solution to avoid any possible source of verbosity is to insert a system message, a strategy which is typically used by AI developers to guide the models into how to interpret conversations with the end users. Since the GroQ inference platform allows for inputting a system message, the following guideline was given to the LLM:

> "You need to translate sentences from English to Italian from two scientific articles on Covid-19, for each we will provide some examples of the style and terminology used, and a translation provided by a machine translation system (MT), which might contain errors. Please provide only your best translation in Italian, with no other explanation, comment or field."

3.5. Assessing Translation Quality

After adapting both systems, a rigorous evaluation framework was implemented to assess translation quality, since evaluation metrics offer an important indication of system

performance. This section will discuss how MT quality has been evaluated, both in terms of automatic and manual assessment.

3.5.1. Automated metrics

As anticipated in 2.2.3., automatic evaluation metrics provide a quantitative assessment of translation quality by comparing MT output with human references. After completing the adaptation process for both systems, quality improvement is measured through three automatic evaluation metrics: BLEU, chrF3 and COMET. The implementation of BLEU is motivated by its widespread use as standard metric in MT research and industry, despite limitations described in Section 2.2.3. and elsewhere (Kocmi et al., 2021; Marie et al., 2021).

Following provisions by Kocmi et al. (2021), chrF3, a more refined version of chrF, is also adopted for better generalization and due to higher correlation of character-based metrics to human assessments. Both BLEU and chrF3 were computed via the SacreBLEU tool (Post, 2018)[44].

Lastly, we chose COMET as our third metric, as well as the only one based on a pre-trained framework. Specifically, the version used for this work corresponds to COMET-22 (Rei et al., 2022)[45]. As mentioned in Section 2.2.3., COMET has also displayed high correlation with human judgements in comparison with most metrics (Freitag et al., 2021; Kocmi et al. 2021). However, manual evaluation is crucial for a comprehensive, fine-grained analysis of system performance, and this framework, as it will be explained in the following subsection, is no exception.

3.5.2. Manual evaluation

To further test the results of automated scores, translation output is evaluated manually for both systems and for each sentence in the test set. The adopted method of choice is Direct Assessment (DA; Graham et al., 2013), in which one or more MT systems are evaluated at sentence level through a score in a continuous range of 0-100. Historically, two instances of DA have been used during evaluation campaigns. The first one assesses

[44] Documentation available at https://github.com/mjpost/sacrebleu [Last accessed 12/02/2025]
[45] Model download at https://huggingface.co/Unbabel/wmt22-comet-da [Last accessed 25/02/2025]

adequacy and is also referred to as DA-src, through which translations are scored based on similarity to the source text. Fluency can be instead evaluated through DA-ref, where the score is assigned to the target sentence by comparison with a reference translation. However, this latter instance of DA has been at the centre of discussions regarding *reference bias,* which can manifest by:

> "giving an implicit boost to candidate translations which are very similar (e.g., in syntax or lexical choice) to the corresponding reference text, or by penalizing good translations because of translation errors affecting the reference itself." (Bentivogli et al., 2018, p. 62).

Studies on reference bias (Bentivogli et al., 2018; Fomicheva & Specia, 2016; Graham et al., 2016;) indicated that DA-ref had indeed lower correlation with other automated metrics and PE, with DA-src proving to be significantly more reliable instead. It should also be noted that DA has been improved more recently by integrating it into more sophisticated evaluation frameworks, like the Scalar Quality Metric (SQM; Kocmi et al., 2022). First implemented at WMT22, this revisitation involves giving both a DA on the translation quality, together with a judgement on a discrete scale in the range 0-6. Nevertheless, DA alone may be more viable for translators at any level, since it is simple and fast to apply. For all the above reasons, it was decided to implement DA-src alone to avoid any possible source of bias, and as a consequence, fluency has not been assessed in this setting.

3.6. Summing up

With the experimental framework now fully outlined, the groundwork has been laid for a detailed evaluation of the domain adaptation strategies implemented. This Chapter described the research hypothesis and questions, dataset selection, preprocessing steps, and the methodologies used to adapt and assess both the NMT and LLM systems. The strategies applied to improve domain adaptation for each system, along with the evaluation metrics chosen to measure translation quality. By structuring the methodology in this way, the study ensures a rigorous and replicable approach to comparing the two translation technologies. In the next Chapter, the results of these experiments will be presented and analysed, offering insights into the effectiveness of domain adaptation for Crisis Translation. Additionally, this evaluation provides a first indicative score for the

English-Italian translation direction within the TICO-19 benchmark, contributing to a broader understanding of system performance in the academic-scientific subdomain.

4. Results

4.1. Introduction

This Chapter presents the evaluation results of the adapted neural machine translation (NMT) system and the large language model (LLM) used for this dissertation. The analysis is divided into two main sections: an automatic evaluation (4.2.), where system performance is assessed using BLEU, chrF3, and COMET scores, and a manual evaluation (4.3.), where human DA-src provides a more fine-grained comparison of translation quality. The findings highlight both the strengths and limitations of domain-adaptive MT and LLM-driven translation, contributing to the broader discussion on their suitability for crisis translation scenarios. Finally, the Chapter reflects on key limitations of this work, considering aspects such as dataset size, quality assurance constraints, and computational limitations of the platform used for model inference (4.4.).

4.2. Automatic evaluation

As already explained in 2.3.4., it is standard practice in MT research and industry to evaluate system performance through both automated metrics and manual assessment. Having completed the domain adaptation for both examined systems, our TICO-19 test set of 100 aligned sentences first underwent an automated evaluation, with BLEU, chrF3 and COMET (see 3.4.1. for more) as our adopted metrics. All three can be expressed in a range 0-100, with the score being directly proportional to the system performance. Namely, the higher the evaluation, the better the quality of the output should be. Breakdown of results for both systems, before and after the adaptation, is provided in Table 5, and the translation direction tested is English-Italian. All scores are to be interpreted on a quantitative basis, meaning that they assess system performance on the whole test set, and not at sentence level.

System	BLEU	chrF3	COMET
ModernMT (*base*)	50.50	73.20	90.3
ModernMT (*adapted*)	***50.77***	***73.85***	***91.1***
LLaMa 3.2 90B (*base*)	47.60	71.14	89.9
LLaMa 3.2. 90B (*2-shot*)	48.60	72.39	90.7

Table 5 - Results of automatic evaluation

Based on the results illustrated above, domain adaptation appears to deliver a slight improvement over the baseline performance for both systems, with ModernMT consistently emerging as the best system. In addition, both BLEU scores obtained by ModernMT on our English-Italian dataset are currently the highest reported for the TICO-19 PubMed subdomain (Anastasopoulos et al., 2020), outperforming all previously recorded results (ibid.). Generally, however, this appears to stand in contrast with the experiments carried out by Moslem et al. (2023a), where traditional MT encoder-decoder models were outperformed by few-shot LLM adaptive MT for high-resourced languages. Looking at values yielded by string-based metrics, ModernMT also seems to show less performance increase after adaptation when compared to LLaMa 3.2 90B. A possible interpretation of this result might be that our chosen language model adjusts its output more swiftly than commercial systems for adaptive MT. However, as BLEU penalizes differences in word order between reference and hypothesis, it might just be that the model tends to produce translations with higher similarity to the reference text in terms of sentence constituents order.

On the other hand, both systems register the same +0.8 increase in COMET scores post-adaptation, suggesting that the process may make impacts on translation output to some degree, even with a small dataset. Still, it is important not to over rely on automated scores results for determining the best system. As recently occurred at the WMT24 (Kocmi et al., 2024b), the system ranking first in the automatic evaluation might not be declared the outright winner after human evaluation has taken place. The next section will illustrate results of the manual evaluation in detail and will provide more fine-grained observations on the translations provided by the systems.

4.3. Manual evaluation

Manual evaluation was adopted to further assess which system provided better translation results. The evaluation framework of choice for manual scoring was Direct Assessment (DA) in its source-based instance (DA-src; see 3.4.2 for more), which is expressed in a range of 0-100. In the same fashion as the implemented automatic metrics, higher scores typically indicate higher translation quality, with reported scores referring to performance at test set level. Although by a small margin, the adapted ModernMT emerged as the better-performing system, scoring 94.33 compared to 93.98 obtained by the adapted LLaMa 3.2 90B in 2-shot setting. Overall, such scores hint at very good performance of both systems, except for some inconsistencies and nuances.

