Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 349-369 recibido 07-04-2025 revisado 05-05-2025 aceptado 18-05-2025

From Rule-Based to LLMs: A Performance and Variability Analysis of Galician Machine Translation Models

De los Sistemas basados en Reglas a los Modelos LLM: un Análisis de Rendimiento y Variabilidad de los Modelos de Traducción Automática para el Gallego

Sofía García González,[1] German Rigau Claramunt,[2] José Ramom Pichel Campos[3]
[1]UPV/EHU & imaxin,
[2]IXA GROUP (UPV/EHU),
[3]Centro de Investigación en Tecnoloxías da Información (CiTIUS)
[1]sofia.garcia@imaxin.com, [2]german.rigau@ehu.es, [3]jramon.pichel@usc.gal

Abstract: This paper evaluates machine translation (MT) for English–Galician, Spanish–Galician, and Portuguese–Galician pairs, with the aim of identifying the most effective models for these language pairs in the general domain. The evaluation encompasses a range of factors, including model quality, performance variance and size. The assessment involves the evaluation of different open-source systems. The results obtained identify that, for Spanish–Galician, both a Rule-Based System and a bilingual Neural Machine Translation model outperform larger multilingual models and LLMs. However, for more distant language pairs, multilingual models demonstrate superior performance. The study underscores the necessity for further research in Portuguese–Galician pair.
Keywords: Galician, Machine Translation, Variability.

Resumen: Este trabajo evalúa la traducción automática (TA) para los pares Inglés–Gallego, Español–Gallego y Portugués–Gallego, con el objetivo de identificar los modelos más efectivos en un dominio generalista. La evaluación abarca factores como calidad, variabilidad del rendimiento y tamaño. Los resultados muestran que, para Español–Gallego, los sistemas basados en reglas y los modelos bilingües superan a los modelos multilingües y LLMs. Sin embargo, en pares de lenguas más distantes, los modelos multilingües ofrecen mejores resultados. Se destaca la necesidad de más investigación para Portugués–Gallego en TA.
Palabras clave: Gallego, Traducción Automática, Variabilidad.

1 Introduction

This paper presents an evaluation of machine translation (MT) systems for English–Galician, Spanish–Galician, and Portuguese–Galician, with the objective of identifying the most promising models for these language pairs, which represents a pioneering endeavour for the Portuguese–Galician pair. The evaluation encompasses a range of factors, including model quality, models performance variability, size, and computational cost, factors that are not usually given due consideration during the process of machine translation evaluation. A range of open-source models are assessed, including a Rule-Based Machine Translation (RBMT) system, Sequence-to-Sequence models, and Large Language Models (LLMs), in order to determine the best trade-offs between accuracy, resource consumption, and scalability.

The findings indicate that for the Spanish–Galician pair, both a Rule-Based System and a bilingual Neural Machine Translation model are competitive in this context when compared to larger multilingual models and LLMs. However, for more distantly related language pairs, such as English–Galician, sequence-to-sequence multilingual models consistently yield the best results. Finally, the performance observed for the Portuguese-Galician pair is particularly noteworthy. Despite the close linguistic relationship between Portuguese and Galician, the results are unexpectedly low. In

ISSN 1135-5948 DOI 10.26342/2025-75-25

fact, the scores are even lower than those obtained for the English–Galician pair, given the significantly greater distance between English and Galician. This counterintuitive outcome suggests that additional research is needed for this pair, highlighting the scarcity of test data sets and the need for further improvements and investigation in MT performance.

The structure of this paper is as follows: Section 2 reviews prior work, Section 3 outlines the methodology, i.e., models, datasets, and metrics used, while Section 4 details the experiments carried out. Section 5 presents the results that were analysed in section 6, and Section 7 includes the conclusions and future work.

2 Related work

Neural Machine Translation (NMT) has evolved significantly, since Transformer-based architectures (Vaswani et al., 2023) set new standards for translation quality. Sequence-to-sequence models (Sutskever, Vinyals, and Le, 2014) have been widely adopted for NMT, demonstrating effectiveness in high-resource languages but struggling with low-resource settings due to data scarcity.

