★Open Machine Translation for Esperanto

Ona de Gibert[♡], **Lluís de Gibert**[★]

[♡]University of Helsinki, Department of Digital Humanities
[★]Kataluna Esperanto-Asocio (KEA)
[★]Sennacieca Asocio Tutmonda (SAT)
ona.degibert@helsinki.fi

Abstract

Esperanto is a widespread constructed language, known for its regular grammar and productive word formation. Besides having substantial resources available thanks to its online community, it remains relatively underexplored in the context of modern machine translation (MT) approaches. In this work, we present the first comprehensive evaluation of open-source MT systems for Esperanto, comparing rule-based systems, encoder–decoder models, and LLMs across model sizes. We evaluate translation quality across six language directions involving English, Spanish, Catalan, and Esperanto using multiple automatic metrics as well as human evaluation. Our results show that the NLLB family achieves the best performance in all language pairs, followed closely by our trained compact models and a fine-tuned general-purpose LLM. Human evaluation confirms this trend, with NLLB translations preferred in approximately half of the comparisons, although noticeable errors remain. In line with Esperanto's tradition of openness and international collaboration, we release our code and best-performing models publicly.

Keywords: Esperanto, Machine Translation, Low-resource

1. Introduction

Constructed languages (*conlangs*) are languages intentionally created for human communication rather than emerging through natural linguistic evolution (Kuhn, 2014). From the perspective of language technology, conlangs occupy an unusual position: they typically attract limited commercial investment, which reduces incentives to develop dedicated tools and resources (Occhini et al., 2026). At the same time, many conlangs are supported by active online communities and maintain a considerable web presence, which leads to their inclusion in large-scale training corpora. Among conlangs, Esperanto is the most prominent and widely used example (Blanke, 2009).

Esperanto represents a unique case among constructed languages. It has a well-developed Wikipedia with over 380,000 articles and a large global community of second-language speakers. In addition, Esperanto ranks 75th in language presence in Common Crawl as of the most recent crawl (CC-MAIN-2026-04).[1] This indicates substantial web representation relative to many other low-resource natural languages. Esperanto is also included in major large-scale pretraining corpora such as MADLAD-400 (Kudugunta et al., 2023) and HPLT (de Gibert et al., 2024; Burchell et al., 2025; Oepen et al., 2025). Consequently, modern Large Language Models (LLMs) are exposed to non-trivial amounts of Esperanto during pre-training and are capable of generating it. However, the absence of dedicated evaluation benchmarks makes it difficult to systematically assess their true proficiency. While Esperanto has non-trivial available resources, we frame Esperanto as under-resourced in terms of its underrepresentation in language technology and limited targeted development of NLP resources and applications.

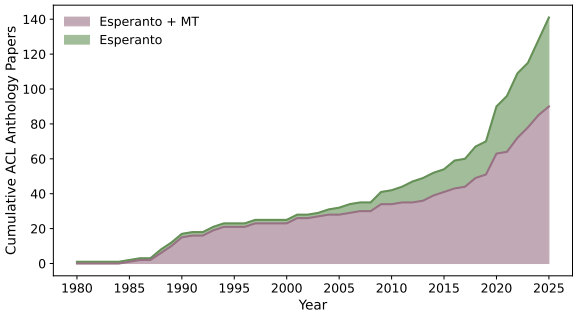


Figure 1: Cumulative number of ACL Anthology papers mentioning 'Esperanto' and 'Esperanto and Machine Translation'. Early work involving Esperanto focused primarily on MT, while more recent work covers a broader range of topics.

Machine translation (MT) provides a practical and controlled evaluation framework. Despite the existence of substantial textual resources, there is currently no state-of-the-art, openly available MT system specifically optimized for Esperanto. While we are aware Esperanto is included in commercial platforms, we disregard them in this work. We focus on evaluating and developing open MT

[1]Data Source: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 312–329
11–12 May 2026.

systems for translation from Esperanto (eo) into Catalan (ca), Spanish (es) and English (en); and vice-versa.

We present the first systematic benchmark of open-source MT systems for Esperanto translation, comparing rule-based systems, encoder–decoder models, and LLMs across model sizes ranging from 600M to 9B parameters. We find that LLMs lag behind specialized MT systems, while the NLLB family achieves the best overall performance. We further train compact Transformer models that remain competitive with substantially larger systems while being more efficient and sustainable. Finally, we conduct a human evaluation and qualitative error analysis to better understand the strengths and weaknesses of the evaluated models.

In our work, we adopt an open-science approach, as we aim to support and develop decentralized, grassroots language technology for Esperanto, reflecting its core values of linguistic equality, international communication, and community collaboration. To support reproducibility, we release our code in a public GitHub repository[2] and our best models on HuggingFace[3].

2. An Introduction to Esperanto

Esperanto is an *a posteriori* conlang (Couturat, 1903), modeled on existing languages. It was proposed by Zamenhof in 1887 with the aim of enabling universal communication in a neutral context. Esperanto represents the most consolidated case of a planned language, designed to be as accessible to as many speakers as possible. It is spoken in more than 100 countries (Poncelas et al., 2020). Estimates of the number of second language speakers vary considerably —from tens of thousands (Eberhard et al., 2026) to several million (Wandel, 2015)— , reflecting the methodological difficulty of obtaining precise demographic data in transnational communities. It is also the only conlang that has developed a stable community of first language speakers, usually estimated at between 1,000 and 2,000 individuals (Eberhard et al., 2026). Despite not having official state status, Esperanto has been the subject of institutional recognition, including UNESCO resolutions since 1954 that highlight its potential as a tool for international understanding (UNESCO, 1954). At the same time, its presence in digital environments and both original and translated literary production contribute to shaping an active communicative ecosystem, with intergenerational continuity and capacity for technological adaptation.

Esperanto uses the Latin alphabet with diacritics. It was designed according to principles of structural regularity and morphological transparency, which make it highly unambiguous. From a typological perspective, it presents a highly systematic agglutinative morphology, based on the productive combination of invariable roots with a restricted and regular set of prefixes and suffixes. This mechanism reduces morphosyntactic irregularity and favors a high degree of lexical compositionality. In terms of vocabulary, the language derives from Indo-European languages, with an approximate proportion of 80% of Romance origin, complemented by Germanic, Slavic and Greek contributions (Parkvall, 2010). It has often been described as structurally transparent and morphologically predictable language, suitable for MT (Schubert, 2002; Gobbo, 2015).