Regarding a more fine-grained analysis of the results, the challenges in document translation described in Section 3.3.2 also had an impact on the adaptation outcomes. Particularly evident was the case for the word "condition" and the associated term "underlying health conditions". As displayed in Table 6, both systems do not provide one single solution consistently[46].

	Sentence	DA-src
Src	Reports from China and Italy suggest that risk factors for severe disease include older age and the presence of at least one of several **underlying health conditions.**	
MMT	Rapporti provenienti da Cina e Italia suggeriscono che i fattori di rischio per la forma grave della malattia includono l'età avanzata e la presenza di almeno una delle numerose **condizioni di salute sottostanti.**	85
LLM	Rapporti provenienti dalla Cina e dall'Italia suggeriscono che i fattori di rischio per la malattia grave includono l'età avanzata e la presenza di almeno una delle varie **patologie pregresse**.	90
Src	These preliminary findings suggest that in the United States, persons with **underlying health conditions** or other recognized risk factors for severe outcomes from respiratory infections appear to be at a higher risk for severe disease from COVID-19 than are persons without these **conditions**.	
MMT	Questi risultati preliminari suggeriscono che negli Stati Uniti le persone con **patologie pregresse** o altri fattori di rischio riconosciuti per esiti gravi da infezioni respiratorie sembrano essere a più alto rischio di contrarre la malattia grave da COVID-19 rispetto alle persone senza queste **patologie**.	90

[46] All examples will also include single sentence scores for the sake of completeness and transparency.

LLM	Questi risultati preliminari suggeriscono che negli Stati Uniti, le persone con **patologie pregresse** o altri fattori di rischio riconosciuti per esiti gravi da infezioni respiratorie sembrano essere a maggior rischio di forme gravi di malattia da COVID-19 rispetto a coloro che non presentano queste **condizioni**.	85
Src	Among all COVID-19 patients with complete information on **underlying conditions** or risk factors, 184 deaths occurred (all among patients aged ≥19 years); 173 deaths (94%) were reported among patients with at least one **underlying condition.**	
NMT	Tra tutti i pazienti affetti da COVID-19 con informazioni complete sulle **condizioni di base** o sui fattori di rischio, si sono verificati 184 decessi (tutti tra i pazienti di età ≥19 anni); sono stati segnalati 173 decessi (94%) tra i pazienti con almeno una **condizione di base**.	79
LLM	Tra tutti i pazienti con COVID-19 con informazioni complete sulle **patologie pregresse** o sui fattori di rischio, si sono verificati 184 decessi (tutti tra pazienti di età ≥19 anni); 173 decessi (94%) sono stati segnalati tra pazienti con almeno una **patologia pregressa**.	92
Src	Strategies to protect all persons and especially those with **underlying health conditions**, including social distancing and handwashing, should be implemented by all communities and all persons to help slow the spread of COVID-19.	
NMT	Tutte le comunità e tutte le persone dovrebbero attuare strategie per proteggere tutte le persone, in particolare quelle con **problemi di salute preesistenti**, tra cui il distanziamento sociale e il lavaggio delle mani, per contribuire a rallentare la diffusione del COVID-19.	85
LLM	Le strategie per proteggere tutte le persone e in particolare quelle con **patologie pregresse**, tra cui il distanziamento sociale e il lavaggio delle mani, dovrebbero essere implementate da tutte le comunità e da tutte le persone per aiutare a rallentare la diffusione del COVID-19.	90

Table 6 - Some instances of the translation of "underlying health condition" across adapted systems

In this specific case, two-shot LLaMa 3.2. 90B seems to deliver greater consistency compared to NMT, but displays major adherence to the source segment overall, as it can be inferred from the last example illustrated.

Another significant setback during manual translation was represented by the grammatical gender used in Italian to address COVID-19. Table 7 illustrates some instances of the output for both systems.

	Sentence	DA-src
Src	All persons should take steps to protect themselves from COVID-19 and to protect others.	
MMT	Tutti dovrebbero adottare misure per proteggersi **dal** COVID-19 e per proteggere gli altri.	95
LLM	Tutte le persone dovrebbero adottare misure per proteggersi **dal** COVID-19 e per proteggere gli altri.	94
Src	According to demographic studies so far, COVID-19 seems to have different epidemiological features from SARS.	
MMT	Secondo gli studi demografici condotti finora, **il** COVID-19 sembra presentare caratteristiche epidemiologiche diverse dalla SARS.	100
LLM	Secondo gli studi demografici finora condotti, **il** COVID-19 sembra avere caratteristiche epidemiologiche diverse dalla SARS	100
Src	As a novel disease, COVID-19 has just started to manifest its full clinical course throughout thousands of patients.	
NMT	Essendo una malattia nuova, **il** COVID-19 ha appena iniziato a manifestare il suo pieno decorso clinico in migliaia di pazienti.	100
LLM	Come malattia nuova, **la** COVID-19 sta iniziando a manifestare il suo intero decorso clinico in migliaia di pazienti.	94
Src	Paul Hunter, et al., estimated that COVID-19, which seems substantially more infectious than SARS, will not end in 2020.	
NMT	Paul Hunter et al. hanno stimato che il COVID-19, che sembra notevolmente più contagioso della SARS, non scomparirà nel 2020.	94
LLM	Paul Hunter e altri hanno stimato che **la** COVID-19, che sembra essere sostanzialmente più infettiva della SARS, non terminerà nel 2020.	86

Table 7 - Case translation regarding the grammatical genders used for "COVID-19"

As explained in 3.3.2, either gender is considered correct when addressing the disease, but efforts to maintain consistency do not seem to be very effective. Although the adapted ModernMT showed strong coherence in using the masculine gender, the same could not apply to the two-shot LLaMa 3.2. 90B. Since all instances of “COVID-19” in our dataset were all addressed with the masculine gender, two possible explanations emerge: either the model incorporated patterns from the fuzzy matches contained in the shots, where all other diseases are assigned the feminine gender, or the translations could reflect the grammatical gender for “COVID-19” that was more frequent in the pre-training textual corpus for LLaMa 3.2 90B. This latter hypothesis, however, is virtually impossible to ascertain, as Meta AI has not (yet) disclosed the pre-training datasets for its more recent LLaMa models. Further cases of incorrect terminology involve single

instances of other terms, as illustrated in Table 8.

	Sentence	DA-src
Src	On February 14, 2020, the Feng Zhang group described a protocol of using the CRISPR-based SHERLOCK technique to detect SARS-CoV-2, which detects synthetic SARS-CoV-2 RNA fragments at 20 × 10-18 mol/L to 200 × 10-18 mol/L (10-100 copies per microliter of input) using a **dipstick** in less than an hour without requiring elaborate instrumentation.	
MMT	Il 14 febbraio 2020, il gruppo di Feng Zhang ha descritto un protocollo di utilizzo della tecnica SHERLOCK basata su CRISPR per rilevare il SARS-CoV-2, che rileva frammenti di RNA sintetici del SARS-CoV-2 da 20 × 10-18 mol/L a 200 × 10-18 mol/L (10-100 copie per microlitro di input) utilizzando un'**astina** di livello in meno di un'ora senza richiedere una strumentazione elaborata.	80
LLM	Il 14 febbraio 2020, il gruppo di Feng Zhang ha descritto un protocollo che utilizza la tecnica SHERLOCK basata su CRISPR per rilevare il SARS-CoV-2, che rileva frammenti di RNA sintetico del SARS-CoV-2 a concentrazioni di 20 × 10-18 mol/L a 200 × 10-18 mol/L (10-100 copie per microlitro di campione) utilizzando un **dipstick** in meno di un'ora senza richiedere strumentazione elaborata.	88
Src	It is possible, similar to SARS-CoV, that SARS-CoV-2 might become weaker in infectivity and eventually die down or become a less pathogenic virus co-existent with humans.	
MMT	È possibile che, analogamente a quanto accaduto con **la SARS-CoV**, **la SARS-CoV-2** possa perdere intensità infettiva e col tempo estinguersi o trasformarsi in un virus meno patogeno in grado di coesistere con gli esseri umani.	89
LLM	È possibile, analogamente a quanto avvenuto per **il SARS-CoV**, che **il SARS-CoV-2** possa diventare meno infettivo e alla fine attenuarsi o trasformarsi in un virus meno patogeno che coesiste con gli esseri umani.	97