To address the challenges of low-resource language translation, researchers have proposed several strategies such as transfer learning (Zoph et al., 2016), multilingual NMT (Artetxe and Schwenk, 2019), and unsupervised learning approaches (Lample et al., 2017). Recent advancements in LLMs, such as GPT-4 (OpenAI et al., 2024) and LLaMA (Touvron et al., 2023), have demonstrated impressive capabilities in zero-shot and few-shot translation tasks. By leveraging massive pretraining corpora, these models offer significant improvements even in the absence of large parallel datasets.

Concerning Galician machine translation, in 2012, (García-Mateo and Arza, 2012) expressed cautious optimism regarding Galician's linguistic technology support, noting the urgent need for further resource development—a sentiment later reinforced by ELE-D1.15 (Sánchez and Mateo, 2022). Galician has been included in various multilingual systems that account for minority languages, such as Apertium (Forcada et al., 2011), M2M (Fan et al., 2021), and NLLB (Costa-jussà et al., 2022). Additionally, Galician has been featured in multilingual benchmarks test sets like Tatoeba (Tiedemann, 2020a) and Flores (Goyal et al., 2022), which assess the performance of machine translation systems across a wide range of languages. These initiatives are critical in promoting the inclusion of Galician in global language technologies and ensuring its presence in multilingual environments.

More recently, initiatives such as *O Proxecto Nós* (The Nos Project) (de Dios-Flores et al., 2022) have advanced the field by providing openly licensed resources, tools, and use cases for Galician, including MT corpora and models.

3 Methodology

In this section, we present the resources employed to carry out the experimental part: the test datasets (Section 3.1), the machine translation systems (Section 3.2) and the evaluation metrics (Section 3.3).[1]

3.1 Datasets Description

In this paper, we used both public 3.1.1 and custom 3.1.2 datasets for the three pairs of languages. For English—Galician and Spanish—Galician, datasets from the general, legal, and health domains were employed, whereas for Portuguese–Galician, only datasets from the general domain were used. See Table 6 in Appendix A for more details about datasets.[2]

The absence of specialised domain datasets for the Portuguese–Galician pair is primarily due to the limited resources available for this language pair. Portuguese–Galician data is predominantly found within multilingual datasets designed for multiple languages. This scarcity of resources, combined with the time constraints in generating specialised datasets, led to the evaluation of this pair being carried out at the generic level. Despite this limitation, we opted to include the Portuguese–Galician pair in this paper, given its relatively overlooked status in the fields of natural language processing and machine translation.

3.1.1 Public Datasets

General Domain

[1]The experiments have been carried out in an Intel Xeon Gold 6326 CPU with 64 cores and a NVIDIA L40S GPU with 46 GB of VRAM.

[2]It is worth noting that all datasets have a single available reference, with the exception of Nós_MT_Gold, which has two. The evaluation was carried out using all available resources. Having additional reference datasets would enable a more robust and comprehensive analysis, which we leave for future work.

Nós_MT_Gold English–Galician and Spanish–Galician test sets consisting of curated sentences divided into two subsets: *Nós_MT_Gold_1*[3] and *Nós_MT_Gold_2.*[4] The distinction lies in the linguistic characteristics of the Galician part: the first subset features Galician sentences closer to Spanish grammar, while the second subset aligns more with Portuguese features.

Flores200[5] is a multilingual machine translation benchmark of 3000 English sentences professionally translated into 200 languages. Originally curated from Wikipedia, the dataset includes two subsets (devtest and dev), both used for all language pairs in our paper (Costa-jussà et al., 2022).

Tatoeba[6] is a multilingual MT benchmark across a wide range of language pairs, including low-resource combinations. Derived from the Tatoeba Project, it provides a community-driven database of example sentences in many languages (Tiedemann, 2020b). We used version v20230412 for the three language pairs.

