UDK 811.163.42'322.4: 811.111'322.4
811.163.42'322.4: 811.134.2'322.4
811.163.42'33+004.85
Izvorni znanstveni članak
Primljeno 1. 4. 2026.
Prihvaćeno za tisak 22. 5. 2026.
https://doi.org/10.22210/suvlin.2026.101.05

Antoni Oliver
Universitat Oberta de Catalunya (UOC)
aoliverg@uoc.edu

Sergi Álvarez–Vidal
Universitat Autònoma de Barcelona (UAB)
Sergi.Alvarez@uab.cat

Croatian Language in the Transition from Neural Machine Translation to Large Language Models

Machine translation (MT) technologies are currently undergoing a paradigm shift, transitioning from specialized Neural Machine Translation (NMT) frameworks to the broader capabilities of Large Language Models (LLMs). This paper examines the current standing of the Croatian language within this technological evolution.

While bilingual NMT models often exhibit high precision, multilingual NMT leverage transfer learning to enhance performance for low–resource language pairs, but with lower performance for high–resource ones. Conversely, LLMs—whether general–purpose or fine–tuned for translation—offer superior multilingual proficiency and context awareness. Unlike NMT, LLMs can process extended discourse, such as full paragraphs or documents, leading to significant improvements in coreference resolution and gender agreement. Despite the substantial computational requirements of LLMs, recent optimization techniques allow for smaller, more efficient versions that maintain high output quality.

This study evaluates the performance of various NMT and LLM architectures specifically for Croatian from/to English and Spanish using several automatic quality evaluation metrics. The findings demonstrate that open–source models can achieve, and occasionally surpass, the quality of Google Translate, a widely used commercial NMT system. Furthermore, while our evaluation focuses on this specific language triad, the multilingual nature of the analysed systems suggests that open–source models provide high–quality translation capabilities for Croatian across dozens, if not hundreds, of language pairs.

Keywords: neural machine translation; large language models; Croatian; English; Spanish

1. Introduction

In recent years, machine translation (MT) has moved beyond its traditional role as a specialized tool, becoming a pervasive component of digital communication infrastructures. This transformation is closely linked to the shift from task–specific Neural Machine Translation (NMT) systems to more general purpose Large Language Models (LLMs), which integrate translation within a broader range of language capabilities. While this evolution has led to substantial improvements in fluency and usability, it has also introduced new imbalances, as current systems remain strongly influenced by the distribution of training data and the dominance of a limited set of high–resource languages (Brown et al., 2020; Caswell et al., 2020).

Languages such as Croatian provide a useful perspective on this transition. As a mid–resourced language, Croatian is typically supported through multilingual modelling rather than through large–scale, language–specific infrastructures. This means that its performance in MT systems depends not only on the intrinsic characteristics of the language, but also on how effectively multilingual architectures distribute capacity across languages and exploit cross–lingual transfer (e.g., Johnson et al., 2017; Aharoni et al., 2019; Conneau et al., 2020). As a result, Croatian serves as a revealing test case for examining the strengths and limitations of current MT paradigms, particularly with regard to capacity sharing and uneven performance across languages (Arivazhagan et al., 2019; Costa–Jussà et al., 2022).

The development of MT has followed a sequence of technological shifts, from rule–based systems to statistical approaches and, more recently, to neural architectures. These neural systems first emerged as encoder/decoder frameworks using Recurrent Neural Networks (RNNs) before evolving into the Transformer architecture. Within this paradigm, translation has progressed from dual–component Transformer models to the use of Large Language Models (LLMs), which are typically based on decoder–only Transformer architectures. These advances have led to substantial gains in translation quality, particularly at the sentence level. Multilingual NMT systems further extended these gains by enabling many–to–many translation within a single model, often improving performance for less represented languages through shared representations (Costa–Jussà et al., 2022). However, such systems also exhibit important limitations, including variability across language pairs and difficulties in capturing discourse level phenomena beyond isolated segments.

LLMs introduce a different perspective on translation. Rather than being designed exclusively for MT, they are trained on large and heterogeneous text collections to model language more generally, and are later adapted to specific tasks through instruction tuning or fine–tuning. One consequence of this design is their ability to process longer spans of text, which can lead to improvements in aspects such as coherence, referential clarity, and stylistic consistency (e.g. Gilabert et al., 2025). These properties are particularly relevant for Croatian, where morphosyn-

tactic features such as agreement and case marking often depend on context extending beyond the sentence level. At the same time, the increased fluency of LLM outputs raises new challenges, as it can obscure translation errors and complicate both automatic and human evaluation.

Evaluating MT systems in this context requires a multidimensional approach. Translation quality involves not only surface similarity to a reference translation but also semantic adequacy, terminological precision, and contextual appropriateness. While manual evaluation remains essential for capturing these aspects, automatic metrics continue to play a central role due to their scalability. In addition, evaluation results are heavily influenced by the choice of benchmark, with datasets such as FLORES and NTREX differing in their treatment of multilingualism and contextual scope (Federmann et al., 2022).

This paper investigates how the shift from NMT to LLM–based MT affects Croatian. We compare a range of neural and language model based systems across translation directions involving Croatian, English, and Spanish, with the aim of assessing their relative performance, and identifying emerging trends. In particular, we examine whether recent open–source models can compete with established commercial systems, and to what extent LLM–based approaches provide tangible improvements over previous multilingual NMT architectures. More broadly, the study contributes to understanding how current MT technologies reshape the translation landscape for mid–resource languages, highlighting the importance of data distribution, model design, and evaluation methodology, with translation quality assessed through a combination of automated metrics.

