Arabic to Romanian Machine Translation: A Case Study on Distant Language Pairs

Ioan Alexandru Hirică [1] and **Ștefana Tăbușcă** [1,2] and **Sergiu Nisioi** [1] *
[1] Human Language Technologies Research Center,
Faculty of Mathematics and Computer Science
[2] Interdisciplinary School of Doctoral Studies,
University of Bucharest
alexandru.hirica2001@gmail.com, stefana.tabusca@s.unibuc.ro,
sergiu.nisioi@unibuc.ro

Abstract

This paper investigates machine translation between two linguistically distant languages, Arabic and Romanian, with a focus on translating from Arabic to Romanian. Dataset cleaning techniques are addressed, offering insights on the impact of translation for a language pair with limited resources. Using publicly available corpora (e.g., OPUS) and manually translated diplomatic texts, filtering methods are applied, such as duplicate removal, embedding similarity analysis (LEALLA), and Large Language Model (LLM)-based validation (Gemini-flash-002). Transformer models are trained and evaluated with diverse preprocessing pipelines that incorporate subword tokenization. Additionally, the performance of two fine-tuned LLMs is assessed for this task and is compared to their pre-trained counterparts. Despite computational limitations, the results emphasize the importance of targeted preprocessing and model adaptation in improving Arabic-Romanian translation quality. Resources and code are available at: https://github.com/HiricaAlexandru/Arabic-to-Romanian-MT

1 Introduction

Machine Translation (MT) has made significant advancements in recent years, achieving strong scores across a variety of evaluation methodologies (Kocmi et al., 2024). Currently, there are two paradigms for developing MT models: 1. by creating sequence-to-sequence translation models with deep neural networks or 2. by fine-tuning generic Large Language Models (LLMs) for the translation. Both approaches are trained massively on multilingual data to cover as many languages as possible. However, most existing research and evaluation frameworks have focused on high-resource languages or high- to low-resource language pairs, leaving distant language pairs relatively underexplored (Isozaki et al., 2010; Tayir et al., 2024). These pairs are characterized by minimal linguistic and cultural contact and pose several challenges, including weak language transfer due to limited lexical and syntactic overlap, a scarcity of high-quality parallel corpora, and difficulties in developing effective bitext mining tools.

Despite these challenges, distant language pairs can offer an interesting research case from a linguistic and computational point of view. Addressing their translation requires novel approaches to dataset creation, model training, and evaluation. Given that available corpora are often limited to official documents (e.g., Embassy translations) or may contain synthetic or low-quality data, robust cleaning and deduplication strategies are needed.

In this work, we train and evaluate models that can translate from Arabic - an Afroasiatic, Semitic language - and Romanian an Eastern Romance language. The two have had little historical contact, do not share the same scripts and have different grammatical properties. According to the parametric comparison method of comparing languages (Longobardi, 2003; Ceolin et al., 2021), both Romanian and Arabic have rich morphological systems with grammaticalized agreement, number, and definiteness marking. In terms of differences, Romanian retains a partially grammaticalized case system and follows an SVO word order, has definite matching pronominal possessives and genitives, as well as genitive licensing iteration. Modern Standard Arabic has largely lost case distinctions, tends to follow a VSO structure (Hewitt, 2006) and features head marking.

The growing popularity of Arabic countries as tourism destinations and employment hubs highlights the importance of learning the Arabic-Romanian language pair to strengthen political and economic ties. According to Kanakri and Ionescu

*Corresponding authors.

Proceedings of Recent Advances in Natural Language Processing,pages 423–432
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-051

(2010), around 1,500 mixed Romanian families live in Jordan, while Libera (2025) reports 6,444 Romanians in the UAE, 2,000 in Qatar, and 850 in Kuwait, further emphasizing the potential benefits of this language pair for fostering international relations.

This paper presents a comprehensive evaluation of machine translation systems with several key contributions. First, we train models from scratch using various pruning strategies, demonstrating that a more selective "less is more" approach is beneficial for translation performance. To assess the capability of large language models (LLMs) in this context, we fine-tune a generic LLM and compare its performance to our established baselines. We further address the challenge of Arabic dialectal variation by evaluating multiple models across diverse test sets, including in-domain, Out-of-Domain, and FLORES+ benchmarks. To support research in this area, we release a cleaned Arabic - Romanian parallel dataset, filling a gap in available resources for this language pair. Finally, we perform an error analysis to better understand the model behaviour and translation quality. Through this research, we contribute to the broader effort of improving MT for distant language pairs, emphasizing the need for targeted strategies in data processing and model optimization.