3. Related Work

We review prior research specifically focused on Esperanto. First, to better understand the evolution of the field, we query the full ACL Anthology for the terms “Esperanto”, and “Esperanto" and "Machine Translation” using the `acl-crawl` tookit.[4] Figure 1 shows the cumulative number of papers matching our queries. Esperanto has been present in the literature since the 1980s. In the early years, most papers mentioning Esperanto focused on MT, but more recently, the topics have diverged. This trend is expected, as research on MT has declined in general and the field has increasingly shifted toward natural language understanding (NLU) tasks.

3.1. Esperanto in NLP

Early work on Esperanto in NLP concentrated on rule-based linguistic modeling. Karlsson (1990) developed a Constraint Grammar framework for Esperanto, Minnaja and Paccagnella (2000) developed a Part-of-Speech tagger, and Manaris et al. (2006) performed a corpus-based linguistic analysis, comparing Esperanto to other European languages. Bick (2016) introduced a morphological lexicon for Esperanto and later a syntactic treebank (Bick, 2020). Beyond core NLP tasks, Fiedler (2018) investigated code-switching phenomena between Esperanto and English. More recently, Oya (2025) introduced Universal Dependencies annotations for Esperanto, while Bick (2025) released a learner corpus with error annotations and Constraint Grammar tags. This shows that the development of core NLP resources for Esperanto is still an active area of research.

[2] https://github.com/onadegibert/EsperantoMT

[3] https://huggingface.co/collections/Helsinki-NLP/open-machine-translation-for-esperanto

[4] https://github.com/Sethjsa/acl-crawl

3.2. Esperanto in MT

Esperanto attracted considerable attention in the early development of MT, particularly during the rule-based and statistical MT eras. Due to its regular grammar and its lexicon derived from multiple European languages, several studies proposed Esperanto as a potential interlingua for multilingual MT systems (Witkam, 1984; Neijt, 1986; Franco Sabarís et al., 2001; Boddington, 2004). However, these works were largely conceptual and exploratory rather than large-scale implemented systems.

Regarding Rule-Based MT (Hutchins and Somers, 1992), Apertium (Forcada et al., 2011), a free and open-source toolkit for developing rule-based MT systems, has included Esperanto translation pairs since its early releases. Additionally, Bick (2011) developed a rule-based system that translated portions of the English Wikipedia into Esperanto. During the statistical MT period, Esperus (Orlova, 2015) was developed as a Russian–Esperanto system using the MOSES toolkit (Koehn et al., 2007). Other systems from that period include Esperantilo (2008), Lingvohelpilo (2009), and Lingvoilo (2015) (Burghelea, 2019). These systems demonstrated practical interest in Esperanto MT but remained relatively small-scale. In recent years, dedicated research on Esperanto in MT has significantly diminished. One notable exception is Poncelas et al. (2020), who explored tokenization strategies for literary translation between Esperanto and English. More recently, Esperanto is included in the No Language Left Behind (NLLB) initiative (Costa-Jussà et al., 2022; NLLB Team, 2024) a highly multilingual model family, capable of translating among more than 200 languages. As a result, Esperanto is also included in the FLORES+ benchmark (Goyal et al., 2022), which provides a standardized evaluation test set for translation between Esperanto and 200 other languages. This represents an important step forward for Esperanto in multilingual MT evaluation. Beyond this, however, the past decade has seen limited focused research on Esperanto translation within modern neural MT paradigms.

4. Experimental Setup

We study MT models for Esperanto to translate both from and into English, Catalan, and Spanish. We are interested in these languages because Esperanto has a strong presence in the Iberian Peninsula, as demonstrated by the development efforts on open-source rule-based MT systems for these language pairs (Forcada et al., 2011) and active groups scattered around Catalonia, the Basque Country, Valencia and Andalusia.

Our experimental framework is divided into two parts. We first introduce our benchmarking setup, where we evaluate existing systems out-of-the-box. We then describe our MT development efforts, where we train encoder-decoder models and fine-tune an LLM. Finally, we present the metrics employed.

4.1. Benchmarking

To assess the current state of Esperanto MT, we evaluate several publicly available systems without any additional fine-tuning. This allows us to establish a realistic performance baseline and measure how well Esperanto is currently supported in multilingual MT systems.

Models We evaluate models representing different architectures and modeling paradigms:

- **Apertium** (Forcada et al., 2011): A rule-based MT system relying on manually crafted linguistic rules and bilingual dictionaries. Apertium is computationally efficient and lightweight. However, it only supports four of the six translation directions considered in this study.
- **NLLB family** (Costa-Jussà et al., 2022; NLLB Team, 2024): We evaluate four models from the NLLB family. These are highly multilingual encoder–decoder Transformer models trained on more than 200 languages. They come in different sizes (from 600M to 54B parameters). We evaluate models up to 3.3B parameters due to computational budget constraints. We also include two variants distilled from their biggest MoE 54B model via Word-level Knowledge Distillation (Kim and Rush, 2016).
- **Llama** (Grattafiori et al., 2024): We evaluate the instruction-tuned variant Llama-3.1-8B-Instruct via zero-shot prompting. These general-purpose decoder-only LLM is not specifically optimized for translation but has shown competitive performance in zero-shot and few-shot settings (Kocmi et al., 2025).
- **Tower family**: We evaluate two models from the Tower family (Alves et al., 2024). The Unbabel/TowerInstruct-7B-v0.2 model is based on Llama 2 (Touvron et al., 2023) and further trained with continued pre-training and supervised fine-tuning for 10 high-resource languages and translation-specific tasks. We also evaluate Tower-Plus (Rei et al., 2025), the model Unbabel/Tower-Plus-9B, a newer variant trained on Gemma 2 (Team et al., 2024) optimized for a broader set of multilingual and instruction-following tasks, covering 27 languages.

	Training Data			Used for Training		FLORES+	
Language Pair	Raw	Filtered	% Removed	Marian	LLM FT	Dev	Test
en–eo	70.34M	45.39M	35.5%	5.0M	100k	997	1012
es–eo	6.21M	4.68M	24.7%	4.7M	100k	997	1012
ca–eo	1.17M	673.93k	42.7%	673k	100k	997	1012
Total	77.73M	50.74M		10.35M	300k	2991	

Table 1: Parallel data statistics before and after filtering, and final training sizes per model family.

All the LLMs are of similar size, between 7B and 9B parameters.

Evaluation Data For evaluation, we use the Flores+ benchmark (Goyal et al., 2022). Flores+ provides professionally translated, multi-domain test sets across a large number of languages, including Esperanto. We evaluate all supported language pairs in both translation directions. We discuss the limitations of using Flores+ in our Limitations Section.

4.2. MT Development

In addition to benchmarking existing systems, we train our own open MT models for Esperanto.

Training We train separate bilingual models for each translation direction (into and from Esperanto), following common practices in low-resource MT (Haddow et al., 2022). We adopt two complementary strategies.