2. Machine translation evolution from NMT to LLMs

The recent history of MT can be understood as a sequence of shifts in both modelling strategy and scale. Earlier paradigms relied first on manually encoded linguistic rules and bilingual lexicons, and later on statistical methods trained on parallel corpora. Phrase–based statistical MT (SMT) dominated the field for many years, but this paradigm was gradually displaced by neural approaches. Initial Neural MT (NMT) systems utilized encoder/decoder frameworks based on Recurrent Neural Networks (RNNs), which were later enhanced by the addition of attention mechanisms. The most decisive shift occurred with the invention of Transformer architecture, which replaced RNNs with self–attention layers while maintaining the dual encoder/decoder structure. This configuration enabled more effective modelling of long–range dependencies, and more scalable training (Bahdanau et al., 2015; Vaswani et al., 2017). This progression marked the consolidation of NMT as the dominant framework for research and deployment.

The success of NMT was not only architectural. It also reflected a broader move toward larger models, larger datasets, and increasingly multilingual training regimes. Instead of building separate systems for individual language pairs, multilin-

gual NMT made it possible to train a single model across dozens or even hundreds of languages, often improving performance for lower resourced directions through transfer learning and shared parameterization (Johnson et al., 2017; Aharoni et al., 2019). At the same time, this multilingual turn introduced its own limitations: improvements were not distributed evenly across languages, and performance often depended on how model capacity was allocated across stronger and weaker language pairs (Arivazhagan et al., 2019; Costa–Jussà et al., 2022).

This gradual change in the field has now given way to a more substantial reconfiguration driven by Large Language Models (LLMs). Unlike conventional NMT systems, which typically rely on the dual encoder/decoder Transformer architecture, modern LLMs are developed as decoder–only Transformers. These models are trained as general purpose language models rather than being designed exclusively for translation. Their development typically involves an initial pretraining stage on very large text collections, followed by instruction tuning and, in many cases, additional optimization procedures such as reinforcement learning from human feedback. Translation can then be further strengthened through task specific fine–tuning on multilingual parallel data. As a result, translation is no longer treated as an isolated task, but as one capability within a broader language modelling framework (Brown et al., 2020; Ouyang et al., 2022).

One of the most visible consequences of this shift is the enhanced capacity to leverage context, addressing a long–standing challenge in MT. Traditional NMT systems have generally been optimized at the segment or sentence level, which limited their ability to model discourse level phenomena, such as referential consistency, lexical cohesion, or document wide stylistic coherence. LLMs, by contrast, are designed to process much longer textual spans and are, therefore, better positioned to handle translation problems that extend beyond the sentence. This is particularly relevant for morphologically rich languages such as Croatian, where agreement, case marking, and reference often depend on wider textual context.

The changing balance between paradigms is also visible in recent evaluation campaigns. The results of recent Conference on Machine Translation (WMT) general translation shared tasks show a clear increase in the presence of LLM–based or LLM–inspired approaches (see Table 1), reflecting a broader shift in research priorities and system design. In this sense, the evolution from NMT to LLMs should not be interpreted simply as a replacement of one architecture by another, but as a change in the overall logic of MT development: from specialized translation engines toward more flexible, instruction–driven, multilingual systems whose performance depends not only on architecture, but also on model size, training mixture, prompting strategy, and evaluation setup. This growing complexity also makes system comparison more difficult, since differences in quality are increasingly shaped by benchmark selection, domain, and task formulation rather than by architecture alone.

Table 1. Percentage of MT methodologies used at WMT shared tasks

Year	NMT	LLM	Hybrid	Other
2022	100%	0%	0%	0%
2023	88.8%	5.6%	0%	5.6%
2024	20%	57.21%	11.4%	11.4%
2025	11.1%	83.3%	5.6%	0%

For Croatian, this transition is particularly consequential. Earlier multilingual NMT systems already improved access to MT for languages with fewer dedicated resources, but LLMs may offer additional gains through their stronger contextual modelling and broader multilingual competence. At the same time, their performance for Croatian cannot be assumed: it remains tied to the quantity and quality of Croatian data within multilingual training pipelines and to the extent to which current benchmarks are able to capture genuine improvements in translation quality. For this reason, the transition from NMT to LLMs is not only a technological story, but also an empirical question that requires careful evaluation.

3. Parallel corpora for Croatian

The primary resources for training NMT systems and fine–tuning LLMs for translation are parallel corpora. A parallel corpus consists of a collection of texts in a source language paired with their translations in a target language. Typically, these texts are segmented into sentences, with each source sentence aligned to its corresponding translation. The predominant source for such data is the Opus Corpus[1] project (Tiedemann, 2009), a comprehensive repository of open parallel corpora that facilitates searches for specific language pairs.

Table 2 details the volume in segments of available corpora for the English–Croatian and Spanish–Croatian pairs. In the context of corpus linguistics and machine translation, a segment typically refers to a sentence, but it can also encompass headings, phrases, or lists elements that function as independent textual units. The corpora have been deduplicated and the given numbers are unique segments.