2 Related Work

The definition of low-resource language is task-specific and can be influenced by factors such as a lack of high-quality parallel data, uneven availability across different domains and dialects, and historical underinvestment in data collection and NLP research (Nigatu et al., 2024). Arabic and Romanian, despite having extensive speakers and monolingual data, can be considered low-resource pairs for machine translation due to the complete absence of high-quality parallel datasets.

Ethical considerations are also crucial when working with low-resource languages, as highlighted by Haroutunian (2022), who recommends collaboration with stakeholders to develop more reliable machine translation tools and consider other forms of language technology, such as strong content moderation policies, alongside machine translation to mitigate potential harms. To address these concerns, our paper aims to incorporate careful dataset curation and analyze the dialectal variation in Arabic data to improve model performance. Both in the aforementioned work and an early study done by Dodita (2012), limitations of automatic translation systems like Google Translate are highlighted, emphasizing the need for rigorous dataset preprocessing.

Previous research on Arabic MT mainly looks at the difference in dialects and various translation strategies; Harrat et al. (2019) compile a meta-study where they explore Arabic dialect translation, highlighting as well that most contributions consider only English as a target language, thus outlining the need for a broader coverage. Baniata et al. (2021) address these dialect challenges by focusing on subword tokenization to handle morphological complexity.

Multilingual machine translation (MT) solutions have traditionally been used to cover efforts for distant language pairs. However, researchers such as Fan et al. (2021) have proposed a many-to-many system that directly translates between non-English language pairs using shared representations and large-scale data. This approach shows improved translation quality for low-resource directions compared to conventional English-pivot methods, highlighting the benefits of a more inclusive and linguistically diverse translation paradigm. Multilingual NMT systems can also share representations across languages to improve performance for underrepresented pairs as demonstrated by Aharoni et al. (2019), a strategy that underpins our approach to Arabic-Romanian translation. The No Language Left Behind (NLLB) model (Team et al., 2022; NLLB Team et al., 2024) shows promise in addressing translation challenges across a wide range of languages. The authors also introduce the FLORES multilingual evaluation benchmark covering 200+ languages, which provides consistent test sets for assessing and improving low-resource language pairs.

The current state-of-the-art techniques for machine translation are dominated by large-language-model (LLM) approaches, with many challenges remaining unsolved, particularly in low-resource and domain-specific translation, as outlined in the WMT24 Shared Task report (Kocmi et al., 2024) as well. To our knowledge this is the first work to address Arabic - Romanian machine translation as a case study of distant language pairs.

3 Dataset

3.1 Evaluation Dataset

To evaluate the models, we use three types of corpora: a custom built high-quality Out-of-Domain dataset, a in-domain dataset created by randomly splitting the training set, and the FLORES+ (NLLB Team et al., 2024) benchmark. The Out-of-Domain Arabic-Romanian translation corpus was created by selecting manually translated texts from diplomatic agreements between Romania and Saudi Arabia, widely recognized documents published by the United Nations, such as those on human rights and refugee status, as well as content from the official X account of Saudi Arabia's embassy in Romania. Since the texts are legal in nature, alignment is straightforward, as they are divided into sentences that follow a formal and structured pattern. The documents are written in official Modern Standard Arabic.

The number of pairs for each dataset type is visible in Table 1.

Document	**Source**	**Sents.**
Saudi-Romania Econ.	State AGT.	189
Geneva Convention	UN	330
Children Rights	UN	217
Economic Rights	UN	124
Human Rights	UN	79
Refugee Status	UN	177
Saudi Embassy	X	86
Total OOD		1202
FLORES+	Huggingface	997
In-Domain	Training dat.	2100
Total	-	**4299**

Table 1: Number of pairs for each data subset. The first seven rows are Out-of-Domain (OOD) samples manually collected for this experiment.

Additionally, we construct an in-domain dataset by selecting 2,100 samples evenly distributed from each source of the training dataset, derived from the dataset after duplicate removal and length filtering, described in the following section.