First, we train standard encoder–decoder Transformer models from scratch using Marian (Junczys-Dowmunt et al., 2018), without relying on pre-trained models. We experiment with two configurations: Transformer-base (60.6M parameters, Vaswani et al. (2017)) and Transformer-tiny (17.4M parameters, Bogoychev et al. (2020)), allowing us to analyze performance–efficiency trade-offs in resource-constrained environments.

Second, we perform supervised fine-tuning of Llama-3.1-8B-Instruct. We apply parameter-efficient fine-tuning using LoRA (Hu et al., 2022), with rank $r = 16$ and scaling factor $\alpha = 32$. These hyperparameters were adopted directly from O'Brien et al. (2025). Fine-tuning is conducted using the `open-instruct` toolkit.[5] We adopt an instruction-style prompting format that includes Flores-like language tags (see Figure 2). The model is trained to generate the target translation directly after the instruction prompt.

Details of training hyperparameters and Transformer architectures can be found in Appendix A.

We also experimented with fine-tuning NLLB models. However, under our multilingual setup and available data scale, we did not observe improvements over the base models. More details about our failed setup can be found in Appendix B.

> Translate the text from English (eng_Latn) into Esperanto (epo_Latn):
>
> Majorcan cuisine, like that of similar zones in the Mediterranean, is based on bread, vegetables and meat (specially pork), and uses olive oil throughout.
>
> Majorka kuirarto, samkiel tiuj de similaj mediteraneaj regionoj, havas kiel bazon panon, legomojn kaj viandon (precipe el porko) kaj ĝi uzadas olivoleon ĉie.

Figure 2: Prompt example for fine-tuning Llama-3.1-8B-Instruct

Data We use the Tatoeba Challenge data (Tiedemann, 2020), a deduplicated aggregation of parallel corpora from the OPUS repository (Tiedemann et al., 2024). Table 1 shows an overview of the data used.

Data cleaning is performed using OpusFilter (Aulamo et al., 2020). Our filtering pipeline includes:

- Length filtering: minimum 3 tokens, maximum 100 tokens.
- Length ratio filtering: maximum ratio of 2 between source and target.
- Removal of sentences containing words longer than 40 characters.
- Removal of HTML tags.
- Language identification filtering with langid.py (Lui and Baldwin, 2012).
- Restriction to Latin alphabet characters.

For training the Transformer models, we subsample the English data to 5M to have a more balanced training set. For LLM fine-tuning, we subsample

[5] https://github.com/allenai/open-instruct

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Rule-based MT						
Apertium	47.03	-	-	50.82	42.68	45.44
Neural MT						
NLLB-200-distilled-600M	65.08	48.06	52.86	58.61	48.32	52.26
NLLB-200-1.3B	66.35	49.36	55.15	59.51	49.22	51.97
NLLB-200-distilled-1.3B	66.81	49.64	55.60	60.04	49.34	51.52
NLLB-200-3.3B	67.27	49.68	56.19	59.91	49.57	52.70
General-purpose LLMs						
Llama-3.1-8B-Instruct	62.94	45.87	48.86	55.39	44.78	49.59
MT-tuned LLMs						
TowerInstruct-7B-v0.2	51.79	38.86	8.22	28.27	25.65	23.97
Tower-Plus-9B	64.63	47.66	49.02	47.02	40.38	42.99
Neural MT from Scratch						
Transformer-base (60.6M)	61.33	46.73	53.27	55.11	45.97	48.73
Transformer-tiny (17.4M)	57.69	45.16	49.02	54.05	45.33	49.57
Fine-tuned General-purpose LLMs						
Llama-3.1-8B-Instruct-FT	61.14	45.56	49.47	52.90	46.64	50.35

Table 2: ChrF++ scores for our benchmarked (above) and trained models (below). Best and worst scores are highlighted for each language direction.

the data to control training cost. We use 100k sentence pairs per language direction. We use FLORES+ for both development and evaluation.

Vocabulary For LLaMA fine-tuning, we use the original LLaMA tokenizer and introduce an additional padding token. For Marian models, we train a multilingual SentencePiece (Kudo and Richardson, 2018) vocabulary with 32k merge operations, learned jointly over the four languages in our setup. The vocabulary is trained on a balanced 50k-sentence sample per language to avoid dominance of higher-resource languages.

4.3. Evaluation Metrics

We report both surface-level and neural evaluation metrics. We compute BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017), as n-gram overlap–based measures. In addition, we report neural metrics, namely COMET[6] (Rei et al., 2022) and MetricX[7] (Juraska et al., 2024). Neither COMET nor MetricX include Esperanto in their fine-tuning stages. As a result, their scores should be interpreted with caution, as their calibration for Esperanto may be suboptimal. Following recent findings in shared tasks (Lavie et al., 2025), we adopt ChrF++ as our primary metric, as it has been shown to correlate more strongly with human judgments than BLEU, especially for morphologically rich or lower-resource languages.

[6]We use the model Unbabel/wmt22-comet-da.

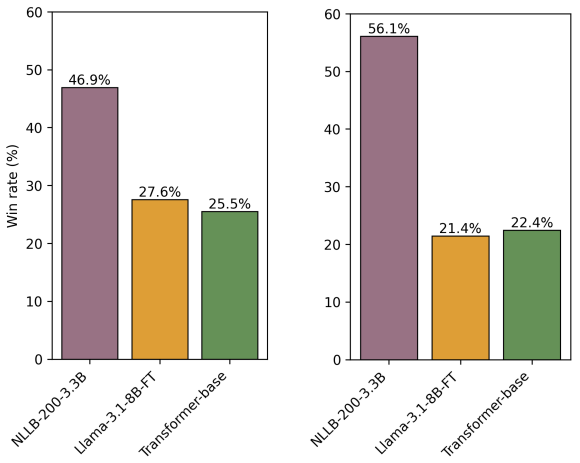
[7]We use the model google/metricx-24-hybrid-large-v2p6.

5. Results

Table 2 reports the ChrF++ scores for both benchmarked models and models trained in this work. Appendix C shows the scores for BLEU, COMET and MetricX. In general, the four metrics follow similar trends, with similar top- and bottom-performing systems. Across systems, translating from Esperanto generally yields higher scores than translating into Esperanto. This asymmetry likely reflects the richer training data available for English, Spanish, and Catalan compared to Esperanto. Consequently, models appear to encode stronger representations for the high-resource languages, while Esperanto generation is more challenging.