Table 8 - Further terminological inconsistencies found during the manual evaluation

In the case for the word "dipstick", the translation provided by ModernMT is not accurate, since "astina" designates an eyeglasses temple (Vocabolario Treccani, 2025). LlaMa 3.2. 90 B does not provide an equivalent in the target language, thus making "dipstick" a loan word in the target sentence. In this context, it is highly likely that the noun refers to a strip of paper use to identify one or more constituents in body fluids (Merriam-Webster Dictionary, 2025)[47]. A potential solution, which was also featured in the manual translation, might thus be offered by "striscia reattiva". Generally speaking, this type of

[47] In particular, the dictionary entry describes the word "dipstick" in relation to urine tests, but the SHERLOCK protocol for identifying Sars-Cov-2 detects presence of the virus in a similar fashion through human saliva. (F. Zhang et al., 2020)

error may be attributed to the low frequency of the word in the experimental dataset, and could be easily solvable by using a larger corpus. On the other hand, addressing SARS-CoVs with the feminine gender is clearly incorrect, since the word "virus" is masculine in Italian. Similarly to the inconsistencies displayed with the term "COVID-19", this other case might also be caused by integrating patterns of the training data to the final output. In a previous work on NMT adaptation, Contarino (2021) observed that such inconsistencies, especially when translating specialized terminology through MT, might also be linked to how instance-based, domain adaptive MT operates. As outlined in 3.4.1, domain adaptation takes place by dynamically selecting sentences with source text similar to the segment being translated. However, its effectiveness can be significantly hindered when only a limited amount of parallel data is available, and no sufficient translation examples can be retrieved.

4.4. Discussion

To better contextualize the results of this work, we provide a recap of the hypothesis and research questions presented in 3.2.:

> ***Hypothesis:*** *Domain adapted, LLM-driven translation can achieve equal or better translation results than a NMT system that has, in its turn, undergone a domain adaptation process.*
>
> ***Research Question 1:*** *"Do LLMs deliver better quantitative results both before and after domain adaptation?"*
>
> ***Research Question 2:*** *"Does domain adaptation yield such an impact on the output to be necessary for future development of translation technologies in crisis scenarios?".*

Given the scores presented in Sections 4.2. and 4.3., our hypothesis might have to be rejected for the time being. Consequently, the answer to our first research question also appears to be negative, as ModernMT proved to be the better system among those examined both pre- and post-adaptation. On a practical note, having to factor in

environmental, ethical and financial issues described in 2.3.3, NMT systems (both in their baseline and adaptive instances) may still represent the best system to implement in crisis management, as their development and deployment is comparatively swifter and easier. Nevertheless, manual assessments seem to align with COMET scores, suggesting that the performance gap between the two systems may be narrower than expected. The situation is not as clear-cut in for our second research question. Based on automatic scores, adapted systems do indeed perform better than a raw NMT system or a pre-trained LLM, which might advocate for integrating domain adaptation strategies during the development of Crisis Translation technologies. On the other hand, the margin between pre- and post-adaptation is still slim, with baseline performance for both systems seemingly on par with other high-resource languages examined in the TICO-19 benchmark. This latter finding also suggests that the English-Italian translation direction may be already well-supported in both specialized and emergency datasets, reducing the need for further fine-tuning. However, this assumption still needs further validation, as we did not assess baseline translation results through manual evaluation.

While results may seem promising at first glance, their interpretation must consider several important limitations. Firstly, our translation quality assurance (QA) workflow did not follow the one outlined by TICO-19 authors (Anastasopoulos et al., 2020), mainly due to time constraints and limited allocation of resources. As already explained in 1.3.4., authors of the original TICO-19 benchmark encourage potential contributors to adopt the following plan: translation by expert translators or Language Service Providers (LSP), a first review by medical experts (if available for the target language), a second review for high priority data, refinement of translations until a satisfactory quality rating is achieved, and a thorough annotation and classification of errors (ibid.). On the other hand, the translation workflow applied to the two documents in this study was limited to the first two stages. Since no quality rating was performed on the translated segments used for this thesis, they may score below the 95% threshold set by the TICO-19 authors (ibid.). Consequently, reported scores might not objective about the performance of both systems. It is also important to note that the TICO-19 authors do not provide a detailed description of the quality ranking framework adopted before dataset release (ibid.). Even with optimal time and resource availability, replicating the full process for the dataset used in this study would have been challenging.

Secondly, the results presented in this thesis are significantly limited by small

sample size. Since each individual assessment has a greater impact on the final scores, overall evaluations may be further distorted, leading to overgeneralized conclusions. Similarly, a reduced sample size diminishes the impact of statistical significance tests. Although such tests are highly recommended to further examine system performance (Kocmi et al., 2021), their outcome would not be generalizable in such a case. For this reason, statistical inference testing was not conducted in the present thesis. Additionally, the dataset size also limits the number of examples that can be used as shots in prompting. Although inconsistencies found during the manual evaluation for LLaMa 3.2. 90B suggest otherwise, small datasets with a low degree of sample diversification could lead to model overfitting. This means that given model memorizes the training set, developing poor generalization capabilities as a consequence.

Thirdly, this thesis represents the first known use of the DA-src score for human assessment within the Department of Translation at the University of Bologna. Previous dissertations on NMT and AI have primarily relied on Likert scales for manual evaluation (Falsone, 2024; Mainardi, 2024; Marvulli, 2023). As a result, DA judgments in this study may be biased toward higher scores.

Fourthly, manual evaluation was not conducted on the baseline systems. While this additional analysis would have delivered a comprehensive answer to the second research question, time constraints prevented its implementation.

Lastly, a set of practical restrictions affects the process of domain adaptation with LLMs. The GroQ platform for model inference imposes token limits per minute and per day, depending on the queried model. As a result, completing adaptation with LLMs may become a lengthy process. Moreover, the platform deletes all user input data upon session closure. While this is a commendable privacy measure, it presents a limitation in terms of result retention and reproducibility, as there is no way to retrieve outputs once the session is ended by the user. We therefore advise ensuring results are stored elsewhere before closing the platform. It is also strongly recommended to keep the working session open for the entire duration of the experiment(s).

Conclusions

This dissertation set out to explore the comparative performance of an adaptive Neural Machine Translation (NMT) system and a Large Language Model (LLM) within the domain of crisis translation. The study was motivated by recent advancements in artificial intelligence and their potential integration into multilingual crisis communication. While NMT systems have long been deployed for real-time translation in humanitarian emergencies, LLMs have emerged as an interesting alternative, offering context-aware translation through in-context learning. However, the extent to which these models can effectively replace (or complement) NMT systems in domain-specific scenarios remains an open question.

To address this, the dissertation formulated the following hypothesis: "*Domain-adapted, LLM-driven translation can achieve equal or better translation results than an NMT system that has, in its turn, undergone a domain adaptation process.*" This hypothesis was examined through two research questions: "*Do LLMs deliver better results both before and after domain adaptation?*", and "*Does domain adaptation yield such an impact on the output to be necessary for the future development of translation technologies in crisis scenarios?*" The methodology adopted to test these questions involved evaluating ModernMT (a domain-adaptive NMT system) and LLaMa 3.2 90B (a domain-adapted LLM) on a set of specialized COVID-19 texts from the TICO-19 benchmark. Given the absence of an Italian version of this dataset, an initial step in the research involved creating a high-quality human reference translation, ensuring an accurate basis for evaluation. The performance of both systems was assessed through a combination of automatic metrics (BLEU, COMET, chrF3) and manual evaluation through source-based Direct Assessment, providing a multi-faceted analysis of translation quality.