3.2 Training Dataset

For training an NMT model from scratch we combine together several datasets available online covering Romanian and Arabic: OPUS-subtitles (Tiedemann, 2009), CC-Matrix (Schwenk et al., 2019), and the data originally used for training NLLB (Team et al., 2022) were used. Various techniques are applied to filter the datasets, aiming to enhance their quality.

The first filtering steps is to remove entries with fewer than 5 words or more than 50 words. Since the NLLB dataset is also created from similar sources (OPUS, CC-Matrix), we implement a fuzzy duplicate removal method using Min-Hash (Broder, 1997) which removes sentences with any Jaccard similarity of at least 0.8.

Dataset Filtering Method	**Number of Pairs**
Removed dup. + len. Filter	13.57M
LEALLA Filtering	10.84M
Gemini Filtering	7.44M
Dialect Removal Filter	5.44M

Table 2: Number of dataset pairs (in millions) remaining after each filtering stage for NMT training

After this initial filtering, we applied LEALLA-base (Mao and Nakagawa, 2023), a lightweight sentence embedding model distilled from LaBSE (Feng et al., 2020). This model provides comparable performance to LaBSE while being significantly smaller in size. Employing this approach, Arabic-Romanian sentence pairs were encoded to generate embeddings for each sentence. The embeddings were normalized using the L2 norm, and cosine similarity was calculated. Sentence pairs with a similarity score below 0.4 were deemed poor translations and removed from the dataset.

Further refinement was carried out using a Large Language Model (LLM) to eliminate ungrammatical or nonsensical sentences. Gemini-flash-002 (Team et al., 2024) is chosen for this task due to its cost-effectiveness and comparable performance to GPT-4. This decision is based on an evaluation of both models on a test dataset, where Arabic sentences were translated into Romanian.

The final data cleaning step is to remove all Arabic sentences that were not in Modern Standard Arabic. This is done to evaluate the impact of dialects in the training, especially since the OOD dataset consists in Modern Standard Arabic (MSA). To distinguish between different Arabic Dialects, we use a classification model introduced by Attieh and Hassan (2022) and built upon AraBERT (Antoun et al., 2020). The model can distinguish between MSA and a wide range of regional dialects. This analysis is conducted to compare with the large corpus used for training the translation

models. MSA is the predominant variety in the training corpus, accounting for over 70% of the dataset, followed by Tunisian Arabic (TN) at just above 10%. Table 2 contains the number of pairs remaining after the application of each filtering method. The largest dataset, even after the removal of duplicates and longer sentences, is still more than twice the size of the data obtained after all filtering methods have been applied.

4 Training and Comparing NMTs

To compare the performance of different machine translation models, we employ several evaluation metrics, including string-based metrics and embedding-based similarity metrics like BERTScore (Zhang et al., 2019). The values of these are reported in the Appendix available in our official repository[1]. In this paper we incorporate reference-based evaluation through MetricX-23-XL (Juraska et al., 2023), and reference-less evaluation using CometKiwi-DA-XL (Rei et al., 2023) by adopting AutoRank following the methodology used in WMT24 (Kocmi et al., 2024). AutoRank uses two top-performing metrics: MetricX-23-XL (Juraska et al., 2023), a reference-based metric built on the mT5 model, and CometKiwi-DA-XL (Rei et al., 2023), a quality estimation metric built on the XLM-R XL model. These two distinct metrics were chosen to minimize bias and potential issues. To create the ranking, scores from both metrics are linearly scaled to a range between 1 and the number of systems in a given language pair. The normalized scores are then averaged to produce the final metric.

4.1 Training Transformers from Scratch

To compare the impact of different filtering methods, we train a transformer system from scratch on the filtered datasets. We use the OpenNMT (Klein et al., 2017) framework to train a transformer model with 6 encoder-decoder layers, 8 heads, a hidden size of 512 and a vocabulary of 50K tokens (Vaswani, 2017). The model contains a total of 100M trainable parameters.

The results of these models can be seen in Table 3 with prefix *Transf*, indicating that data filtering can substantially enhance translation quality.