5.1. Benchmarked Models

The NLLB family consistently outperforms all other models by a clear margin. Performance increases with model size, with NLLB-200-3.3B achieving the highest scores in most directions. The distilled 1.3B variant performs competitively, even slightly surpassing the 3.3B model in one direction (en–eo). As hypothesized, LLMs have sufficient pretraining information to understand and produce Esperanto. However, they underperform when



(a) Spanish → Esperanto (b) Esperanto → Spanish

Figure 3: Human evaluation win rates for both translation directions.

compared to dedicated MT systems. While Llama-3.1-8B-Instruct produces reasonable results, both Tower variants lag considerably behind. In particular, TowerInstruct-7B-v0.2 exhibits extremely low scores in several directions, suggesting that its instruction tuning on specific high-resource languages may have negatively affected its general translation capabilities. A qualitative inspection of its outputs showed that the model often failed to generate Esperanto or Catalan. While it performed somewhat better in directions involving English and Spanish, it still struggled to maintain the requested target language, frequently mixing languages or producing malformed output. TowerPlus-9B performs more competitively but still falls short of NLLB and strong neural baselines. The model achieves higher performance when translating from Esperanto than when translating into Esperanto. Rule-based Apertium performs substantially worse than neural approaches overall, though, surprisingly, it still surpasses the Tower variants in several directions.

5.2. Trained Models

Among the models trained in this work, fine-tuning Llama-3.1-8B-Instruct yields only modest improvements. Gains are more noticeable when generating Esperanto, especially for Catalan and Spanish as a source. Performance into English decreases slightly. This is consistent with the expectation that the base model already possesses strong English representations due to extensive pretraining. Despite its substantially smaller size, the Transformer-base model performs comparably to Llama-3.1-8B-Instruct-FT across most directions, surpassing it in 4 out of 6 pairs. Notably, the Transformer-tiny model achieves surprisingly competitive results,

Spanish → Esperanto				
Metric	τ (mean)	τ (pooled)	P-value	Accuracy
BLEU	0.061	-0.023	0.610	0.522
ChrF++	0.116	-0.049	0.276	0.558
COMET	0.191	-0.110	**0.015***	0.595
MetricX	0.327	0.165	**2.6x10-4***	0.663

Esperanto → Spanish				
Metric	τ (mean)	τ (pooled)	P-value	Accuracy
BLEU	0.361	-0.123	**6.24x10-3***	0.668
ChrF++	0.347	-0.146	**1.18x10-3***	0.672
COMET	0.422	-0.177	**8.6x10-5***	0.711
MetricX	0.415	0.200	**9.5x10-6***	0.708

Table 3: Correlation between human judgments and automatic evaluation metrics. Results are reported as mean Kendall's τ over per-sentence rankings, pooled Kendall's τ over metric scores with the corresponding p-value, and pairwise accuracy. Highest and lowest values are highlighted. Statistically significant values ($p < 0.05$) are bolded and marked with an asterisk (*).

particularly given its limited parameter count. This suggests that compact, task-specific architectures remain strong contenders in low-resource multilingual settings.

6. Human Evaluation

To complement the automatic evaluation, we conducted a human assessment of translation quality for the Spanish–Esperanto language pair. We randomly sampled 100 source sentences and extracted the corresponding translations produced by the top three models for each architecture, namely, Transformer-base, NLLB-200-3.3B, and Llama-3.1-8B-Instruct-FT. For every source sentence, the three system outputs were presented in randomized order to a human annotator, who was asked to rank them by selecting the best and the worst translation[8]. An optional comment field allowed the annotator to justify their choice or note specific errors briefly, without any specific guidelines. This pairwise ranking setup (Läubli et al., 2018) enables direct comparison between models while keeping the annotation procedure simple and intuitive. The annotation guidelines can be found in Appendix D.

6.1. Results

Figure 3 shows the win rates achieved by the three models on the human evaluation task for both

[8]Two human annotators were employed, one for each translation direction. L1: Catalan, Spanish. L2: Esperanto.

language directions[9]. These results confirm the trends observed in the automatic metrics. In both cases, NLLB stands out as the clear winner, selected as the best translation around 50% of the time. The other two systems perform considerably worse and at a similar level. The advantage of NLLB is particularly pronounced in the Esperanto → Spanish direction, which suggests once more that generating Esperanto is generally more challenging across models. However, these results reflect relative differences between systems rather than absolute translation quality, and even the best-performing system still produces noticeable errors.

6.2. Metric Correlations with Human Judgements

We compute correlations of the human judgments with the four automatic metrics with Kendall's τ (Macháček and Bojar, 2013; Deutsch et al., 2023). We use three complementary measures: (i) Kendall's τ over model rankings, computed per sample and averaged, to measure how well each metric reproduced the human ordering of translations; (ii) pooled Kendall's τ over model scores, computed across all translations, to measure the global monotonic relationship between metric scores and human quality; and (iii) pairwise accuracy, measuring how often a metric correctly identified the better translation in each pairwise comparison.

Table 3 shows the results for both translation directions. The results reveal a consistent hierarchy of metric quality across directions. MetricX and COMET achieve the strongest agreement with human judgments, with MetricX performing best overall and COMET showing comparable performance, particularly in the Esperanto→Spanish direction. These metrics show statistically significant agreement in both directions. In contrast, ChrF++ shows weaker agreement with human rankings, while BLEU performs close to random in the Spanish→Esperanto direction but shows moderate correlation in the other. Overall, the learned metrics correlate substantially better with human judgments than traditional n-gram-based metrics, even though they have not been directly exposed to Esperanto in the fine-tuning stage.

6.3. Qualitative Error Analysis

We summarize the free comments from the human evaluation and provide qualitative insights into recurring error patterns. Appendix E provides illustrative examples.

[9]We removed two sentences for each translation direction where two of translations resulted in a tie.

For the Transformer-base model, we observe a range of lexical and grammatical errors. The model occasionally produces non-existent words or leaves source words untranslated. Grammatical problems include incorrect verb forms, agreement errors between articles and nouns, and missing verbs. The model also struggles with named entity translation and occasionally mistranslates relatively simple lexical items. In contrast, the NLLB-200-3.3B model generally produces fluent and semantically adequate translations. Most errors appear to stem from the compositional nature of Esperanto word formation. For example, *nekredantoj* is rendered as *incrédulos*, while *neĝtabulo* and *flugaparatoj* are translated compositionally as *tabla de nieve* and *aparatos de vuelo*. Although these translations remain understandable, they are less appropriate than conventional equivalents (*snowboard*, *aviones*). In one case, it omits semantically relevant information. Finally, the Llama-3.1-8B-Instruct-FT model exhibits the widest range of error types. The model recurrently adds, omits, or invents information. In addition, it sometimes introduces English words into the output and produces grammatical errors of varying severity, including agreement errors and semantic distortions such as incorrectly assigning the subject.