Our findings indicate that the initial hypothesis could not hold: the domain-adapted NMT system outperformed the LLM both before and after adaptation. However, the gap between the two systems was narrower than expected, particularly when considering COMET scores and manual assessment. This suggests that while adapted NMT could be the most effective solution for crisis translation, LLMs might still play a supporting role in certain applications. From a practical standpoint, this dissertation highlights several factors that must be considered when developing crisis translation

technologies. As illustrated in Chapter 2, both NMT engines and LLMs require extensive computational resources. Additionally, the ethical and environmental costs associated with training, particularly for LLMs, pose a significant challenge for their large-scale adoption in crisis management. However, the primary focus of language technology development should remain on what enables its success in the first place: high-quality, in-domain parallel language data. This involves either creating new language resources or curating existing ones, including parallel corpora, translation memories, and terminologies. As discussed in Chapter 1, the management and long-term storage of parallel datasets are essential to ensuring optimal performance of automated systems. Moreover, with the contemporary world facing a growing number of local and global crises, we emphasize once again that proactively creating and managing language resources can bolster crisis preparedness and provide a crucial advantage when a new emergency arises

Building on this and the limitations posed by our corpus' size, the most immediate development for this dissertation could involve expanding the experimental dataset on crisis-related content. Multiple solutions are available in this sense, like providing more translations from the TICO-19 benchmark, or leverage existing corpora through parallel sentence mining. Potential downstream impacts of this development are countless: better generalization on systems' performance and learning capabilities, further evaluation with statistical inference testing and more insight on the impact of repeated fuzzy matches in instance-based adaptation, and a resource that can be leveraged to reach linguistic minorities in Italy by relying on Italian as a pivot language. Specifically, shall dataset expansion be achieved through additional translations of TICO-19 documents, we recommend, time and resources allowing, to carry out a thorough QA procedure before carrying out experiments on the material. Future analyses could also take advantage from integrating other NMT systems and, most importantly different LLMs, like the lightweight LLaMa 3.2 1B and 3B and/or the latest instance of LLaMa, the 3.3 version. On a geneal note, a possible evolution of the present dissertation may regard the metrics used for automated evaluation. Specifically, a recent study by Zouhar et al. (2024) has raised concerns about training bias, different software versions, and influences of translationese on scores displayed by the COMET framework. Also based on recent recommendations by Kocmi et al (2024a), future works could 1) use COMET-kiwi, a

reference free instance of COMET, as a main metric[48] 2) using a metric of a different type, preferably BLUERT[49] (Sellam et al., 2020) 3) discard BLEU and chrF to evaluate unrelated systems. As for manual evaluation, future works could also consider keeping up the use of DA judgements or switching to DA-SQM (Kocmi et al., 2022) for more fine-grained analyses.

All in all, these results contribute to the broader discussion on how AI-driven translation technologies can be integrated into multilingual crisis response strategies. While NMT continues to offer the most efficient and reliable solution for specialized translation, LLMs may serve as a complementary tool, particularly when domain adaptation is not feasible due to time or data constraints.

[48] or alternatively compute COMET via the SacreCOMET tool, available at: https://github.com/PinzhenChen/sacrecomet [Last accessed 25/02/2025]

[49] Documentation available at: https://github.com/google-research/bleurt?tab=readme-ov-file [Last access 25/02/2025]

References

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., & Larochelle, H. (2024). *Many-Shot In-Context Learning* (arXiv:2404.11018). arXiv. https://doi.org/10.48550/arXiv.2404.11018

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., & Ghazvininejad, M. (2022). *In-context Examples Selection for Machine Translation* (arXiv:2212.02437). arXiv. https://doi.org/10.48550/arXiv.2212.02437

Alves, D., Guerreiro, N., Alves, J., Pombal, J., Rei, R., De Souza, J., Colombo, P., & Martins, A. (2023). Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. Findings of the Association for Computational Linguistics: EMNLP 2023, 11127–11148. https://doi.org/10.18653/v1/2023.findings-emnlp.744

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). TICO-19: The Translation Initiative for COvid-19. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online.

Aycock, S., & Bawden, R. (2024). Topic-guided Example Selection for Domain Adaptation in LLM-based Machine Translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 175-195).

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate* (arXiv:1409.0473). arXiv. https://doi.org/10.48550/arXiv.1409.0473

Balakrishnan, V., Ng, W. Z., Soo, M. C., Han, G. J., & Lee, C. J. (2022). Infodemic and fake news - A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International journal of disaster risk reduction : IJDRR*, *78*, 103144. https://doi.org/10.1016/j.ijdrr.2022.103144

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., & Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P.

Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 1–61). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., … Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, & M. Negri (Eds.), *Proceedings of the fifth conference on machine translation* (pp. 1–55). Association for Computational Linguistics. https://aclanthology.org/2020.wmt-1.1/

Basharin, G. P., Langville, A. N., & Naumov, V. A. (2004). The life and work of AA Markov. *Linear algebra and its applications*, *386*, 3-26.

Bawden, R., & Yvon, F. (2023). Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bengio, Y., Ducharme, R., & Vincent, P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3* 1137-1155

Bentivogli, L., Cettolo, M., Federico, M., & Federmann, C. (2018). Machine translation human evaluation: An investigation of evaluation based on post-editing and its relation with direct assessment. In M. Turchi, J. Niehues, & M. Frederico (Eds.), *Proceedings of the 15th international conference on spoken language translation* (pp. 62–69). International Conference on Spoken Language Translation. https://aclanthology.org/2018.iwslt-1.9/

Bertoldi, N., Caroselli, D., & Federico, M. (2018). The ModernMT Project. In J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popović, C. Rico, A. Martins, J. Van den Bogaert, & M. L. Forcada (Eds.), *Proceedings of the 21st annual conference of the european association for machine translation* (p. 365). https://aclanthology.org/2018.eamt-main.46/

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., … Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 131–198. https://doi.org/10.18653/v1/W16-2301

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., & Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). 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: Shared task papers* (pp. 272–303). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401

Boulanger, A. M. (2024). The use of machine translation and AI in medical translation: Pros and cons. Medical Writing, 33(1), 62–65. https://doi.org/10.56012/fcbh4324

Bowker, Lynne and Jairo Buitrago Ciro. (2019). Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community. Emerald Publishing. https://doi.org/10.1108/9781787567214

Briakou, E., Liu, Z., Cherry, C., & Freitag, M. (2024). *On the Implications of Verbose LLM Outputs: A Case Study in Translation Evaluation* (arXiv:2410.00863). arXiv. https://doi.org/10.48550/arXiv.2410.00863

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 33 1877–1901. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.14165

Cadwell, P., O'Brien, S., & DeLuca, E. (2019). More than tweets: A critical reflection on developing and testing crisis machine translation technology. *Translation Spaces*, 8, 300–333. https://doi.org/10.1075/ts.19018.cad

Cambridge Dictionary (2025) Condition. In *Camrbidge Dictionary* Retrieved February 21, 2025 from https://dictionary.cambridge.org/dictionary/english/condition

Carolan, K., Fennelly, L., & Smeaton, A. F. (2024). A Review of Multi-Modal Large Language and Vision Models (*arXiv:2404.01322*). arXiv. https://doi.org/10.48550/arXiv.2404.01322

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics, 108, 109–120. https://doi.org/10.1515/pralin-2017-0013

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment* (Vol. 1, pp. 9–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2

CDC COVID-19 Response Team (2020). Preliminary Estimates of the Prevalence of Selected Underlying Health Conditions Among Patients with Coronavirus Disease 2019 - United States, February 12-March 28, 2020. *MMWR. Morbidity and mortality weekly report*, *69*(13), 382–386. https://doi.org/10.15585/mmwr.mm6913e2

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, *26*(2), 137–161. https://doi.org/10.1017/S1351324919000469

Chen, B., Kuhn, R., Foster, G., Cherry, C., & Huang, F. (2016). Bilingual methods for adaptive training data selection for machine translation. In S. Green & L. Schwartz (Eds.), *Conferences of the association for machine translation in the americas: MT researchers' track* (pp. 93–106). The Association for Machine Translation in the Americas. https://aclanthology.org/2016.amta-researchers.8/

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53.