Table 2. Size in unique segments for the larger corpora available in Opus Corpus for English–Croatian and Spanish–Croatian

Corpus	English–Croatian	Spanish–Croatian
Open Subtitles 2024	46,010,567	37,923,208
HPLT v.3.	26,377,859	–
NLLB	18,797,414	–
Paracrawl v8	11,063,213	–
MultiCCAligned v11	–	2,774,615
Multiparacrawl v9b	–	1,817,219

1 https://opus.nlpl.eu

As evidenced by the data, there is a significant disparity in the volume of resources available for the two language pairs. While the English–Croatian pair benefits from several large–scale datasets exceeding ten million segments, the Spanish–Croatian pair relies more heavily on the Open Subtitles corpus, with other specialized resources being notably scarcer or absent. This imbalance underscores the challenges of developing direct translation models for language pairs involving Croatian that do not include English as a pivot. Furthermore, the massive scale of Open Subtitles highlights the importance of domain specific filtering, as the conversational nature of subtitles may not always align with the formal requirements of general purpose LLM fine–tuning. A significant driver for the availability of resources is the European Union's multilingual policy. As the EU generates a vast volume of official documentation—including legislative acts, parliamentary proceedings, and administrative reports—all of which must be translated into its official languages, there is a steadily increasing corpus of parallel texts involving Croatian. This institutional output ensures a growing supply of high–quality, professionally aligned data across all official EU language pairs.

4. Machine translation evaluation

Evaluating machine translation systems has become increasingly complex as the field has evolved from statistical models to neural architectures and, more recently, to LLMs. Translation quality cannot be adequately captured by a single metric, as it involves multiple dimensions, including semantic adequacy, fluency, terminological accuracy, and contextual coherence. This multidimensional nature of translation has led to the coexistence of different evaluation approaches, each capturing only part of the overall quality.

Manual evaluation remains the most reliable way to assess translation quality, as it allows for the identification of subtle errors related to meaning, terminology, discourse, and pragmatic appropriateness. Frameworks such as MQM (Multidimensional Quality Metrics) provide structured taxonomies for classifying translation errors and assessing their severity (Lommel et al., 2014). However, manual evaluation is time consuming, costly, and difficult to scale, especially when comparing multiple systems, languages, and configurations. As a result, automatic evaluation metrics continue to play a central role in MT research and benchmarking.

Among automatic metrics, BLEU has historically been the most widely used, measuring n–gram overlap between system output and reference translations (Papineni et al., 2002). Despite its limitations, including its sensitivity to surface variation and its limited ability to capture meaning, BLEU remains relevant due to its simplicity and comparability across studies. SacreBLEU was later introduced to standardize BLEU computation and improve reproducibility by controlling for tokenization and evaluation settings (Post, 2018). In addition to BLEU, character based metrics, such as chrF2, provide a complementary perspective by operating at

the character n–gram level rather than the word level (Popović, 2015). This makes chrF2 particularly well suited to languages with rich morphology, such as Croatian, where inflectional variation may lead to different surface forms that still preserve the same underlying meaning. By focusing on sub–word units, chrF2 is more robust to minor lexical or morphological differences and can better capture partial matches between hypothesis and reference translations.[2]

Edit–based metrics such as TER (Translation Edit Rate) offer a different perspective by measuring the number of edits required to transform the system output into the reference translation (Snover et al., 2006). These edits include insertions, deletions, substitutions, and shifts, and the resulting score can be interpreted as an approximation of post–editing effort. Unlike BLEU and chrF2, which are similarity based, TER directly models the notion of correction cost, making it particularly relevant in workflows where MT output is subsequently post–edited. However, TER also depends on the reference translation and may penalize valid alternative renderings that differ structurally from the reference.

More recently, neural evaluation metrics have gained prominence. COMET (Rei et al., 2020) represents a new generation of metrics that leverage pretrained language models to estimate translation quality based on semantic similarity and contextual information. Unlike traditional metrics, COMET integrates information from the source text, the hypothesis, and the reference, and has been shown to correlate more strongly with human judgments. This makes it particularly suitable for evaluating modern systems, including LLM–based approaches, whose outputs may diverge lexically from references while preserving meaning.

Evaluation outcomes are also strongly influenced by the choice of benchmark. Standardized evaluation datasets provide a controlled basis for comparing systems, but they also define the scope of what is being measured. FLORES, developed within the No Language Left Behind (NLLB) project, has become one of the most widely used multilingual benchmarks, offering high–quality human translations across a large number of languages (Costa–Jussà et al., 2022). Its main advantage lies in enabling consistent comparison across many language pairs, including those with limited resources. However, FLORES is primarily sentence–based, which limits its ability to capture document–level phenomena, such as coherence, consistency, and discourse structure.

To address some of these limitations, alternative benchmarks have been proposed. NTREX (Federmann et al., 2022) provides high–quality human reference translations for a large number of languages and is particularly suitable for evaluating multilingual systems in a more context aware setting. Unlike strictly sentence–level benchmarks, NTREX allows for a more realistic assessment of translation performance, especially when comparing systems that can leverage broader context,

2 In machine translation evaluation, the hypothesis refers to the automatic translation generated by the system, while the reference is the independent, human–vetted translation used as the gold standard for comparison.

such as LLMs. For this reason, NTREX is used in this study as the primary evaluation dataset.