The results are not conclusive on all test sets, i.e., on the out-of-domain data the models trained from scratch are below LLM-based and NLLB models; on the FLORES+ the models are better than NLLB and on the in-domain data the models are slightly below proprietary LLMS (GPT-4, Gemini).

Notably, filtering the dataset using Gemini has a negative impact and the best results tend to be the ones where only duplicates are being removed. The only exception is the out-of-domain data, where the removal of dialects yields the best transformer-from-scratch model, confirming that the OOD data consists mostly of Modern Standard Arabic.

4.2 Pre-trained Models

JAIS (Sengupta et al., 2023) is a bilingual large language model (LLM) created at MBZUAI trained on Arabic and English texts, currently considered the best "Arabic-speaking" LLM. However, the model performs poorly on Romanian and we use Jais to translate into English (pivot language). Subsequently, to translate into Romanian, a pre-trained transformer model, trained on an English-Romanian corpus, was employed (Tiedemann, 2020). This ensemble performed similarly to the models trained from scratch, the only significant performance increase being on social media posts from the Saudi Embassy in Romania.Overall, this approach performs mediocrely on all test sets and on the in-domain data generating extremely poor translations.

NLLB-600M (Team et al., 2022) is a transformer-based architecture distilled to 600 million parameters from a 3.3B Sparsely Gated Mixture of Experts designed for massively multilingual translation.

We compare both the the Pre-trained NLLB and a Fine-tuned version on the dataset with the dialects removed (the smallest version). For finetuning, the evaluation metric is BERTScore and early stopping is employed, so training would halt if the score did not improve after 6000 steps. The model reached its best performance after just 1800 steps, with a batch size of 16.

The out-of-domain dataset is the only one where the NLLB models outperform our models trained from scratch (see Table 3). The Out-of-Domain dataset is mostly made of legal documents, unlike the FLORES+ and in-domain datasets which primarily consist also of informal sentences, such as those found in film subtitles.

[1] https://github.com/HiricaAlexandru/Arabic-to-Romanian-MT/

Dataset	Model	CometKiwi ↑	MetricX ↓	AutoRank ↓
Out-of-Domain	Human	0.615	0.019	2.469
	Gemini-flash-002	0.624	1.100	2.699
	GPT-4	0.624	1.472	2.807
	Fine-tuned RoLLama3.1-8b	0.573	2.890	3.733
	Fine-tuned RoMistral-7b	0.557	2.926	3.905
	Pre-trained RoLLama3.1-8b	0.530	3.764	4.421
	Jais+Transf	0.529	4.036	4.513
	Pre-trained NLLB-600M	0.535	4.443	4.577
	Fine-tuned NLLB-600M	0.537	4.703	4.638
	Transf (Dialects removed)	0.535	5.554	4.908
	Pre-trained RoMistral-7b	0.486	4.471	5.067
	Transf (Duplicates removed)	0.495	5.017	5.140
	Transf (Gemini filter)	0.477	5.099	5.350
	Transf (Lealla filter)	0.469	5.265	5.473
Dataset	**Model**	**CometKiwi ↑**	**MetricX ↓**	**AutoRank ↓**
FLORES+	Human	0.687	0.022	1.748
	GPT-4	0.736	0.986	1.551
	Gemini-flash-002	0.729	0.958	1.614
	Fine-tuned RoLLama3.1-8b	0.664	1.979	2.559
	Fine-tuned RoMistral-7b	0.653	2.496	2.823
	Pre-trained RoLLama3.1-8b	0.644	2.606	2.950
	Transf (Duplicates removed)	0.633	2.922	3.148
	Transf (Lealla filter)	0.632	3.114	3.216
	Fine-tuned NLLB-600M	0.628	3.045	3.242
	Jais+Transf	0.603	2.720	3.394
	Pre-trained NLLB-600M	0.617	3.336	3.435
	Transf (Gemini filter)	0.615	3.309	3.442
	Transf (Dialects removed)	0.610	3.260	3.478
	Pre-trained RoMistral-7b	0.534	4.597	4.632
Dataset	**Model**	**CometKiwi ↑**	**MetricX ↓**	**AutoRank ↓**
In domain	Human	0.571	0.033	2.901
	GPT-4	0.693	1.415	2.110
	Gemini-flash-002	0.681	1.409	2.227
	Transf (Duplicates removed)	0.625	1.798	2.894
	Transf (Lealla filter)	0.620	1.849	2.963
	Transf (Dialects removed)	0.618	1.865	2.989
	Fine-tuned RoMistral-7b	0.623	2.089	3.005
	Transf (Gemini filter)	0.616	1.874	3.011
	Fine-tuned NLLB-600M	0.611	2.081	3.117
	Pre-trained NLLB-600M	0.600	2.241	3.280
	Fine-tuned RoLLama3.1-8b	0.586	2.871	3.600
	Pre-trained RoLLama3.1-8b	0.584	3.097	3.689
	Pre-trained RoMistral-7b	0.531	3.954	4.474
	Jais+Transf	0.140	18.700	12.745