CMU LTI Haitian Creole Text Data (n.d.) CMU Speech Group. Retrieved December 03, 2024 from: http://www.speech.cs.cmu.edu/haitian/text/ [Last accessed 03/12/2024]

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. https://doi.org/10.48550/arXiv.1911.02116

Contarino, G. A. (2021). *Neural machine translation adaptation and automatic terminology evaluation: a case study on Italian and South Tyrolean German legal texts* [Master's thesis, Università di Bologna] *AMS Tesi di Laurea - AlmaDL - Università di Bologna*. https://amslaurea.unibo.it/24989/

Content Team (2023, July 10). *On-the-Fly Machine Translation Domain Adaptation - UNBaBel.* Unbabel. Retrieved February 04, 2025, https://unbabel.com/on-the-fly-machine-translation-domain-adaptation/

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). *The rising costs of training frontier AI models*. https://arxiv.org/abs/2405.21015

Council of Europe, European Convention on Human Rights, as amended by Protocols Nos. 11, 14 and 15, ETS No. 005, 4 November 1950, https://www.refworld.org/legal/agreements/coe/1950/en/18688 [accessed 11 December 2024]

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t

Custom Machine Translation - Tilde.ai. (2025, January 24) Tilde.ai Retrieved February 07, 2025, from https://tilde.ai/custom-machine-translation/

De Sa, V. (1993). Learning classification with unlabeled data. *Advances in neural information*

processing systems, *6*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Diandaru, R., Susanto, L., Tang, Z., Purwarianti, A., & Wijaya, D. (2024). *Could We Have Had Better Multilingual LLMs If English Was Not the Central Language?* (arXiv:2402.13917). arXiv. https://doi.org/10.48550/arXiv.2402.13917

Di Gangi, M. A., Nguyen, V.-N., Negri, M., & Turchi, M. (2020). Instance-based model adaptation for direct speech translation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7914–7918. https://doi.org/10.1109/ICASSP40776.2020.9053901

Eschbach-Dymanus, J., Essenberger, F., Buschbeck, B., & Exel, M. (2024). Exploring the effectiveness of LLM domain adaptation for business it machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 610-622).

European Commission (2018, April 25). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe. (Report COM/2018/237) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN Last accessed 10/12/2024

European Commission. Joint Research Centre. (2024). *Artificial intelligence applied to disasters and crises management.* Publications Office. https://data.europa.eu/doi/10.2760/0323818

Falsone, I. (2024). *Sfide e risultati tra traduzione tecnica e traduzione automatica adattiva: Il caso dell'azienda spagnola ESPASEME*. [Master's thesis, Università di Bologna] AMS Tesi di Laurea https://amslaurea.unibo.it/id/eprint/31170/

Farajian, M. A., Turchi, M., Negri, M., & Federico, M. (2017). Multi-domain neural machine translation through unsupervised adaptation. 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. 127–137). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4713

Federici, F. M., & Cadwell, P. (2018). Training citizen translators: Design and delivery of bespoke

training on the fundamentals of translation for New Zealand Red Cross. *Translation Spaces*, *7*(1), 20-43.

Federici, F. M., O'Hagan, M., O'Brien, S., & Cadwell, P. (2019). Crisis Translation Training: Challenges Arising from New Contexts of Translation. In *Cultus Journal,* 12(1), 246-279. https://www.cultusjournal.com/files/Archives/Cultus_2019_12_013_ Federici_et-al.pdf

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., & Germann, U. (2014). The MateCat Tool. In L. Tounsi & R. Rak (Eds.), *Proceedings of COLING 2014, the 25th international conference on computational linguistics: System demonstrations* (pp. 129–132). Dublin City University and Association for Computational Linguistics. https://aclanthology.org/C14-2028/

FEMA. (2018). A proposed research agenda for the emergency management higher education community. U.S. Federal Emergency Management Agency. https://training.fema.gov/hiedu/docs/latest/2018_fema_research_agenda_final-508%20(march%202018).pdf

Fernicola, F. (2022). *Return to the Source: Assessing Machine Translation Suitability based on the Source Text using XLM-RoBERTa* [Master's thesis, Università di Bologna] AMS Tesi di Laurea https://amslaurea.unibo.it/id/eprint/25307/

Fomicheva, M., & Specia, L. (2016). Reference Bias in Monolingual Machine Translation Evaluation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 77–82. https://doi.org/10.18653/v1/P16-2013

Freitag, M., Grangier, D., & Caswell, I. (2020). BLEU might be Guilty but References are not Innocent. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 61–71). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.5

Freitag, M., Rei, R., Mathur, N., Lo, C., Stewart, C., Foster, G., Lavie, A., & Bojar, O. (2021). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, & C. Monz (Eds.), *Proceedings of the sixth conference on machine*

translation (pp. 733–774). Association for Computational Linguistics. https://aclanthology.org/2021.wmt-1.73/

Freitag, M., Mathur, N., Lo, C., Avramidis, E., Rei, R., Thompson, B., Kocmi, T., Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., & Foster, G. (2023). Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 578–628). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.51

Galiero, L. (2025) Original texts, translation resources, dateset subsets and notebooks for translating TICO-19 (EN-IT) PubMed texts with NMT and LLMs. GitHub. [Repository] https://github.com/lucia-galiero/TICO-19_NMT_LLM

Gao, D., Chen, K., Chen, B., Dai, H., Jin, L., Jiang, W., Ning, W., Yu, S., Xuan, Q., Cai, X., Yang, L., & Wang, Z. (2024). LLMs-based machine translation for E-commerce. *Expert Systems with Applications*, *258*, 125087. https://doi.org/10.1016/j.eswa.2024.125087

Gao, J., & Lin, C.-Y. (2004). Introduction to the special issue on statistical language modeling. *ACM Transactions on Asian Language Information Processing*, *3*(2), 87–93.

Ghazvininejad, M., Gonen, H., & Zettlemoyer, L. (2023). *Dictionary-based phrase-level prompting of large language models for machine translation*. https://arxiv.org/abs/2302.07856

Gimpel, L. (2024). Toward Open-Source AI Systems as Digital Public Goods: Definitions, Hopes and Challenges. In M. Streit-Bianchi & V. Gorini (Eds.), *New Frontiers in Science in the Era of AI* (pp. 129–142). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61187-2_8

Giovine, S. (July 3, 2020) *Il Covid-19 o La Covid-19? – Consulenza linguistica* Accademia della Crusca https://accademiadellacrusca.it/it/consulenza/il-covid19-o-la-covid19/2787

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In A. Pareja-Lora, M. Liakata, & S. Dipper (Eds.), *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 33–41). Association for Computational Linguistics. https://aclanthology.org/W13-2305/

Graham, Y., Baldwin, T., Dowling, M., Eskevich, M., Lynn, T., & Tounsi, L. (2016). Is all that

glitters in machine translation quality estimation really gold? In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3124–3134). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1294/

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). *The llama 3 herd of models*. https://arxiv.org/abs/2407.21783

Groq, Inc. (2025, February 24). *Groq is Fast AI Inference*. Groq. Retrieved July 18, 2024 from: https://groq.com/

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). *A survey on self-supervised learning: Algorithms, applications, and future trends*. https://arxiv.org/abs/2301.05712

Gupta, R., Lambert, P., Patel, R., & Tinsley, J. (2019). Improving robustness in real-world neural machine translation engines. In M. Forcada, A. Way, J. Tinsley, D. Shterionov, C. Rico, & F. Gaspari (Eds.), *Proceedings of machine translation summit XVII: Translator, project and user tracks* (pp. 142–148). European Association for Machine Translation. https://aclanthology.org/W19-6727/

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., & Ranzato, M. (2019). The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 6097–6110. https://doi.org/10.18653/v1/D19-1632

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage*. https://doi.org/10.36227/techrxiv.23589741.v1

Han, L., Jones, G. J. F., & Smeaton, A. F. (2021). *Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods* (arXiv:2105.03311). arXiv. https://doi.org/10.48550/arXiv.2105.03311

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. (2021). *Pre-Trained Models: Past, Present and Future* (arXiv:2106.07139). arXiv. https://doi.org/10.48550/arXiv.2106.07139

Hapke, H., Howard, C., & Lane, H. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon and Schuster.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation* (arXiv:2302.09210). arXiv. https://doi.org/10.48550/arXiv.2302.09210

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. https://arxiv.org/abs/1503.02531

Huang, J., & Chang, K. C.-C. (2023). *Towards reasoning in large language models: A survey*. https://arxiv.org/abs/2212.10403

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., & Wei, F. (2023). *Language Is Not All You Need: Aligning Perception with Language Models* (arXiv:2302.14045). arXiv. https://doi.org/10.48550/arXiv.2302.14045

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. (2016) Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

Interact – International Network on Crisis Translation Crisis Translation Retrieved December 16, 2024 from: https://sites.google.com/view/crisistranslation/home

Jelinek, F. (1990). Self-organized language modeling for speech recognition. *Readings in speech recognition*, 450-506.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine* (arXiv:2301.08745). arXiv. https://doi.org/10.48550/arXiv.2301.08745

Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1(67), 1.