These error patterns are in line with our expectations. The Transformer-base model tends to produce accurate but literal translations, it sometimes generates unusual constructions and wrong named entity translations. Both NLLB-200-3.3B and Llama-3.1-8B-Instruct-FT produce fluent output; however, NLLB-200-3.3B can occasionally be overly literal, while Llama more frequently modifies or invents information, which may pose risks in real-world deployment.

7. Discussion

In this section, we discuss the main findings of our experiments and their implications for Esperanto translation and, more broadly, low-resource MT.

7.1. NLLB Remains a Strong Baseline

Although the NLLB model family is now several years old, it remains by far the strongest system in our experiments. We hypothesize that this is due to NLLB's highly multilingual training, which allows it to benefit from transfer learning, as well as its explicit training for translation. This finding is consistent with recent work on low-resource MT, where NLLB continues to outperform a wide range of alternative approaches (de Gibert et al., 2025; Scalvini et al., 2025; Tapo et al., 2025; Aycock et al., 2025). The distilled 1.3B model performs slightly better than its non-distilled counterpart. Even the

smallest model in the family achieves strong results across language directions. These observations support a clear practical recommendation: for low-resource MT, NLLB should be considered the default choice, with model size selected according to available computational resources, prioritizing distilled variants.

7.2. General-Purpose LLMs Outperform MT-Tuned LLMs

The MT-tuned LLMs evaluated in this work, which are specifically designed for translation tasks, consistently underperform the general-purpose LLM baseline. In particular, TowerInstruct-7B-v0.2, which was fine-tuned primarily on 10 high-resource languages, is largely unable to produce meaningful output beyond English; even though its fine-tuning also includes Spanish. Tower-Plus-9B performs similarly to Llama-3.1-8B-Instruct-FT but is poor at Esperanto generation. These findings suggest that, for low-resource scenarios where NLLB coverage is unavailable, general-purpose LLMs appear to be a more reliable choice than translation-specialized LLMs, since language-specific fine-tuning may reduce their general translation abilities.

7.3. Data-Hungry vs. Compute-Hungry Models

In our training setup, we compare a fine-tuned Llama model with encoder–decoder models. When only limited training data is available, fine-tuning a pretrained Llama model is the most effective approach: even with as little as 100k sentence pairs per language direction, fine-tuning yields competitive results. However, under constrained computational budgets, training models from scratch becomes an attractive alternative, provided that sufficient training data is available. Our smallest Transformer models are more than 500 times smaller than Llama-3.1-8B-Instruct-FT, yet achieve comparable and, in one case (en-eo), superior performance. This result highlights a broader tendency in the field toward increasingly large architectures, even in scenarios where smaller models can achieve similar quality. Moreover, our compact models can run efficiently on standard CPUs, making them suitable for deployment on personal devices. This aligns well with the principles of Esperanto as a language intended to facilitate universal communication, as lightweight models lower the computational barriers to MT. To support accessibility and reproducibility, we release our Transformer models on HuggingFace.

7.4. Neural Metrics are Effective in Zero-Shot Settings

Our analysis in Table 3 shows that learned metrics correlate substantially better with human judgments than traditional metrics. Neither metric has been explicitly fine-tuned on Esperanto data. MetricX is based on mT5 (Xue et al., 2021), while COMET builds on XLM-RoBERTa-base (Conneau et al., 2020); both pre-trained models include Esperanto and Spanish in their multilingual training data. However, the subsequent fine-tuning of these metrics on WMT datasets (Kocmi et al., 2025) does not involve Esperanto. The strong performance observed in our experiments, therefore, suggests that transfer learning enables neural metrics to generalize effectively to previously unseen language pairs. We hypothesize that the strong performance of these metrics may be partly explained by the linguistic characteristics of Esperanto, which is largely derived from Romance languages. A systematic evaluation across a broader set of low-resource languages would be necessary to assess the generalization of these findings.

8. Conclusions

Esperanto is a widely used conlang whose community aligns with ideals of linguistic and technological sovereignty. We systematically study Esperanto translation with a particular emphasis on open models. We evaluate a range of existing systems and develop compact models of our own, demonstrating that high translation quality can be achieved with remarkably small architectures. Our results confirm that NLLB remains the strongest overall system, while general-purpose LLMs perform similarly to our task-specific Transformer models despite being orders of magnitude larger. The efficiency of these smaller models makes them faster, more accessible, and environmentally sustainable, aligning closely with the practical and ideological goals of the Esperanto community. Ensuring the continued development of free MT systems is essential for maintaining a digital linguistic infrastructure that can be governed, audited, and adapted by members of the community who rely on it. This work represents a first step toward that goal.

For future work, we are interested in studying whether Esperanto is inherently easier to model than natural languages due to its regular morphological and syntactic structure, following work by Ploeger et al. (2025). Another important direction would be to revisit the original idea of Esperanto as an interlingua for pivot-based MT. Finally, given the availability of rule-based resources in Apertium, future work could explore hybrid approaches that leverage RBMT knowledge (De Gibert et al., 2024).

Limitations

We are aware of our limited evaluation setup covering only one test set, Flores+. Flores+ was created by translating directly from English, which may introduce biases that affect evaluation. However, Esperanto is not present in any other MT benchmark. To compensate for this, we perform human evaluation and report a diverse set of evaluation metrics. Furthermore, our human evaluation is limited to 100 samples per language direction and one annotator per direction due to a lack of resources.

Ethical considerations

All annotators are authors of this paper, and the total time spent on individual annotations did not exceed four hours.

Acknowledgments

We thank Seth Aycock and Joseph Attieh for their insightful feedback and valuable comments.

This project has received funding from the Digital Europe programme of the European Union under Grant No. 101195233. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

9. Bibliographical References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. Opusfilter: A configurable parallel corpus filtering toolbox. In *2020 Annual Conference of the Association for Computational Linguistics*, pages 150–156. The Association for Computational Linguistics.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book? In *Proceedings of The Thirteenth International Conference on Learning Representations*, pages 12334–12357.

Eckhard Bick. 2011. Wikitrans: the english wikipedia in esperanto. In *Constraint Grammar Applications, Workshop Proceedings at Nodalida*, volume 14, pages 8–16.

Eckhard Bick. 2016. A morphological lexicon of Esperanto with morpheme frequencies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1075–1078, Portorož, Slovenia. European Language Resources Association (ELRA).

Eckhard Bick. 2020. Syntax and semantics in a treebank for Esperanto. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5120–5127, Marseille, France. European Language Resources Association.

Eckhard Bick. 2025. An annotated error corpus for Esperanto. In *Proceedings of the 9th Workshop on Constraint Grammar and Finite State NLP*, pages 1–8, Tallinn, Estonia. University of Tartu Library.