Within this framework, our evaluation combines multiple automatic metrics—BLEU, chrF2, TER, and COMET—in order to capture complementary aspects of translation quality. This combination reflects the current consensus in MT evaluation, according to which no single metric is sufficient on its own. At the same time, it is important to acknowledge that even this combination remains an approximation of translation quality, as it cannot fully capture higher–level phenomena, such as discourse coherence, stylistic adequacy, or cultural appropriateness.

Overall, the evaluation of MT systems must be understood as a methodological design choice rather than a purely technical procedure. The interaction between metrics, benchmarks, and system architectures plays a decisive role in shaping the results. This is particularly relevant in the comparison between NMT and LLM–based systems, where differences in context handling, output variability, and fluency may not be equally reflected across all metrics. Consequently, the results presented in this paper should be interpreted in light of these methodological constraints, as part of a broader effort to understand how current evaluation practices interact with emerging MT paradigms. This study explores how the transition from task–specific NMT systems to general–purpose LLMs impacts translation performance for mid–resource languages within current automated frameworks. It specifically investigates whether open–source models can deliver quality comparable to that of a leading commercial system.

5. Experimental part

5.1. Analysed models

We have evaluated a selection of NMT systems and LLMs, both general and specifically fine–tuned for translation. We have evaluated the models for the following language pairs and directions: English–Croatian, Croatian–English, Spanish–Croatian and Croatian–Spanish.

- Google Translate[3]: a very popular machine translation system that can be accessed both through a webpage or through an API. The translations were performed using the API.[4]
- NLLB (Costa–Jussà et al., 2022). This is a multilingual NMT model able to translate between about 200 languages, including Croatian. We have evaluated several versions of the model with different model sizes: 600M–distilled, 1.3B–distilled. 1.3B and 3.3B. The M means millions of parameters and the B means billions (thousand millions) of parameters.

3 https://translate.google.com

4 Queries performed on March 13, 2026.

- SalamandraTA–instruct (Gilabert et al., 2025) is a translation–oriented LLM derived from an internal, unpublished base model (SalamandraTA–7b–base), which was continually pre–trained on parallel data. This instruction–tuned model is proficient in 35 European languages—including Croatian—and supports a wide range of tasks beyond sentence–level translation. Its capabilities extend to paragraph and document–level translation, automatic post–editing, grammar checking, and context aware translation, among other linguistic evaluation tasks. SalamandraTA–instruct is distributed in two sizes: 2B i 7B.
- Madlad400 (Kudugunta et al., 2023) is a T5–based multilingual model trained on 1 trillion tokens across more than 450 languages—including Croatian—using publicly available data. Despite its relatively compact parameter count, it remains competitive with significantly larger models, offering a robust open–source alternative for wide–scale multilingual translation. This model is distributed in three sizes: 3B, 7B and 10B.
- EuroLLM (Martins et al., 2025) is developed within the EuroLLM project. This project aims to develop a suite of models tailored for all European Union languages and other high–relevance global scripts. Available in two scales—9B and 22B parameters—the models were trained on a massive 4–trillion–token corpus encompassing web content, parallel data, and curated high–quality datasets. Its instruction–tuned counterpart, EuroLLM–Instruct, has been further optimized using the EuroBlocks dataset to enhance general instruction following and machine translation performance. The model provides robust support for 35 languages, including Croatian.
- Translategemma (Finkelstein et al., 2026) represents a specialized family of lightweight, state–of–the–art open translation models developed by Google, built upon the multimodal Gemma 3 architecture. Engineered to support translation tasks across 55 languages—including Croatian—these models are capable of both text–to–text and image–to–text processing. This multimodal functionality allows the models to perform sophisticated tasks such as Visual Question Answering (VQA) and the translation of embedded text within visual contexts. Their relatively compact footprint allows for seamless deployment in resource constrained environments, such as standard laptops or private cloud infrastructures, thereby lowering the hardware barriers to entry and fostering innovation in diverse linguistic applications. This model is available in three sizes: 4B, 12B and 27B.

General purpose LLMs, whether open–source or proprietary (such as Gemini or GPT–4), were not included in this evaluation because their performance is highly sensitive to prompt engineering, leading to significant variability in output quality based on the specific instructions used. In contrast, the translation oriented LLMs

selected for this study utilize a fixed, standardized prompt format explicitly defined in their technical documentation. This approach ensures a consistent evaluation framework and guarantees that the results are reproducible, as the models are operated according to their intended design for translation tasks.

5.2. Evaluation results

This section presents the evaluation results for the analysed models across four language pairs: English–Croatian, Croatian–English, Spanish–Croatian, and Croatian–Spanish. The NTREX corpus was employed as the evaluation set, with performance assessed through several automatic metrics. Specifically, SacreBLEU was used to calculate BLEU, chrF2, and TER, alongside COMET (utilizing the wmt22–comet–da model). While the results for each language pair are detailed in this subsection, a broader discussion and a comprehensive analysis of the findings are provided in the next subsection.

5.2.1. English–Croatian

As can be observed in Table 3, where the best values are marked in bold, the evaluation results show that while the commercial system Google Translate still leads with a peak BLEU score of 37.2, several open–source models are proving to be remarkably competitive. For instance, EuroLLM–22b matches Google's BLEU score exactly, demonstrating that open–source architectures can now achieve the same level of raw performance. Although Google Translate maintains a slight advantage in terms of fluency—indicated by its higher chrF2 (64.3) and COMET (0.9123) scores—the gap is narrowing significantly. It is also worth noting that while Google's TER (49.1) suggests it requires the least manual correction, the EuroLLM and Gemma suites are not far behind, offering high–quality alternatives to proprietary software.