Table 3: Aggregated results for all models. Saudi-Romania Econ., Geneva Convention, Children Rights, Economic Rights, Human Rights, Saudi Embassy, Refugee Status subsets represent Out-of-Domain data (OOD), results are averaged. Values have been rounded - to 3 decimals. The models are sorted by AutoRank score. MetricX and AutoRank are metrics where lower values are better, while CometKiwi is a reference-less similarity score where higher values are better.

4.3 Fine-tuning LLMs

RoLLama3.1-8b-Instruct (Masala et al., 2024) is a model that has been instruction tuned with Romanian data. We fine-tune it using LoRA (Low-Rank Adaptation (Hu et al., 2022)) on a sample of 1 million examples extracted from the dataset after applying the Dialect Removal Filter. The model is loaded in 8-bit precision, utilizing the AdamW optimizer, LoRA modules are applied to the LM head and Embedding tokens layers, with a rank of 16.

A second LLM fine-tuning uses **RoMistral-7b-Instruct** (Masala et al., 2024) with the same parameters as in the fine-tuning of RoLLama-3.1.

Table 3 shows that these models achieve the best evaluation scores on FLORES+ and Out-of-Domain datasets from all our open-sources models. Fine-tuning is beneficial for all models, but even the base pre-trained LLama 3.1 obtains a strong performance. However, the same LLM models significantly underperform on the custom in-domain dataset, where transformers trained from scratch are leading in evaluation.

Open source models are surpassed only by proprietary systems like GPT-4 and Gemini-flash-002 by a considerable margin. Since the training data of proprietary models is not known, the comparison is not entirely fair.

Proprietary Large Language Models (LLMs) such as Gemini-flash-002 (Team et al., 2024) and

GPT-4 (Achiam et al., 2023) are used in single shot to translate sentences. Both models use the same prompt and significantly outperformed other approaches on the evaluation datasets and their performance comparable is at a close (insignificant) margin. The models are very large and may have seen a lot of web data, which may have included our test sets, therefore the comparison is not entirely just. For example, Gemini demonstrates a substantial advantage in BLEU scores over GPT-4 on the Saudi Embassy dataset,[2] meaning that the string overlap with the true translation is very high. Furthermore, the transformers trained from scratch show the best BLEU and chrF++ scores on the in-domain datasets, indicating that a high string similarity could be a source of overfitting on a particular type of language / dataset.

5 Discussion

The rows prefixed with "Human" in Table 3 reflect the scores assigned by the reference based evaluation metric to the human-translated baseline. For MetricX we use the same texts as reference and hypothesis and we observe that MetricX does not obtain a perfect zero score when the references and hypotheses are identical, but a lower bound of 0.033.[3]

The CometKiwi quality estimation metric does not assign the best scores to the Human translations with a difference as large as 0.122 between the in-domain reference Human translations and the GPT-4 translations. This clearly shows that the metric is biased towards more literal translations and, possibly, that the efficiency of the metric for the Arabic-Romanian language pair is questionable. Regardless of this bias against human translations, the overall ranking from AutoRank of machine translation systems is not significantly different from MetricX with the exception of small differences between Gemini and GPT-4.

Figure 1 presents the AutoRank scores for various models and each out-of-domain sub-corpus. The selection includes the best-performing model in each proprietary category: the top proprietary LLM, the best locally run LLM, Fine-tuned NLLB-600M, and a Transformer-based architecture trained on the smallest dataset. This plot provides a comparison for each test subset, highlighting instances where GPT-4 surpasses the Human translations.