Ntrex–128[7] (News Test References for MT Evaluation) is a multilingual dataset of English source sentences professionally translated into 128 target languages with document-level information (Federmann, Kocmi, and Xin, 2022). It originates from 'newstest2019' released for WMT19 (Barrault et al., 2019).

Test Suite English–Galician[8] and Spanish–Galician[9] test suites classified according to challenging linguistic phenomena.

Health Domain

Covid19-Health-Wikipedia[10] is an English–Galician corpus acquired from Wikipedia focused on health and the Covid-19 domain.

Legal Domain

TaCon[11] is a multilingual dataset from the legal domain, including translations of the Spanish Constitution into Basque, Catalan, Galician, and English (de Gibert Bonet et al., 2022).

LEGA[12] is a Galician–Spanish legal parallel corpus from SLI CLUVI.[13]

3.1.2 Custom Dataset

GalicianHealthMT[14] We constructed a custom Spanish–Galician test set by randomly selecting 1000 sentences from the Spanish Biomedical Crawled Corpus (Carrino et al., 2021).[15] After a thorough cleaning process, 959 sentences remained, which were manually translated into Galician.

3.2 Model Architectures

We evaluated three types of machine translation systems: RBMT (3.2.1), sequence-to-sequence (3.2.2), and LLMs (3.2.3). See Table 7 in Appendix A for models details.

For the biggest multilingual seq-to-seq models execution, we used the Easy-Translate framework (García-Ferrero, Agerri, and Rigau, 2022),[16] and for LLMs, the vLLM library (Kwon et al., 2023).[17] In the case of LLMs, different prompt formulations were employed depending on the nature of the model. For LLMs specifically fine-tuned or instructed for machine translation, we used the official prompts provided by the model developers, see Listings 1, 2 and 3. In contrast, for instructed LLMs, we designed a single custom prompt tailored to the machine translation task, see Listing 4 in Appendix subsection A.3. All prompt templates used in the experiments are included in the appendices to ensure reproducibility.

3.2.1 RBMT System

Apertium is an open-source RBMT system designed for closely related languages (Forcada

[3]Source: https://zenodo.org/records/7658009#.Y_O0x9LMJ3k, https://zenodo.org/records/7657887#.Y_OvX9LMJ3k

[4]Source: https://zenodo.org/records/7658033#.Y_O2o9LMJ3k, https://zenodo.org/records/7657993#.Y_Ozr9LMJ3k

[5]Source: https://github.com/facebookresearch/flores/blob/main/flores200/README.md

[6]Source: https://opus.nlpl.eu/Tatoeba/es&gl/v2023-04-12/Tatoeba

[7]Source: https://github.com/MicrosoftTranslator/NTREX

[8]Source: https://zenodo.org/records/7658249#.Y_O6bdLMJ3k

[9]Source: https://zenodo.org/records/7658052#.Y_O4fNLMJ3k

[10]Source: https://live.european-language-grid.eu/catalogue/corpus/3538/overview/

[11]Source: https://live.european-language-grid.eu/catalogue/corpus/19785

[12]Source: https://live.european-language-grid.eu/catalogue/corpus/12187/download/

[13]https://ilg.usc.gal/cluvi/

[14]This test dataset is publicly available at Zenodo: https://zenodo.org/records/15510935

[15]https://zenodo.org/record/5510033##.ZA5i_BzMJH5

[16]https://github.com/ikergarcia1996/Easy-Translate

[17]https://github.com/vllm-project/vllm

et al., 2011).[18]

The system used to conduct the experiments in this paper is the same one that powers *Gaio*.[19]

The system specifications for each language pair are deployed in Table 1.

3.2.2 Sequence-to-Sequence

Multilingual models

mBART is the multilingual extension of BART covering 50 languages (Tang et al., 2020). In this paper, we used the *mbart-50-many-to-many* model.[20]

M2M-100 is a multilingual model enabling direct translation between 100 languages. We employed the 418M,[21] 1.2B,[22] and 12B variants[23] (Fan et al., 2021).