Table 3. Evaluation results for the English–Croatian language pair

System	BLEU	chrF2	TER	COMET
Google Translate	**37.2**	**64.3**	**49.1**	**0.9123**
NLLB 600 dist.	27.4	56.7	58.2	0.8533
NLLB 1.3B dist.	31.3	59.6	54.4	0.8771
NLLB 1.3B	30.6	59.1	55.3	0.8749
NLLB 3.3B	32.3	60.2	53.9	0.8792
SalamandraTA–2b–instruct	30.2	59.1	56.8	0.8881
SalamandraTA–7b–instruct	32	60.6	54.7	0.9049

Madlad400–3b	31	59.8	55.2	0.8906
Madlad400–7b	31	59.8	55.3	0.8949
Madlad400–10b	31	59.8	54.8	0.8938
EuroLLM–9b	33.2	60.9	55.5	0.8925
EuroLLM–22b	**37.2**	63.5	51.4	0.905
TranslateGemma–4b	20.3	52.5	68.5	0.8779
TranslateGemma–12b	23.8	56.9	65.9	0.905
TranslateGemma–27b	27	59.1	63.9	0.9119

Furthermore, the results in Table 3 highlight how effective scaling has become for these open–source models. In the NLLB and SalamandraTA series, the larger versions consistently outperform the smaller ones, with the SalamandraTA–7b model reaching a very strong COMET score of 0.9049. Perhaps the most impressive growth is seen in the TranslateGemma series; despite the 4b and 12b versions starting with lower scores, the 27b variant achieves a COMET of 0.9119. This is nearly identical to Google Translate's performance, proving that free, high–capacity models are now reaching a professional standard. This trend suggests that for the English–Croatian pair, the reliance on commercial giants may soon become unnecessary as open–source models continue to improve.

5.2.2. Croatian–English

The evaluation results for translation from Croatian into English, presented in Table 4, reveal a significant shift in the performance hierarchy compared to the English–to–Croatian pair. While Google Translate remained the benchmark in the previous analysis, the COMET scores for this translation direction are much closer, and in several cases, open–source models actually outperform the commercial system. Specifically, TranslateGemma–27b (0.6264) and the 12b version (0.6242) both achieve the highest COMET scores, surpassing Google Translate (0.6189). Similarly, both EuroLLM–22b (0.6208) and EuroLLM–9b (0.6193) exceed the proprietary system's performance. This suggests that for translating into English, these open–source architectures benefit from extensive training on high resource target language data.

Table 4. Evaluation results for the Croatian–English language pair

System	**BLEU**	**chrF2**	**TER**	**COMET**
Google Translate	**45.1**	**68.9**	**42.4**	0.6189
NLLB 600 dist.	37.5	63.4	49.8	0.5995
NLLB 1.3B dist.	40.7	65.5	46.5	0.6098
NLLB 1.3B	40	65.3	47.1	0.6081

NLLB 3.3B	41.5	66.3	45.9	0.6105
SalamandraTA–2b–instruct	40.8	65.3	47.1	0.6115
SalamandraTA–7b–instruct	44.1	67.6	43.7	0.6148
Madlad400–3b	37.4	63.9	47.9	0.6138
Madlad400–7b	38.6	64.9	46.6	0.6149
Madlad400–10b	37.8	64.3	47.1	0.6142
EuroLLM–9b	36.8	64.4	51.5	0.6193
EuroLLM–22b	38.4	65.2	49.9	0.6208
TranslateGemma–4b	30.4	59.8	59.5	0.6174
TranslateGemma–12b	34.1	63.2	54.9	0.6242
TranslateGemma–27b	33.8	63.5	54.2	**0.6264**

However, it is noteworthy that this superiority is primarily reflected in the neural based COMET metric. When examining lexical based metrics like BLEU, Google Translate (45.1) still maintains a lead over the top–performing open–source model in that category, SalamandraTA–7b–instruct (44.1). This discrepancy highlights the importance of using multidimensional evaluation frameworks, as neural metrics often capture semantic accuracy and fluency that traditional n–gram overlaps may overlook.

This trend reinforces findings regarding the effectiveness of model scaling. In this direction, the gap between mid–range and high–capacity models is narrower, yet the refined performance of the Gemma and EuroLLM suites is evident. While the NLLB and Madlad400 series remain reliable, they fall slightly behind the latest instruction–tuned models in terms of semantic quality. The fact that accessible, open–source models are now delivering results that match or exceed those of a major commercial provider represents a major milestone. It confirms that for Croatian–to–English tasks, open–source solutions have transitioned from being competitive alternatives to becoming a leading choice for high–quality translation.

5.2.3. Spanish–Croatian

The evaluation results for the Spanish–Croatian language pair, as presented in Table 5, where the best values are marked in bold, reveal a particularly unexpected and significant trend. While the commercial system Google Translate maintains a leading position in surface–level fluency metrics like BLEU (27.9), it is remarkably challenged by open–source models in terms of semantic accuracy. Most notably, TranslateGemma–27b achieves a COMET score of 0.8982, actually surpassing Google Translate's 0.8949. As can be observed in Table 5, this is highly surprising

given that Spanish–to–Croatian is a lower resource language pair compared to English; typically, one would expect proprietary systems to hold a much larger advantage due to their massive data–harvesting capabilities. Instead, these results suggest that modern open–source architectures are becoming exceptionally efficient at handling complex, less common language combinations.