Detlev Blanke. 2009. Causes of the relative success of esperanto. *Language Problems and Language Planning*, 33(3):251–266.

Richard Boddington. 2004. Evaluation of an esperanto-based interlingua multilingual survey form machine translation mechanism incorporating a sublanguage translation methodolgy.

Nikolay Bogoychev, Roman Grundkiewicz, Alham Fikri Aji, Maximiliana Behnke, Kenneth Heafield, Sidharth Kashyap, Emmanouil-Ioannis Farsarakis, and Mateusz Chudyk. 2020. Edinburgh's submissions to the 2020 machine translation efficiency task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 218–224, Online. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Manuela Burghelea. 2019. On not being lost in translation: Creative strategies to approach multiculturalism in esperanto. *Język. Komunikacja. Informacja*, (13):159–174.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

L Couturat. 1903. Histoire de la langue universelle. *Hildesheim, Zürich, & New York: Olms*.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Ona De Gibert, Mikko Aulamo, Yves Scherrer, and Jörg Tiedemann. 2024. Hybrid distillation from RBMT and NMT: Helsinki-NLP's submission to the shared task on translation into low-resource languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, pages 908–917, Miami, Florida, USA. Association for Computational Linguistics.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Alison J. Robinson, editors. 2026. *Ethnologue: Languages of the World*, 29 edition. SIL Global, Dallas, Texas. Online version.

Sabine Fiedler. 2018. Linguistic and pragmatic influence of english: Does esperanto resist it? *Journal of Pragmatics*, 133:166–178.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Marcos Franco Sabarís, José Luis Rojas Alonso, C. Dafonte, and B. Arcay. 2001. Multilingual authoring through an artificial language. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Federico Gobbo. 2015. Machine translation as a complex system: The role of esperanto. *Interdisciplinary Description of Complex Systems: INDECS*, 13(2):264–274.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli-Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

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.

William John Hutchins and Harold L Somers. 1992. An introduction to machine translation. *(No Title)*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, et al. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Tobias Kuhn. 2014. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Bill Z Manaris, Luca Pellicoro, George J Pothering, and Harland Hodges. 2006. Investigating esperanto's statistical proportions relative to other languages using neural networks and zipf's law. In *Artificial Intelligence and Applications*, pages 102–108.

Carlo Minnaja and Laura Paccagnella. 2000. A part-of-speech tagger for Esperanto oriented to MT. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*, University of Exeter, UK.

A Neijt. 1986. Esperanto as the focal point of machine translation. *Multilingua*, 5(1):9–13.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Dayyán O'Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Suzhou, China. Association for Computational Linguistics.

Giulia Occhini, Kumiko Tanaka-Ishii, Anna Barford, Refael Tikochinski, Songbo Hu, Roi Reichart, Yijie Zhou, Hannah Claus, Ulla Petti, Ivan Vulić, et al. 2026. Artificial intelligence is creating a new global linguistic hierarchy. *arXiv preprint arXiv:2602.12018*.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, et al. 2025. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono-and bi-lingual data, multilingual evaluation, and pre-trained models. *arXiv preprint arXiv:2511.01066*.

Darja Orlova. 2015. Esperus: the first step to build a statistical machine. translation system for esperanto and russian languages. *AINL FRUCT, Saint Petersburg, Russia*.

Masanori Oya. 2025. UD treebanks for Esperanto as a natural language. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 22–29, Ljubljana, Slovenia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mikael Parkvall. 2010. How european is esperanto?: A typological study. *Language Problems and Language Planning*, 34(1):63–79.

Esther Ploeger, Johannes Bjerva, Jörg Tiedemann, and Robert Östling. 2025. A cross-lingual perspective on neural machine translation difficulty. In *Proceedings of the Tenth Conference on Machine Translation*, pages 340–354, Suzhou, China. Association for Computational Linguistics.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. 2020. Using multiple subwords to improve english-esperanto automated literary translation quality. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 108–117.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. Rethinking low-resource MT: the surprising effectiveness of fine-tuned multilingual models in the LLM age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Klaus Schubert. 2002. Esperanto as an intermediate language for machine translation. In *Computers in translation*, pages 98–115. Routledge.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique, and Marcos Zampieri. 2025. Bayelemabaga: Creating resources for Bambara NLP. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12060–12070, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

UNESCO. 1954. Records of the general conference, eighth session, montevideo 1954; resolutions. UNESDOC Database. P. 36. Archived from the original (PDF) on February 2, 2011. Retrieved May 16, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Amri Wandel. 2015. How many people speak esperanto? esperanto on the web. *Interdisciplinary Description of Complex Systems: INDECS*, 13(2):318–321.

A. P. M. Witkam. 1984. Distributed language translation, another MT system. In *Proceedings of the International Conference on Methodology and Techniques of Machine Translation: Processing from words to language*, Cranfield University, UK.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A. Training details

Below, we list the hyperparameters used during training for Transformer-base (Table 5), Transformer-tiny (Table 6), and Llama fine-tuning (Table 7). We also report the architectures for the tiny and base Transformer models in Table 4.

	base	tiny
N_{enc}	6	6
N_{dec}	6	2
d_{emb}	1024	256
d_{ff}	2048	1536
h	8	8
Params (M)	60.6	17.4
Size (MB)	232	67

Table 4: Transformer architectures for base and tiny. The table lists the number of encoder and decoder layers (N_{enc} and N_{dec}), embedding dimensions (d_{emb}), feed-forward dimensions (d_{ff}), number of attention heads (h), parameters in millions, and model size in MB.

Hyperparameter	Value
Optimizer	Adam
Adam β	(0.9, 0.98)
Adam ϵ	1e-9
Learning rate	3e-4
LR warmup	16000 steps
LR decay strategy	epoch+stalled
LR decay start	(10 epochs, 1 stalled)
Optimizer delay	1
Validation frequency	2500 updates
Early stopping	10 (on perplexity)
Seed	1111

Table 5: Hyperparameters for Transformer-base

Hyperparameter	Value
Optimizer	Adam
Adam β	(0.9, 0.98)
Adam ϵ	1e-9
Learning rate	3e-4
LR warmup	16000 steps
LR decay strategy	epoch+stalled
LR decay start	(10 epochs, 1 stalled)
Optimizer delay	2
LR inv-sqrt factor	32000
Validation frequency	5000 updates
Early stopping	20 (on cross-entropy)
Seed	0

Table 6: Hyperparameters for Transfomer-tiny

Hyperparameter	Value
Learning Rate	5e-05
LR Scheduler Type	Linear
Warmup Ratio	0.3
Weight Decay	0.0
Per Device Train Batch Size	4
Gradient Accumulation Steps	4
Number of Train Epochs	1
LoRA Rank	16
LoRA Alpha	32
Seed	123