Khayrallah, H., & Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 74–83. https://doi.org/10.18653/v1/W18-2709

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

Kim, S., Joo, S. J., Kim, D., Jang, J., Ye, S., Shin, J., & Seo, M. (2023). *The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning*. https://arxiv.org/abs/2305.14045

Ko, W.-J., El-Kishky, A., Renduchintala, A., Chaudhary, V., Goyal, N., Guzmán, F., Fung, P., Koehn, P., & Diab, M. (2021). Adapting high-resource NMT models to translate low-resource related languages without parallel data. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 802–812). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.66

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., & Menezes, A. (2021). *To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation* (arXiv:2107.10821). arXiv. https://doi.org/10.48550/arXiv.2107.10821

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., & Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, … M. Zampieri (Eds.), *Proceedings of the seventh conference on machine translation (WMT)* (pp. 1–45). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.1/

Kocmi, T., Zouhar, V., Federmann, C., & Post, M. (2024a). Navigating the metrics maze:

Reconciling score magnitudes and accuracies. 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. 1999–2014). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.110

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024b). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. *Proceedings of the Ninth Conference on Machine Translation*, 1–46. https://doi.org/10.18653/v1/2024.wmt-1.1

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. *In Proceedings of the First Workshop on Neural Machine Translation*, 28–39. https://doi.org/10.18653/v1/W17-3204

Koehn, P. (2020). *Neural Machine Translation* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108608480

Kombrink, S., Mikolov, T., Karafiát, M., & Burget, L. (2011). Recurrent neural network based language modeling in meeting recognition. *Interspeech 2011*, 2877–2880. https://doi.org/10.21437/Interspeech.2011-720

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Kukreja, S., Kumar, T., Purohit, A., Dasgupta, A., & Guha, D. (2024). A literature survey on open source large language models. *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, 133–143. https://doi.org/10.1145/3647782.3647803

Landahl, M. R., Bennett, D. M., & Phillips, B. D. (2019). Disaster Research: Past, present, and future. In P. Leavy (Ed.), *The Oxford Handbook of Methods for Public Scholarship*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274481.013.35

Lankford, S., & Way, A. (2024). *Leveraging LLMs for MT in Crisis Scenarios: A blueprint for low-resource languages* (arXiv:2410.23890). arXiv. https://doi.org/10.48550/arXiv.2410.23890

Language Weaver, Translation Technology - RWS. (n.d.) RWS. Retrieved February 04, 2025, from https://www.rws.com/language-weaver/

Lardilleux, A., & Lepage, Y. (2017). CHARCUT: Human-targeted character-based MT evaluation with loose differences. In S. Sakti & M. Utiyama (Eds.), *Proceedings of the 14th international conference on spoken language translation* (pp. 146–153). International Workshop on Spoken Language Translation. https://aclanthology.org/2017.iwslt-1.20/

Lewis, W. (2010). 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. https://aclanthology.org/2010.eamt-1.37/

Lewis, W., Munro, R., & Vogel, S. (2011). Crisis MT: Developing a cookbook for MT in crisis situations. In C. Callison-Burch, P. Koehn, C. Monz, & O. F. Zaidan (Eds.), *Proceedings of the sixth workshop on statistical machine translation* (pp. 501–511). Association for Computational Linguistics. https://aclanthology.org/W11-2164/

Lexical Comptung CZ s.r.o. (n.d.) *COVID-19 corpus from Open Research Dataset (CORD-19)* Sketch Engine. Retrieved December 4, 2024 from: https://www.sketchengine.eu/covid-19-corpus/

Lexical Computing CZ s.r.o. (n.d.). *enTenTen — English corpus from the web* Sketch Engine. Retrieved February 21, 2025 from https://www.sketchengine.eu/ententen-english-corpus/

LF Aligner donwload (2025) SourceForge.net. Retrieved February 14, 2025 from: https://sourceforge.net/projects/aligner/

Lin, Z., Feng, M., Santos, C. N. dos, Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). *A Structured Self-attentive Sentence Embedding* (arXiv:1703.03130). arXiv. https://doi.org/10.48550/arXiv.1703.03130

Liu, Z., Qiao, A., Neiswanger, W., Wang, H., Tan, B., Tao, T., Li, J., Wang, Y., Sun, S., Pangarkar, O., Fan, R., Gu, Y., Miller, V., Zhuang, Y., He, G., Li, H., Koto, F., Tang, L., Ranjan, N., … Xing, E. P. (2023). *LLM360: Towards fully transparent open-source llms*. https://arxiv.org/abs/2312.06550

LLaMa (n.d.) Meta Retrieved February 13, 2025 from: https://www.llama.com/

LlaMa 3.2: Revolutionizing edge AI and vision with open, customizable models (2024, September

25). MetaAI blog. Retrieved February 13, 2025 from: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

Lu, X., Zhao, Y., Qin, B., Huo, L., Yang, Q., & Xu, D. (2024). *How does architecture influence the base capabilities of pre-trained language models? A case study based on FFN-wider and MoE transformers*. https://arxiv.org/abs/2403.02436

Machine Translation Amazon Translate (n.d.) AWS. Retrieved February 04, 2025, from https://aws.amazon.com/translate/

Maffulli, S. (July 20, 2023) *Meta's LLaMa 2 license is not Open Source.* Open Source Initiative. Retrieved February 13, 2025 from: https://opensource.org/blog/metas-llama-2-license-is-not-open-source

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. arXiv. https://doi.org/10.48550/arXiv.2006.07264

Mainardi, P. (2024) *Bias di genere e traduzione automatica. Esperimenti con il linguaggio non binario diretto nella traduzione dall'inglese all'Italiano* [Master's thesis, Università di Bologna] AMS Tesi di Laurea https://amslaurea.unibo.it/id/eprint/32051/

Marie, B., Fujita, A., & Rubino, R. (2021). Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7297–7306. https://doi.org/10.18653/v1/2021.acl-long.566

Marvulli, F. (2023) *LA TECNOLOGIA MEDICA TRA TRADUZIONE UMANA E AUTOMATICA: IL CASO HERSILL* [Master's thesis, Università di Bologna] AMS Tesi di Laurea https://amslaurea.unibo.it/id/eprint/32051/

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4984–4997). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448

Merriam-Webster Dictionary (n.d.) Dipstick. In *Merriam-Webster Dictionary.* Retrieved February 22, 2025 from: https://www.merriam-webster.com/dictionary/dipstick

Merriam-Webster Dicitionary (2025) Condition. In *Merriam-Webster.com Dictionary*. Retrieved February 21, 2025 from https://www.merriam-webster.com/dictionary/condition

meta-llama (Meta Llama) (n.d.) HugginFace Retrieved February 13, 2025 from: https://huggingface.co/meta-llama

Meta Llama (2025) *Inference code for LLaMa models*. Github [Repository]. Retrieved February 07, 2025, from https://github.com/meta-llama/llama

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech 2010*, 1045–1048. https://doi.org/10.21437/Interspeech.2010-343

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. https://arxiv.org/abs/1301.3781

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large Language Models: A Survey* (arXiv:2402.06196). arXiv. https://doi.org/10.48550/arXiv.2402.06196

Mission 4636 (n.d.) Mission 4636. Retrieved December 12, 2024 from: https://www.mission4636.org

ModernMT | Welcome. (n.d.) ModernMT. Retrieved February 04, 2025, from https://www.modernmt.com/

Mohammadshahi, A., Nikoulina, V., Berard, A., Brun, C., Henderson, J., & Besacier, L. (2022). SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages (arXiv:2210.11621). arXiv. https://doi.org/10.48550/arXiv.2210.11621

Moorkens, J., Toral, A., Castilho, S., & Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. Translation Spaces, 7, 240–262. https://doi.org/10.1075/ts.18014.moo

Moorkens, J., & Lewis, D. (2019). Research Questions and a Proposal for the Future Governance of Translation Data. The Journal of Specialised Translation, 2–25.