When comparing the data in Table 5 to the English–Croatian results, it is evident that overall scores are lower across all metrics, reflecting the inherent difficulty of translating between two non–English languages. However, the performance gap between commercial and open–source systems is much narrower than anticipated. For instance, EuroLLM–22b (26.4 BLEU) and TranslateGemma–12b (0.8913 COMET) perform remarkably close to the commercial benchmark.

Furthermore, the scaling efficiency noted in previous sections is confirmed once more in Table 5. The SalamandraTA–7b model shows strong performance (0.8884 COMET), proving to be a highly efficient choice for Spanish–sourced tasks despite its smaller parameter count. In conclusion, while commercial systems still hold a slight lead in metrics like BLEU and chrF2, the fact that open–source solutions such as Gemma and EuroLLM can match or even exceed them in a lower resource context is a major finding for the research community.

Table 5. Evaluation results for the Spanish–Croatian language pair

System	**BLEU**	**chrF2**	**TER**	**COMET**
Google Translate	**27.9**	**57.7**	**60.4**	0.8949
NLLB 600 dist.	19.9	50.7	68.9	0.8458
NLLB 1.3B dist.	23.1	53.2	65	0.8694
NLLB 1.3B	21.9	52.3	66.3	0.8673
NLLB 3.3B	23.2	53.6	65.3	0.8768
SalamandraTA–2b–instruct	23	53.2	65.9	0.8793
SalamandraTA–7b–instruct	23.3	53.5	64.8	0.8884
Madlad400–3b	23	53.2	65.9	0.8747
Madlad400–7b	22.4	52.4	66.9	0.8717
Madlad400–10b	23.1	52.9	65.7	0.8743
EuroLLM–9b	23.7	54.3	67.2	0.8763
EuroLLM–22b	26.4	56.1	63.2	0.8877
TranslateGemma–4b	16.1	48.7	75.7	0.8616
TranslateGemma–12b	19.5	52.8	71.7	0.8913
TranslateGemma–27b	22.3	54.8	69.7	**0.8982**

5.2.4. Croatian–Spanish

The evaluation results for the translation from Croatian into Spanish, presented in Table 6, confirm that this language pair is highly competitive. While Google Translate maintains the top position in terms of BLEU (34.1) and chrF2 (60.2), the gap with open–source models is remarkably narrow. Specifically, EuroLLM–22b follows closely with a BLEU score of 33.3, and TranslateGemma–27b achieves a COMET score of 0.8542, which is practically identical to Google's 0.8544. As can be observed in Table 6, the scores for this direction (hr → es) are notably higher than those for the inverse (es → hr), suggesting that these models are generally more proficient at generating Spanish text.

When comparing these findings with the previous tables, a significant trend emerges regarding language resources. Although the Spanish–Croatian pair suffers from a lack of parallel data—as previously discussed in the corpus section—modern open–source LLMs prove to be exceptionally capable of translating from a medium–resource language like Croatian into a high–resource language like Spanish. This suggests that the vast amount of monolingual data available in Spanish helps these models compensate for the scarcity of direct parallel bilingual examples.

Furthermore, the data in Table 6 reinforces the scaling efficiency noted throughout this study. While proprietary systems still hold a slight lead in raw fluency, the latest generation of open–source models, such as the Gemma and EuroLLM suites, demonstrate an impressive ability to capture linguistic nuances. In conclusion, the results from Table 6 show that despite the limited resources for this specific pair, open–source solutions have become highly effective at producing high–quality Spanish output, often reaching a professional standard that is indistinguishable from established commercial services.

Table 6. Evaluation results for the Croatian–Spanish language pair

System	BLEU	chrF2	TER	COMET
Google Translate	**34.1**	**60.2**	**54.7**	0.8544
NLLB 600 dist.	29	55.7	59	0.8217
NLLB 1.3B dist.	31.6	57.6	56.2	0.8393
NLLB 1.3B	31.3	57.4	56.5	0.8381
NLLB 3.3B	32.2	58	56.2	0.8429
SalamandraTA–2b–instruct	29.3	56	59.7	0.8452
SalamandraTA–7b–instruct	28.9	55.5	60.1	0.8477
Madlad400–3b	29.3	56.1	57.4	0.8392
Madlad400–7b	29.6	56.5	56.8	0.839

Madlad400–10b	29.7	56.4	56.5	0.8389
EuroLLM–9b	32.1	58.2	56.9	0.8482
EuroLLM–22b	33.3	59.2	55.3	0.8509
TranslateGemma–4b	27.1	55.3	63	0.8392
TranslateGemma–12b	30.2	58.2	59.2	0.8516
TranslateGemma–27b	30.7	58.6	59.3	0.8542

5.3. Discussion

The comprehensive evaluation across all four language pairs demonstrates a transformative shift in the MT landscape. The most salient finding is that open–source models now provide a level of quality that is fully comparable to well–established commercial systems, like Google Translate. While proprietary engines often maintain a slight lead in surface–level fluency (BLEU), the semantic accuracy (COMET) of high–capacity models such as TranslateGemma–27b and EuroLLM–22b frequently matches or even surpasses the commercial benchmark. This indicates that the technological gap that once protected proprietary systems is closing, particularly as open–source architectures become more efficient at capturing deep linguistic nuances.