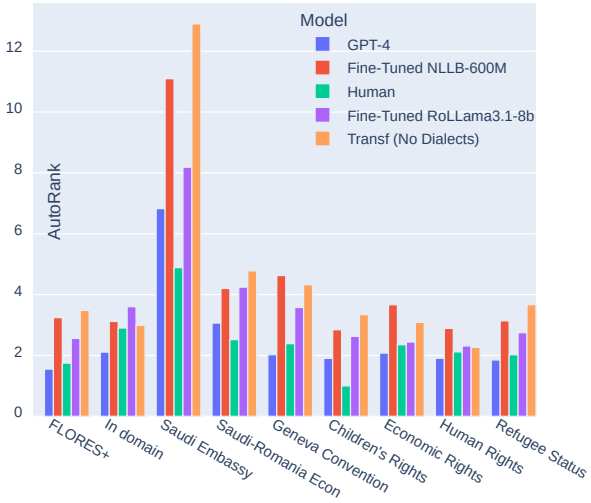

Figure 1: AutoRank (lower is better) scores of selected models on each Out-of-Domain dataset. Saudi Embassy documents, Saudi-Romanian Econ., and Children's Rights are the only three subsets where the human translations are ranked significantly better than GPT-4. The differences come from the bias of CometKiwi against human translations for Arabic-Romanian language pair. Additionally, the Transformer trained from scratch performs on par (or even better) than the Fine-tuned NLLB model on multiple test sets.

Notably, on the Saudi Embassy dataset, Human translations achieves a much higher score than GPT-4, likely due to its manual sourcing from X, featuring recent data that GPT-4 would not have processed during training. Similarly Children's Rights and Saudi-Romanian Economic cooperation agreements are datasets where the human translations are ranked better than automatic translations.

For locally run models, Fine-tuned RoLLAMA-3.1-8b appears to be an excellent choice, as it outperforms all other locally run models. However, NLLB-600M is not far behind and could be a viable option for systems with limited resources, given its smaller size of 600 million parameters compared to RoLLAMA's 8 billion parameters.

6 Error Analysis

An automatic analysis of machine translation output quality is conducted across a diverse set of Arabic–Romanian models using the XCOMET-XL framework.[4] This approach provides both scalar

[2] See Appendix in our github repository: https://github.com/HiricaAlexandru/Arabic-to-Romanian-MT/

[3] We hypothesize that a MetricX difference smaller than 0.066 between systems is not relevant.

[4] https://huggingface.co/Unbabel/XCOMET-XL

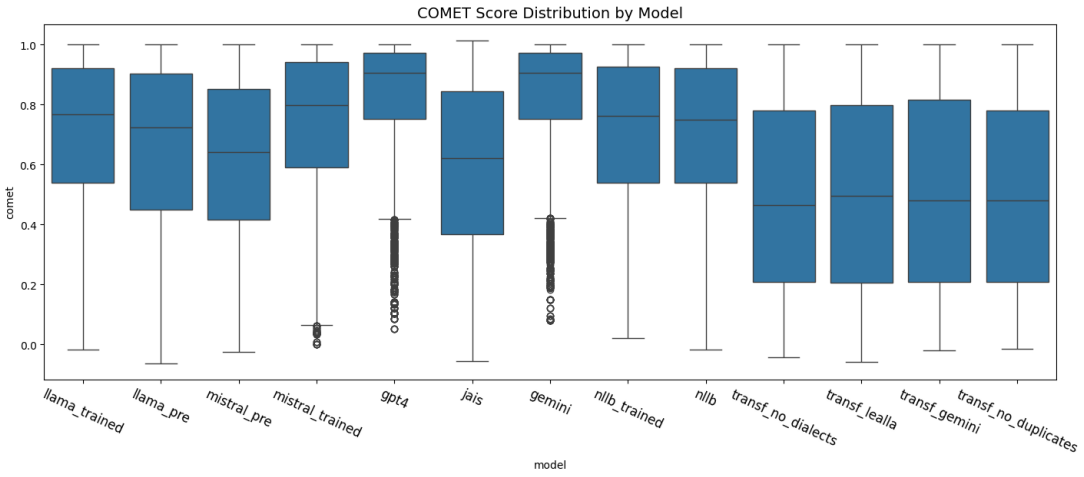

Figure 2: CometKiwi-DA-XL score distribution across all evaluated models.