Moslem, Y., Haque, R., Kelleher, J., & Way, A. (2023a). Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European*

Association for Machine Translation, Tampere, Finland p. 227-237. ACL Anthology. https://aclanthology.org/2023.eamt-1.22

Moslem, Y., Romani, G., Molaei, M., Kelleher, J. D., Haque, R., & Way, A. (2023b). Domain terminology integration into machine translation: Leveraging large language models. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the eighth conference on machine translation* (pp. 902–911). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.82

Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access*, *11*, 144497–144529. https://doi.org/10.1109/access.2023.3344653

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023). *Crosslingual Generalization through Multitask Finetuning* (arXiv:2211.01786). arXiv. https://doi.org/10.48550/arXiv.2211.01786

Mulloch. G. (2020, May 31). *Covid-19 Is History's Biggest Translation Challenge* WIRED Retrieved December 17, 2024 from: https://www.wired.com/story/covid-language-translation-problem/

Mytton, D. (2021). Data centre water consumption. *Npj Clean Water*, *4*(1), 11. https://doi.org/10.1038/s41545-021-00101-w

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models* (arXiv:2307.06435). arXiv. https://doi.org/10.48550/arXiv.2307.06435

Nayak, P., Kelleher, J., Haque, R., & Way, A. (2023). Instance-based domain adaptation for improving terminology translation. In M. Utiyama & R. Wang (Eds.), *Proceedings of machine translation summit XIX, vol. 1: Research track* (pp. 222–234). Asia-Pacific Association for Machine Translation. https://aclanthology.org/2023.mtsummit-research.19/

NGA (1979) Comprehensive Emergency Management: A Governor's Guide. Washington, D.C.: National Governor's Association

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19

news translation task submission. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 314–319). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5333

NHS Confederation, (2004). *Multilingual emergency phrasebook*. British Red Cross.

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of the second international conference on language resources and evaluation (LREC`00)*. European Language Resources Association (ELRA). https://aclanthology.org/L00-1210/

Nurminen, M., & Koponen, M. (2020). Machine translation and fair access to information. In *Translation Spaces*, 9(1), 150–169. https://doi.org/10.1075/ts.00025.nur

O'Brien, S., Federici, F., Cadwell, P., Marlowe, J., & Gerber, B. (2018). Language translation during disaster: A comparative analysis of five national approaches. *International Journal of Disaster Risk Reduction,* 31, 627–636. https://doi.org/10.1016/j.ijdrr.2018.07.006

O'Brien, S. (2019). Translation technology and disaster management. In Minako O'Hagan (Ed.), Routledge *Handbook of Translation and Technology* (pp. 304-318). Routledge. https://doi.org/10.4324/9781315311258-18

O'Brien, S., & Federici, F. M. (2019). Crisis translation: Considering language needs in multilingual disaster settings. Disaster Prevention and Management: An International Journal, 29(2), 129–143. https://doi.org/10.1108/DPM-11-2018-0373

O'Brien, S., & Federici, F. M. (2020). Crisis Translation: Considering Language Needs in Multilingual Disaster Settings". In *Disaster Prevention and Management*, 29(1). https://doi:10.1108/DPM-112018-0373

O'Brien, S. (2022). Crisis Translation: A snapshot in time. *In INContext: Studies in Translation and Interculturalism, 2(1).* p. 84-108. https://doi.org/10.54754/incontext.v2i1.12

Ojha, A. Kr., Liu, C.-H., Kann, K., Ortega, J., Shatam, S., & Fransen, T. (2021). Findings of the LoResMT 2021 shared task on COVID and sign language for low-resource languages. In J. Ortega, A. Kr. Ojha, K. Kann, & C.-H. Liu (Eds.), *Proceedings of the 4th workshop on technologies for MT of low resource languages (LoResMT2021)* (pp. 114–123).

Association for Machine Translation in the Americas. https://aclanthology.org/2021.mtsummit-loresmt.11/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Open Source Initiative (n.d.) *The Open Source AI Definition – 1.0*. Retrieved February 07, 2025, from https://opensource.org/ai/open-source-ai-definition

Otal, H. T., Stern, E., & Canbaz, M. A. (2024). LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. *2024 IEEE Conference on Artificial Intelligence (CAI)*, 851–859. https://doi.org/10.1109/CAI59869.2024.00159

Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling neural machine translation. 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. 1–9). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6301

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Öktem, A., DeLuca, E., Bashizi, R., Paquin, E., & Tang, G. (2021). *Congolese Swahili Machine Translation for Humanitarian Response* (arXiv:2103.10734). arXiv. https://doi.org/10.48550/arXiv.2103.10734

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Papineni, K., Roukos. S., Ward, T., Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* Philadelphia, Pennsylvania, 311-318.

Parra Escartín, C., & Moniz, H. (2019). Ethical considerations on the use of machine translation and crowdsourcing in cascading crises. In F. M. Federici & S. O'Brien (Eds.), *Translation in Cascading Crises* (1st ed., pp. 132–151). Routledge. https://doi.org/10.4324/9780429341052-7

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon emissions and large neural network training*. https://arxiv.org/abs/2104.10350

Perry, R.W. (2007). What Is a Disaster?. In: Handbook of Disaster Research. Handbooks of Sociology and Social Research. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32353-4_1

Perry, R. W. (2018). Defining Disaster: An Evolving Concept. In H. Rodríguez, W. Donner, & J. E. Trainor (Eds.), *Handbook of Disaster Research* (pp. 3–22). Springer International Publishing. https://doi.org/10.1007/978-3-319-63254-4_1

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (arXiv:1802.05365). arXiv. https://doi.org/10.48550/arXiv.1802.05365

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. https://doi.org/10.18653/v1/W15-3049

Popovic, M. (2018). *Error classification and analysis for machine translation quality assessment*. https://doi.org/10.1007/978-3-319-91241-7_7

Post, M. (2018). 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. https://doi.org/10.18653/v1/W18-6319

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., & Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In M. Walker, H. Ji, & A. Stent (Eds.), Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers) (pp. 529–535). Association for Computational Linguistics.

https://doi.org/10.18653/v1/N18-2084

Quarantelli, E. L. (1998). *What is a disaster? - Perspectives on the question*. Routledge.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2024). *Direct preference optimization: Your language model is secretly a reward model*. https://arxiv.org/abs/2305.18290

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. https://doi.org/10.48550/arXiv.1910.10683

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Online: Association for Computational Linguistics, 2685–2702.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., & Martins, A. F. T. (2022). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, … M. Zampieri (Eds.), *Proceedings of the seventh conference on machine translation (WMT)* (pp. 578–585). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.52/

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-networks*. https://arxiv.org/abs/1908.10084

Riekters, M. (2018). Impact of corpora quality on neural machine translation. In *Human language technologies–The Baltic perspective* (pp. 126-133). IOS Press.

Rios, M., Chereji, R.-M., Secara, A., & Ciobanu, D. (2022). *Impact of domain-adapted multilingual neural machine translation in the medical domain*.

https://arxiv.org/abs/2212.02143

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). *ChatGPT MT: Competitive for high- (but not low-) resource languages*. https://arxiv.org/abs/2309.07423

Roussis, D. G. (2022). Building End-to-End Neural Machine Translation Systems for Crisis Scenarios: The Case of COVID-19. [Master's thesis, National and Kapodistrian University of Athens] https://core.ac.uk/download/pdf/580001711.pdf

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, *88*(8), 1270–1278. https://doi.org/10.1109/5.880083

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., … Rush, A. M. (2022). *Multitask Prompted Training Enables Zero-Shot Task Generalization* (arXiv:2110.08207). arXiv. https://doi.org/10.48550/arXiv.2110.08207

Saunders, D. (2022). *Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey* (arXiv:2104.06951). arXiv. https://doi.org/10.48550/arXiv.2104.06951

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., … Wolf, T. (2023). *BLOOM: a 176B-parameter open-access multilingual language model*. https://inria.hal.science/hal-03850124

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of The Acm*, *63*(12), 54–63. https://doi.org/10.1145/3381831

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7881–7892). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Sennrich, R., & Zhang, B. (2019). *Revisiting Low-Resource Neural Machine Translation: A Case Study* (arXiv:1905.11901). arXiv. https://doi.org/10.48550/arXiv.1905.11901

Sgroi, S. C. (July 29, 2020) *Il Covid o la Covid? Ma è un problema?* Treccani

https://www.treccani.it/magazine/lingua_italiana/articoli/scritto_e_parlato/Covid.html

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. The Bell System Technical Journal. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shu, P., Chen, J., Liu, Z., Wang, H., Wu, Z., Zhong, T., Li, Y., Zhao, H., Jiang, H., Pan, Y., Zhou, Y., Owl, C., Zhai, X., Liu, N., Saunt, C., & Liu, T. (2024). *Transcending Language Boundaries: Harnessing LLMs for Low-Resource Language Translation* (arXiv:2411.11295). arXiv. https://doi.org/10.48550/arXiv.2411.11295

Smith, C. S. (2024, January 1). What Large Models Cost You – There Is No Free AI Lunch. *Forbes*. Retrieved February 5, 2025 from https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25/

Son, J., & Kim, B. (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information*, *14*(10), 574. https://doi.org/10.3390/info14100574

Sun, K., & Dredze, M. (2025). *Amuro and Char: Analyzing the Relationship between Pre-Training and Fine-Tuning of Large Language Models* (arXiv:2408.06663). arXiv. https://doi.org/10.48550/arXiv.2408.06663

Tarkowski (2023, August 11). *The Mirage of Open-Source AI: Analyzing Meta's Llama 2 Release Strategy – Open Future.* Open Future. Retrieved 07 February, 2025, from https://openfuture.eu/blog/the-mirage-of-open-source-ai-analyzing-metas-llama-2-release-strategy/

Tarkowski (2032, October 25). *Falcon 180B, open source AI and control over compute – Open Future*. Open Future. Retrieved 07 February, 2025, from https://openfuture.eu/blog/falcon-180b-open-source-ai-and-control-over-compute/

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5-21.