Furthermore, these results highlight the remarkable versatility and availability of open–source alternatives. Users are no longer restricted to a single provider for high–quality translations involving Croatian. The models evaluated—such as NLLB, Madlad400, and EuroLLM—are inherently multilingual, supporting not only English and Spanish but, in many cases, hundreds of other languages. This multilingual nature is particularly beneficial for medium–resource languages like Croatian, as the models can leverage cross–lingual knowledge to improve translation performance even when direct parallel data is scarce.

Finally, the data proves that these models excel regardless of the translation direction. Whether translating from a high–resource language into Croatian or vice versa, the open–source ecosystem offers a diverse array of professional grade tools. The ability of these models to produce high–quality output in a variety of linguistic contexts—from the syntactically complex Slavic structures to the resource–rich Romance languages—confirms their readiness for integration into professional workflows. In summary, the reliance on commercial giants is no longer a necessity, as the open–source community now provides robust, transparent, and equally capable global translation solutions.

Despite the remarkable performance of open–source models—many of which have been developed through high–level European research initiatives—there remains a significant gap between their potential and their actual adoption in professional and industrial workflows. Currently, proprietary systems from large tech-

nology corporations continue to dominate the market, largely due to their ease of access and established ecosystems. However, as the results in this study demonstrate, this trend must be reversed to prevent high–quality, publicly funded European research from remaining 'on the shelf' without being put to practical use.

A key factor in bridging this gap is the emergence of projects like MTUOC[5] (Machine Translation from the Universitat Oberta de Catalunya) (Oliver, 2025). These initiatives are essential because they provide the necessary infrastructure and tools to facilitate the deployment and everyday use of open–source models, making them accessible to translators and researchers, since one of the primary concerns often cited regarding open–source Large Language Models (LLMs) is the high hardware requirement for execution. While it is true that the largest variants demand significant computational power, this study highlights two important counterpoints:

- Efficiency of Small–to–Mid–Sized Models: Models with fewer parameters (such as the 7b or 9b versions) offer a notable balance between performance and resource consumption, making them perfectly viable for standard professional hardware.
- Quantization: The use of quantized versions of larger models (like the 22b or 27b variants) allows for high–quality inference on mid–range computers without a critical loss in accuracy.

In conclusion, the transition toward open–source translation solutions is not only a matter of performance—which is already proven—but a strategic necessity. Supporting local projects that facilitate their use will ensure that Europe remains at the forefront of the AI revolution, utilizing its own resources to maintain linguistic diversity and digital independence.

6. Conclusions and future work

This study has examined the current state of the Croatian language during the paradigm shift from specialized NMT to LLMs. By evaluating a diverse range of architectures—including commercial benchmarks and open–source suites like EuroLLM, TranslateGemma, and SalamandraTA—we have demonstrated that open–source solutions are no longer just alternatives, but are now capable of matching or even exceeding the quality of proprietary systems. This is particularly evident in the Croatian–to–Spanish and Croatian–to–English directions, where high–capacity models leverage their extensive multilingual training to produce professional grade results despite limited parallel resources.

Our results confirm that translation quality for mid–resource languages, such as Croatian, depends on how the model architecture manages its capacity and lev-

5 https://mtuoc.github.io/

erages cross–lingual knowledge (Aharoni et al., 2019; Conneau et al., 2020). We found that modern open–source models no longer need to be designed exclusively for translation; their nature as general purpose language models allows them to act as highly effective learners through instructions and examples (Brown et al., 2020; Ouyang et al., 2022). In this regard, the strong performance of models like SalamandraTA demonstrates that well–balanced multilingual training is key to overcoming data limitations and delivering professional grade results (Gilabert et al., 2025).

It is important to acknowledge, however, that this article is inherently incomplete. The rapid pace of innovation in the field means that new models are released almost weekly, and several emerging architectures remain to be tested. This evaluation represents a snapshot of a highly dynamic landscape where the boundaries of what open–source models can achieve are constantly being redefined.

Looking forward, the primary challenge lies in the transition of these open–source models into productive environments. While their performance is proven, the industry still relies heavily on commercial giants due to ease of integration. Bridging this gap requires a concerted effort to support infrastructure projects, such as MTUOC, which facilitate the deployment of these models in real–world professional workflows.

Future research should focus on two main areas:

- Specialized Fine–tuning: Moving beyond general purpose translation toward the adaptation of these models for specific domains (legal, medical, or technical), where terminological precision and stylistic consistency are paramount.
- Efficiency and Hardware Accessibility: Further exploring optimization techniques such as quantization to ensure that high–performance translation is accessible on standard professional hardware, thereby fostering technological autonomy for local industries.

In conclusion, while the technological gap between commercial and open–source translation is closing, the future of the Croatian MT landscape will depend on our ability to implement, specialize, and maintain these free resources within a sustainable and independent European ecosystem.

Acknowledgements

This research was conducted within the framework of the coordinated project AI–Driven Translation for Low–Resource Languages and Cultures (AI–TraLow). Specifically, this work is part of the sub–project “Large Language Models for Translating Low–Resource Romance Languages of the Iberian Peninsula” (Project PID2024–158157OB–C33), funded by MICIU/AEI/10.13039/501100011033/FEDER, UE.