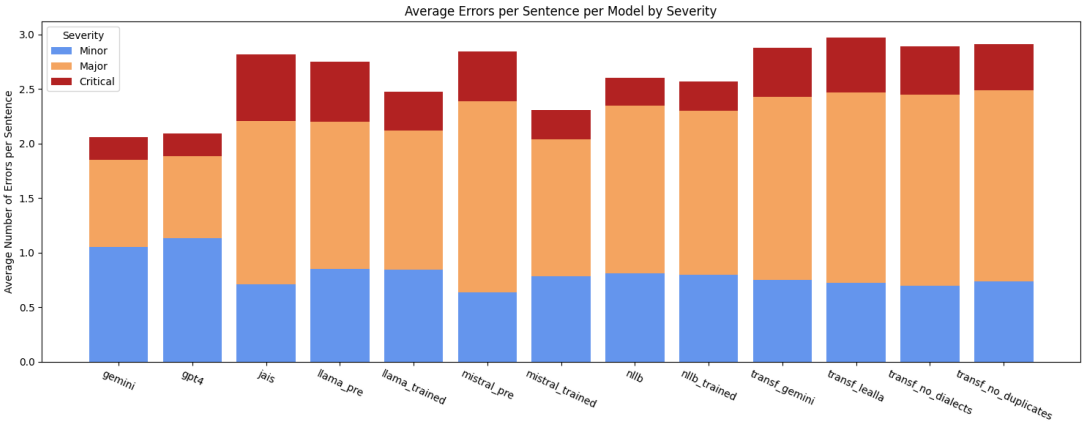

Figure 3: Distribution of error severity levels per model.

quality estimations and token-level error segmentations with associated confidence scores. Distributions of XCOMET-XL score and error severity for each model are visible in Figure 2 and Figure 3.

Across all evaluated models, the average number of errors per sentence was between 2 and 3, as visible in Figure 3; however, the distribution and nature of translation errors varied significantly depending on both the dataset and the architecture of the model. Proprietary models like Gemini and GPT4 have a significantly low number of critical errors, and a noticeable balance of major and minor errors, suggesting a more uniform and qualitative translation. However, for the Transformer-based models trained from scratch, as well as for Jais and the non-finetuned versions of Llama, NLLB, and Mistral, major and critical errors are more common on average, indicating a higher risk of serious disparities. These values are in agreement with the results obtained for the metrics in Table 3.

A more in-depth analysis of the errors in the OOD datasets reveals, however, interesting observations. For instance, Gemini exhibits a large volume of low-confidence errors classified as "critical" or "major" when applied to formal diplomatic text (e.g., Saudi Embassy and Geneva Convention). In these cases, a pattern can be observed: mistranslations often involve the misrendering of official titles, named entities, or domain-specific expressions. This suggests that while the general fluency of the output may be preserved (as partially reflected by moderate XCOMET-XL scores), semantic fidelity is somewhat compromised in areas important to official translations.

Some models attain relatively high XCOMET-XL scores even in the presence of severe errors. This discrepancy points to a limitation in how sentence-level scoring alone reflects translation adequacy, particularly when the surface form remains grammatical but semantically distorts the source.

On the other hand, minor errors are frequently associated with high XCOMET-XL values, reflect-

ing issues that are largely stylistic or morphological in nature - such as gender agreement or article usage. These could be addressed through post-editing rather than retraining, especially in production workflows where fluency is acceptable and semantic drift is minimal.

Error confidence scores provide an additional layer of diagnostic value. In critical segments, low confidence often aligns with deeply flawed translations or hallucinations, such as the introduction of fictitious roles ("croitor" Eng. *tailor*, "Slujitor al Slujitorilor" Eng. *servant of servants*) in otherwise formulaic diplomatic statements. This behavior is especially visible in models that were either undertrained or applied out of domain, such as general-purpose transformer models evaluated on official political text.

The error analysis also reveals that models fine-tuned on domain-adjacent data tends to reduce the frequency of severe errors and increased the proportion of accurate renderings in technical and diplomatic subdomains.