Table 7: Hyperparameters for Llama fine-tuning

Hyperparameter	Value
Optimizer	Adafactor
Learning Rate	1e-4
LR Scheduler Type	Constant
LR Warmup	1000 steps
Weight Decay	1e-3
Per Device Train Batch Size	32
Gradient Accumulation Steps	2
Maximum Sequence Length	128
Number of Train Epochs	4
Validation Frequency	1000 updates
Early Stopping	5
Seed	42

Table 8: Hyperparameters for NLLB fine-tuning

B. Details of the NLLB fine-tuning experiments

We conducted exploratory fine-tuning experiments with Hugging Face implementations of NLLB. We fine-tuned the 600M- and 3.3B-parameter models in a multilingual setup, training one model for each Esperanto language direction. The training data were the same as those used for Marian. Details of the fine-tuning hyperparameters are given in Table 8. Under this configuration, fine-tuning did not improve over the base NLLB models, with results generally comparable to or slightly below those of the original checkpoints.

C. Automatic Evaluation Results

We report automatic evaluation results using BLEU (Table 9), COMET (Table 10), and MetricX (Table 11).

D. Annotation Guidelines

Figure 4 shows the guidelines presented to the annotators for the human evaluation task.

Annotation Task. For each source sentence, you will see three possible translations (T1, T2, and T3). Read them carefully and indicate which translation is the best and which is the worst. Write 1, 2, or 3 in the corresponding columns.

You may also add a short optional comment explaining your decision (e.g., if something sounds unnatural or contains a clear error). There is no need for technical analysis, simply choose the translation that sounds most natural and most faithful to the original meaning.

Figure 4: Annotation guidelines shown to the human annotator.

E. Qualitative Error Analysis

Tables 12 and 13 present illustrative examples of recurring errors in the evaluated models for translation into Spanish and Esperanto, respectively.

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Rule-based MT						
Apertium	19.94	-	-	20.80	10.99	14.34
Neural MT						
NLLB-200-distilled-600M	43.04	21.74	27.98	31.86	18.18	24.20
NLLB-200-1.3B	44.66	23.37	30.82	33.11	18.83	24.26
NLLB-200-distilled-1.3B	45.49	23.63	31.63	33.52	18.98	24.26
NLLB-200-3.3B	46.05	23.50	32.10	33.47	19.25	24.54
General-purpose LLMs						
Llama-3.1-8B-Instruct	40.05	19.08	23.03	27.57	13.52	19.70
MT-tuned LLMs						
TowerInstruct-7B-v0.2	27.28	14.61	0.35	4.66	3.08	2.72
Tower-Plus-9B	42.74	21.43	23.01	17.80	10.80	14.10

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Neural MT from Scratch						
Transformer-base (60.6M)	37.47	20.00	28.35	26.42	16.25	21.43
Transformer-tiny (17.4M)	33.13	18.49	23.58	25.69	15.04	20.78
Fine-tuned General-purpose LLMs						
Llama-3.1-8B-Instruct-FT	38.55	19.61	24.98	25.14	17.17	22.33

Table 9: BLEU scores for our benchmarked (above) and trained models (below). Best and worst scores are highlighted for each language direction.

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Rule-based MT						
Apertium	70.43	-	-	77.67	76.02	71.77
Neural MT						
NLLB-200-distilled-600M	87.80	82.10	81.99	88.92	86.09	85.89
NLLB-200-1.3B	88.58	83.71	84.22	89.74	86.80	86.62
NLLB-200-distilled-1.3B	88.72	83.85	84.52	89.85	86.93	86.24
NLLB-200-3.3B	88.82	84.03	85.07	89.82	86.96	86.80
General-purpose LLMs						
Llama-3.1-8B-Instruct	87.23	80.69	77.93	87.09	82.64	82.92
MT-tuned LLMs						
TowerInstruct-7B-v0.2	75.54	65.93	34.61	50.86	54.82	56.36
Tower-Plus-9B	87.19	81.79	77.79	76.62	73.71	73.63

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Neural MT from Scratch						
Transformer-base (60.6M)	86.07	80.66	82.30	85.95	83.51	80.99
Transformer-tiny (17.4M)	81.94	76.44	73.43	84.12	81.59	80.72
Fine-tuned General-purpose LLMs						
Llama-3.1-8B-Instruct-FT	86.97	81.27	81.65	85.61	85.40	85.68

Table 10: COMET scores for our benchmarked (above) and trained models (below). Best and worst scores are highlighted for each language direction.

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Rule-based MT						
Apertium	9.90	-	-	8.04	7.36	8.68
Neural MT						
NLLB-200-distilled-600M	2.88	3.26	4.15	4.09	4.56	4.90
NLLB-200-1.3B	2.57	2.73	3.36	3.74	4.19	4.71
NLLB-200-distilled-1.3B	2.57	2.67	3.35	3.59	4.08	4.85
NLLB-200-3.3B	2.52	2.63	3.12	3.64	4.10	4.57
General-purpose LLMs						
Llama-3.1-8B-Instruct	3.05	3.62	5.12	4.98	5.67	5.87
MT-tuned LLMs						
TowerInstruct-7B-v0.2	7.33	8.81	13.03	12.48	14.18	12.37
Tower-Plus-9B	3.08	3.28	5.10	7.82	8.36	8.54

	eo-en	eo-es	eo-ca	en-eo	es-eo	ca-eo
Neural MT from Scratch						
Transformer-base (60.6M)	3.47	3.61	3.94	4.68	5.34	6.73
Transformer-tiny (17.4M)	5.15	5.11	6.89	5.53	6.03	6.73
Fine-tuned General-purpose LLMs						
Llama-3.1-8B-Instruct-FT	3.33	3.77	4.42	5.50	4.86	5.15

Table 11: MetricX scores for our benchmarked (above) and trained models (below). Best and worst scores are highlighted for each language direction. Lowest is best for this metric.