References

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3874–3884). Association for Computational Linguistics.

Arivazhagan, N., Bapna, A., Firat, O., Cherry, C., Macherey, W., Foster, G., Krikun, M., Johnson, M., & Chen, Z. (2019). *Massively multilingual neural machine translation in the wild: Findings and challenges*. arXiv. https://doi.org/10.48550/arXiv.1907.05019

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few–shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Caswell, I., Breiner, T., Van Esch, D., & Bapna, A. (2020). Language ID in the wild: Unexpected challenges on the path to a thousand–language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6588–6608). International Committee on Computational Linguistics.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics.

Costa–Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia Gonzalez, G., Hansanti, P., Hoffman, J., Holtzman, S., Ituarte–Vazquez, I., Kuo, J., Miller, D., Ngan, H., Pira, S., Schwenk, H., Shao, J., Tang, K. K., Zaccagna, S., & Wang, C. (2022). *No Language Left Behind: Scaling human–centered machine translation*. arXiv. https://doi.org/10.48550/arXiv.2207.04672

Federmann, C., Kocmi, T., & Xin, Y. (2022). NTREX–128 – News test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation* (pp. 21–24). Association for Computational Linguistics.

Finkelstein, M., Caswell, I., Domhan, T., Peter, J., Juraska, J., Riley, P., Deutsch, D., Kovacs, G., Dilanni, C., Cherry, C., Briakou, E., Nielsen, E., Luo, J., Black, K., Mullins, R., Agrawal, S., Xu, W., Kats, E., Jaskiewicz, S., Freitag, M., Vilar, D. (2026). *TranslateGemma technical report*. arXiv. https://doi.org/10.48550/arXiv.2601.09012

Gilabert, J. G., Liao, X., Da Dalt, S., Bohman, E., Mash, A., Fornaciari, F. D. L., ... & Melero, M. (2025). From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 614–637). Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural

machine translation system: Enabling zero–shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065

Kudugunta, S., Caswell, I., Zhang, B., Garcia X., Choquette–Choo, Christopher A., Lee, K., Xin, D., Kusupati, A., Stella, R., Bapna, A., & Firat, O. (2023). MADLAD–400: A multilingual and document–level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 67284–67296). Curran Associates Inc.

Lommel, A. R., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, *12*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J. G. C., Birch, A., Martins, A. F. T. (2025). EuroLLM: Multilingual language models for Europe. *Procedia Computer Science*, *255*, 53–62. https://doi.org/10.1016/j.procs.2025.02.260

Oliver, A. (2025). MTUOC server: Integrating several NMT and LLMs into professional translation workflows. In *Proceedings of Machine Translation Summit XX: Volume 2* (pp. 81–82). European Association for Machine Translation.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., 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. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.–J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.

Popović, M. (2015). chrF: Character n–gram F–score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation* (pp. 186–191). Association for Computational Linguistics.

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)* (pp. 2685–2702). Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A study of translation edit rate with targeted human annotation*. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*.

Tiedemann, J. (2009). News from OPUS — A collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2009)* (Vol. 5, pp. 237–248).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Hrvatski jezik u prijelazu s neuronskog strojnog prevođenja na velike jezične modele

Tehnologije strojnoga prevođenja (MT) trenutno prolaze kroz promjenu paradigme, prelazeći sa specijaliziranih sustava neuronskoga strojnog prevođenja (NMT) na šire mogućnosti velikih jezičnih modela (LLM). Ovaj rad ispituje trenutni položaj hrvatskoga jezika unutar te tehnološke evolucije.

Dok dvojezični NMT modeli često pokazuju visoku preciznost na segmentnoj razini, višejezični NMT koristio se učenjem prijenosom (engl. *transfer learning*) kako bi poboljšao izvedbu za jezične parove s manje resursa, iako je ta tehnika postizala slabije rezultate za jezike bogate resursima. Za razliku od toga, LLM–ovi, bilo oni opće namjene ili fino podešeni (engl. *fine–tuned*) za prevođenje, nude vrhunsku višejezičnu točnost uzimajući u obzir kontekst. Za razliku od NMT–a, LLM–ovi mogu obrađivati dijelove veće od rečenice, poput cijelih odlomaka ili dokumenata, što dovodi do značajnih poboljšanja u razrješavanju koreferencije i slaganja u rodu. Unatoč znatnim računalnim zahtjevima LLM–a nedavno uvedene tehnike optimizacije omogućuju manje, učinkovitije inačice koje zadržavaju visoku kvalitetu rezultata.

U ovome radu procjenjuje se kvaliteta prijevoda različitih NMT i LLM arhitektura s hrvatskoga jezika na engleski i španjolski i obratno uporabom nekoliko metrika za automatsku procjenu kvalitete. Rezultati pokazuju da modeli otvorenoga koda mogu postići, a povremeno i nadmašiti, kvalitetu Google Prevoditelja, široko korištenoga komercijalnog NMT sustava. Osim toga, iako se ovo istraživanje usredotočuje na tri navedena jezika, višejezična priroda analiziranih sustava sugerira da modeli otvorenoga koda pružaju visokokvalitetne mogućnosti strojnoga prevođenja za hrvatski jezik za desetke, pa čak i stotine jezičnih parova.

Ključne riječi: neuronsko strojno prevođenje; veliki jezični modeli; hrvatski; engleski; španjolski