7 Conclusions and Future Work

Proprietary systems are performing strongly across different metrics and corpora for Arabic-Romanian language pair and none of the fine-tuned models exceeded these performances.

Our experiments demonstrate that data filtering is an essential step for improving Machine Translation quality, especially for distant language pairs, such as Arabic-to-Romanian. The most important data cleaning process is the removal of duplicates followed by filtering with a cross-language embedding model such as LEALLA. Filtering the data based on prompting and LLM such as Gemini does not bring any significant improvements over LEALLA while removing the Arabic dialects hurts the ability of the model to translate diverse texts. This can lead to incidental high evaluation scores on Modern Standard Arabic official documents (e.g., in Table 3 the best Transformer from scratch model on Out-of-Domain data). Training a transformer model with only 100M parameters from scratch on a filtered corpus—reduced by up to 60%—requires significantly less training time, yet still achieves competitive and sometimes superior results compared to NLLB or even 8-billion-parameter LLMs.

Compared to NLLB models and proprietary large language models (LLMs), the from-scratch approach performs competitively on the FLORES+ and In-Domain datasets, while maintaining lower computational and resource costs. However, such methods have major limitations on out-of-domain data.

We found that quality estimation metrics such as CometKiwi and MetricX can sometimes misjudge human translations. CometKiwi may exhibit biases, occasionally rating human translations as worse than machine-generated ones, while MetricX may fail to assign a perfect score even when the hypothesis is identical to the reference.

Error analysis performed using XCOMET-XL highlights that proprietary LLMs produce fewer critical and major errors, particularly in formal or structured domains. In contrast, the use of English as a pivot via Jais (an Arabic LLM) did not yield significant performance improvements for this language pair.

Looking forward, expanding the evaluation scope to include additional domains, such as medical and technical fields, will enable more comprehensive assessments of model robustness and generalization. Incorporating such domain-specific corpora is essential for improving specialized translation quality.

A second avenue for future work involves deeper analysis of dialectal variation within Arabic, which remains a critical factor in achieving more accurate and culturally aware translations. Finally, reversing the translation direction to Romanian-to-Arabic introduces unique linguistic challenges that warrant exploration. Addressing this direction will contribute to a more holistic understanding of MT performance for low-resource and structurally divergent languages.

Acknowledgments

This research is partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and partially by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007.

References

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

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Joseph Attieh and Fadi Hassan. 2022. Arabic dialect identification and sentiment classification using transformer-based models. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 485–490, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

L. H. Baniata, I. K. E. Ampomah, and S. Parl. 2021. A transformer-based neural machine translation model for arabic dialects that utilizes subword units. *Sensors*, 21(19).

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, Monica-Alexandrina Irimia, Luca Bortolussi, and Andrea Sgarro. 2021. At the boundaries of syntactic prehistory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1824):20200197. Epub 2021 Mar 22.

A. R. Dodita. 2012. Limitations of automatic translations: Google translate (arabic to romanian). *Academia*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54, Dublin, Ireland. Association for Computational Linguistics.

S. Harrat, K. Meftouh, and K. Samili. 2019. Machine translation for arabic dialects: A survey. *Information Processing and Management*, 56(2):262–273.

Steve Hewitt. 2006. Arabic: verb-subject-object or verb-given-new? implications for word order typology. In *Conference on Communication and Information Structure in Spoken Arabic*, pages 1–22.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Mahmoud Kanakri and Violeta Ionescu. 2010. Prototypes of code-switching in the speech of romanian/arabic bilinguals in jordan. *Jordanian Journal for Language studies and literary works*, 2(2):179–194.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Viata Libera. 2025. Pe toate continentele lumii: câți români au părăsit românia pentru a trăi în străinătate. Accessed: 2025-04-18.

Giuseppe Longobardi. 2003. Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, 3(1):101–138.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Mihai Masala, Denis C. Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian-Dan, Andrei Terian-Dan, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "vorbeşti româneşte?" a recipe to train powerful romanian llms with english instructions.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Turghun Tayir, Lin Li, Xiaohui Tao, Mieradilijiang Maimaiti, Ming Li, and Jianquan Liu. 2024. Visual pivoting unsupervised multimodal machine translation in low-resource distant language pairs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5596–5607, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.