Error Type	Source	Translation
	Transformer-base	
Lexical	*La epidemio igis la baratan registaron entrepreni tiajn rimedojn, kiel instalado de porkokaptiloj en la serioze damaĝitaj areoj, disdonado de miloj da kontraŭmoskitaj retoj kaj ŝprucado de pesticidoj.*	*La epidemia llevó al gobierno indio a emprender tales medios, como la instalación de trampas de cerdo en las áreas seriamente dañadas, la distribución de miles de redes anti-moscoquios y el brote de pesticidas.*
Grammatical	*Malgraŭ foresto de cunama minaco, loĝantoj ekpanikis kaj komencis forlasadi siajn komercojn kaj hejmojn.*	*A pesar de la ausencia de la amenaza tsunami, los residentes comenzaron a pánico y comenzaron a dejar sus negocios y casas.*
Untranslated	*Gregon konsistigas de unu ĝis tri parencaj plenkreskaj maskloj kaj multaj ĝis tridek inoj kaj idoj.*	*Gregon consiste de uno a tres varones adultos relacionados y muchos a treinta hembras y crías.*
	NLLB-200-3.3B	
Mistranslation	*Nekredantoj vivantaj en kristanaj landoj aŭ aliloke en la tuta mondo adoptis multajn el la tradicioj pri ĉi tiu festo.*	*Los incrédulos que viven en países cristianos o en otras partes del mundo han adoptado muchas de las tradiciones de esta fiesta.*
	Malsupreniraj vintrosportoj, kiuj inkluzivas skiadon kaj neĝtabuladon, estas popularaj sportoj, en kiuj oni glitas malsupren sur neĝokovrita tereno havante skiojn aŭ neĝtabulon fiksitaj al la piedoj.	*Los deportes de invierno de descenso, que incluyen el esquí y el snowboard, son deportes populares en los que se desliza hacia abajo en un terreno cubierto de nieve con esquís o una tabla de nieve fijada a los pies.*
	Pro internaciaj sankcioj, novaj flugaparatoj ne estas aĉeteblaj.	*Debido a las sanciones internacionales, no se pueden comprar nuevos aparatos de vuelo.*
	Llama-3.1-8B-Instruct-FT	
Addition	*Ni penos malaltigi eligon de karbona dioksido po unuo de MEP je rimarkinda diferenco ĝis la jaro 2020, kompare al la nivelo de 2005", Hu diris.*	*"Haremos todo lo posible para reducir las emisiones de dióxido de carbono por unidad de PIB en un 20% hasta el 2020, en comparación con el nivel de 2005", dijo Hu.*
Grammatical	*Post kiam oficialuloj konfirmas identecon de la voĉdonanto, la voĉdonanto enmetas la koverton en la balotujon kaj subskribas la voĉdonan liston.*	*Después de que los funcionarios verifican la identidad del votante, el votante coloca el sobre en el urna y firma la lista de votación.*
Hallucination	*Gregon konsistigas de unu ĝis tri parencaj plenkreskaj maskloj kaj multaj ĝis tridek inoj kaj idoj.*	*Una colonia consta de un macho adulto y varias hembras y hasta treinta crías.*
Lexical	*Del Potro havis fruan avantaĝon en la dua serio, sed ankaŭ tiu bezonis decidludon, kiam la rezulto atingis 6-6.*	*Del Potro tenía ventaja en el segundo set, pero necesitó un tiebreaker, ya que el resultado se igualó en 6-6.*
Mistranslation	*Tio nomiĝas pH de kemiaĵo. Vi povas fari indikanton per uzo de suko de ruĝa beto.*	*Se llama pH de un compuesto. Puedes hacer un indicador con jugo de betún.*
Omission	*Tigro estas en la sama grupo (Genus Panthera) kun leonoj, leopardoj kaj jaguaroj.*	*El tigre pertenece al mismo género (Panthera) que los leones, leopardos y jaguar.*
	Tiuj ĉi kvar katoj estas la solaj, kiuj povas rori.	

Table 12: Illustrative examples of error categories identified in the qualitative analysis for Esperanto into Spanish. Highlighted spans mark erroneous content.

Error Type	Source	Translation
	Transformer-base	
Grammatical	*En verdad, el formato de 35 mm es algo confuso, ya que sus medidas son 36 mm de anchura por 24 mm de alto.*	*Fakte, la 35mm formato estas iom konfuzita, ĉar ĝiaj mezuroj estas 36 mm larĝa de 24 mm alta.*
Mistranslation	*Conforme surge de las imágenes infrarrojas, las variaciones de temperatura durante la noche y el día indican que, probablemente, se trate de cuevas.*	*Ĉar ĝi ekestiĝas de infraruĝaj bildoj, temperaturkatastrofoj dum la nokto kaj tago indikas ke ili estas verŝajne kavernoj.*
Named Entities	*El día de hoy se originó en el Océano Atlántico la tormenta número diez de su temporada de huracanes en ser nombrada, a la que se denominó Tormenta subtropical Jerry.*	*La atlantika ŝtormo numero dek el sia uragansezono estis nomita Jerry Storm.*
	La cueva en sí misma, que perduró en el tiempo, ilustra de forma muy realista las inclinaciones espirituales de Mahoma.	*La kaverno mem, kiu daŭris en tempo, tre racie ilustras la spiritajn inklinojn de Mohamedo.*
	NLLB-200-3.3B	
Mistranslation	*Los científicos creen que los ocelotes rastrean y encuentran animales para comer (presas) a través del olfato, oliendo el suelo donde han pasado.*	Sciencistoj kredas, ke la ocelotoj spuras kaj trovas bestojn por manĝi (ofero) per la flaro, flarante la grundon, kiun ili pasis.
Omission	*A pesar de que había tres personas en el interior de la vivienda en el momento del impacto del automóvil, ninguna de ellas sufrió lesiones.*	*Kvankam tri homoj estis en la domo, neniu el ili vundiĝis.*
	Llama-3.1-8B-Instruct-FT	
Grammatical	*Duvall, que está casado y tiene dos hijos adultos, no causó una buena impresión a Miller, que fue a quien le relató la historia.*	*Duvall, kiu estis edziĝinta kaj havis du plenkreskajn infanojn, ne faris bonan impreson al Miller, kiu rakontis la historion.*
Lexical	*El mismo mes, otro avión de pasajeros sobrepasó una pista en Mashhad y colisionó contra una pared, dejando 17 muertos.*	*En la sama monato, alia pasaĝera aviadilo preterflugis Mashhad's runvojo kaj koliziis kun muro, mortigante 17 homojn.*
Mistranslation	*Serían las fuerzas de marea entre la materia oscura presente en la galaxia y la de la Vía Láctea lo que provocaría el efecto esperado por el grupo de científicos.*	*La fortoj de la maroj inter la malluma materio en la galaksio kaj tiu de la Vida Laktone estas tiuj kiuj kaŭzus la atendatan efikon, laŭ la sciencistoj.*
Omission	*Hacia 1976 ya se había restaurado el treinta por ciento del Machu Picchu y los trabajos continúan en la actualidad.*	*En 1976 estis jam restarigita la triono de la konstruaĵo kaj la laboroj daŭras.*

Table 13: Illustrative examples of error categories identified in the qualitative analysis for Spanish into Esperanto. Highlighted spans mark erroneous content.