Tay, Y., Wei, J., Chung, H. W., Tran, V. Q., So, D. R., Shakeri, S., Garcia, X., Zheng, H. S., Rao, J., Chowdhery, A., Zhou, D., Metzler, D., Petrov, S., Houlsby, N., Le, Q. V., & Dehghani, M. (2022). *Transcending Scaling Laws with 0.1% Extra Compute* (arXiv:2210.11399). arXiv. https://doi.org/10.48550/arXiv.2210.11399

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., & Stojnic, R. (2022). *Galactica: A Large Language Model for Science* (arXiv:2211.09085). arXiv. https://doi.org/10.48550/arXiv.2211.09085

Thrun, S., & Pratt, L. (1998). Learning to Learn: Introduction and Overview. In S. Thrun & L. Pratt (Eds.), *Learning to Learn* (pp. 3–17). Springer US. https://doi.org/10.1007/978-1-4615-5529-2_1

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation, 2012*(1) 2214-2218

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT–building open translation services for the world. In Proceedings of the 22nd annual conference of the European Association for Machine Translation (pp. 479-480).

Tokenization. (n.d.). The Stanford Natural Language Processing Group. Retrieved January 14, 2025 from: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Translation Initiative for COVID-19 (n.d.) ["Translation Initiative for COVID-19"]. Retrieved 17/11/2024 from: https://tico-19.github.io/

Treccani (n.d.) Condizióne - Significato ed etimologia. In *Vocabolario on line - Treccani*. Retrieve February 21, 2025 from: https://www.treccani.it/vocabolario/condizione/

TWB glossary for COVID-19 (n.d.) Translators Without Borders Retrieved December 4, 2024 from: https://translatorswithoutborders.org/resource/twb-covid-19-glossary/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv.

https://doi.org/10.48550/arXiv.1706.03762

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., & Foster, G. (2023). Prompting PaLM for Translation: Assessing Strategies and Performance. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15406–15427. https://doi.org/10.18653/v1/2023.acl-long.859

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-Level Machine Translation with Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16646–16661. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Wang, P. (2019). Translation in the COVID-19 health emergency in Wuhan. In The Journal of Internationalization and Localization (Vol. 6, Issue 2, pp. 86–107). John Benjamins. https://doi.org/10.1075/jial.00014.wan

Wassie, A. K., Molaei, M., & Moslem, Y. (2024). *Domain-Specific Translation with Open-Source Large Language Models: Resource-Oriented Analysis* (arXiv:2412.05862). arXiv. https://doi.org/10.48550/arXiv.2412.05862

Wang, W., Peter, J.-T., Rosendahl, H., & Ney, H. (2016). CharacTer: Translation edit rate on character level. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. J. Yepes, A. Névéol, M. Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann, & M. Turchi (Eds.), *Proceedings of the first conference on machine translation: Volume 2, shared task papers* (pp. 505–510). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2342

Wang, Z.ichong, Chu, Z., Doan, T. V., Ni, S., Yang, M., & Zhang, W. (2024). History, development, and principles of large language models: An introductory survey. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00583-7

Wang, Ziheng., Wohlwend, J., & Lei, T. (2020). Structured pruning of large language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.496

Wassie, A. K., Molaei, M., & Moslem, Y. (2024). *Domain-Specific Translation with Open-Source Large Language Models: Resource-Oriented Analysis* (arXiv:2412.05862). arXiv. https://doi.org/10.48550/arXiv.2412.05862

Way, A. (2018). *Quality expectations of machine translation*. in J. Moorkens, S. Castilho, F.

Gaspari, and S. Doherty (eds). *Translation Quality Assessment: From Principles to Practice* (pp. 159-178), Cham: Springer International Publishing,

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. https://arxiv.org/abs/2201.11903

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Wołk, K., & Marasek, K. (2015). Neural-based machine translation for medical text domain. based on European Medicines Agency Leaflet texts. Procedia Computer Science, 64, 2–9. https://doi.org/10.1016/j.procs.2015.08.456

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A Large Language Model for Finance* (arXiv:2303.17564). arXiv. https://doi.org/10.48550/arXiv.2303.17564

Zhu-chen, & Feng, J. (2018). Development and Application of Artificial Neural Network. *Wireless Personal Communications*, *102*(2), 1645–1656. https://doi.org/10.1007/s11277-017-5224-x

Xu, H., Kim, Y. J., Sharaf, A., & Awadalla, H. H. (2024). *A paradigm shift in machine translation: Boosting translation performance of large language models*. https://arxiv.org/abs/2309.11674

Xu, H. (2021). Transformer-based NMT: modeling, training and implementation [Doctoral dissertation, Universität des Saarlandes] SciDok – Der Wissenschaftsserver der Universität des Saarlandes doi:10.22028/D291-34998

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Yi, Y., Lagniton, P. N. P., Ye, S., Li, E., & Xu, R. H. (2020). COVID-19: what has been learned and to be learned about the novel coronavirus disease. *International journal of biological sciences*, *16*(10), 1753–1766. https://doi.org/10.7150/ijbs.45134

Zafar, I., Tzanidou, G., Burton, R., Patel, N., & Araujo, L. (2018). *Hands-on convolutional neural networks with TensorFlow: Solve computer vision problems with modeling in TensorFlow and Python*. Packt Publishing Ltd.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Zhang, P., Dong, Y., & Tang, J. (2023). *GLM-130B: An Open Bilingual Pre-trained Model* (arXiv:2210.02414). arXiv. https://doi.org/10.48550/arXiv.2210.02414

Zhang, Biao, Ghorbani, B., Bapna, A., Cheng, Y., Garcia, X., Shen, J., & Firat, O. (2022). *Examining scaling and transfer of language model architectures for machine translation*. https://arxiv.org/abs/2202.00528

Zhang, Biao, Haddow, B., & Birch, A. (2023). Prompting large language model for machine translation: A case study. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 41092–41110). PMLR. https://proceedings.mlr.press/v202/zhang23m.html

Zhang, F., Abudayyeh, O. O., & Gootenberg, J. S. (2020, March 15) *A protocol for detection of COVID-19 using CRISPR diagnostics.* Retrieved February 24, 2025 from: https://www.broadinstitute.org/files/publications/special/COVID-19%20detection%20(updated).pdf

Zhang, Jie, & Wu, Yuqin (2020). Providing multilingual logistics communication in COVID-19 disaster relief. Multilingua, 39(5), 517-528. https://doi.org/10.1515/multi-2020-0110

Zhang, Jiajun, & Zong, C. (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, *63*(10), 2028-2050. https://doi.org/10.1007/s11431-020-1632-x

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating text generation with BERT*. https://arxiv.org/abs/1904.09675

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2024). *A Survey of Large Language Models* (arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zhu, D., Chen, P., Zhang, M., Haddow, B., Shen, X., & Klakow, D. (2024). *Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice?*

https://arxiv.org/abs/2404.14122

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2024). *Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis* (arXiv:2304.04675). arXiv. https://doi.org/10.48550/arXiv.2304.04675

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). *Fine-tuning language models from human preferences*. https://arxiv.org/abs/1909.08593

Zouhar, V., Chen, P., Lam, T. K., Moghe, N., & Haddow, B. (2024). 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. https://doi.org/10.18653/v1/2024.wmt-1.121