NLLB (No Language Left Behind) is a multilingual model for 200 languages, optimised for low-resource scenarios. We used all its available sizes: 600M,[24] 1.3B,[25] 3.3B,[26] and 54B[27] (Costa-jussà et al., 2022).

Madlad400 was trained on 450 languages with up to 250B tokens (Kudugunta et al., 2023). We used the *madlad-3b*,[28] *madlad-7b*,[29] and *madlad-10b* models.[30]

Small100 is a distilled version of M2M-100 tailored for efficiency (Mohammadshahi et al., 2022).[31]

Nós_MT-OpenNMT-multilingual is a transformer model trained on the languages of Spain and English (Outeirinho et al., 2024).[32]

Bilingual models

PlanTL-es-gl is the official MT system for Spanish and the co-official languages in Spain.[33]

OpusMT consists of Marian-based models trained on OPUS data (Tiedemann and Thottingal, 2020).[34]

Nós_MT-OpenNMT-en-gl/es-gl are bilingual MT Galician models trained using both authentic and synthetic data (Outeirinho et al., 2024).[35]

Translate-en-gl-v1.0-hplt_opus is a bilingual English–Galician model trained on OPUS and HPLT data.[36]

3.2.3 Large Language Models

Instructed for MT

Plume refers to a family of 2B models trained on parallel Catalan-centric data for machine translation,[37] available with vocabulary sizes of 32k,[38] 128k,[39] and 256k[40] (Gilabert et al., 2024).

SalamandraTA-2B is a multilingual model trained on 70B tokens across 30 languages, including Galician (Gonzalez-Agirre et al., 2025).[41]

LLaMAX is a LLaMA-based multilingual model series supporting over 100 languages

[18] https://github.com/apertium

[19] Source: https://tradutorgaio.xunta.gal/TradutorPublico/traducir/index. *Gaio* is the open-source machine translation platform provided by the *Xunta de Galicia* (Galician Government). The system was developed and is maintained by the company **imaxin**.

[20] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[21] https://huggingface.co/facebook/m2m100_418M

[22] https://huggingface.co/facebook/m2m100_1.2B

[23] https://huggingface.co/facebook/m2m100-12B-last-ckpt

[24] https://huggingface.co/facebook/nllb-200-distilled-600M

[25] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[26] https://huggingface.co/facebook/nllb-200-distilled-3.3B

[27] https://huggingface.co/facebook/nllb-moe-54b

[28] https://huggingface.co/google/madlad400-3b-mt

[29] https://huggingface.co/google/madlad400-7b-mt

[30] https://huggingface.co/google/madlad400-10b-mt

[31] https://huggingface.co/alirezamsh/small100

[32] https://huggingface.co/proxectonos/Nos_MT-OpenNMT-multilingual

[33] https://administracionelectronica.gob.es/ctt/verPestanaGeneral.htm?idIniciativa=plata

[34] https://huggingface.co/Helsinki-NLP, https://huggingface.co/Helsinki-NLP/opus-mt-en-ROMANCE

[35] https://huggingface.co/proxectonos/Nos_MT-OpenNMT-en-gl, https://huggingface.co/proxectonos/Nos_MT-OpenNMT-es-gl

[36] https://huggingface.co/HPLT/translate-en-gl-v1.0-hplt_opus

[37] The languages included in the training process were: Spanish, French, Italian, Portuguese, Galician, German, English, and Basque.

[38] https://huggingface.co/projecte-aina/Plume32k

[39] https://huggingface.co/projecte-aina/Plume128k

[40] https://huggingface.co/projecte-aina/Plume256k

[41] https://huggingface.co/BSC-LT/salamandraTA-2B

Pair	Monolingual Entries		Bilingual Entries		Transfer Rules
	Source	Target	All domains	Health domain	src → tgt / tgt → src
English–Galician	23814	11413	29185	–	364 / 117
Spanish–Galician	113775	76740	67432	62131	128 / 109
Portuguese–Galician	20536		21687	–	213 / 150

Table 1: Apertium system specifications per language pair. For the Spanish–Galician pair, a specialised bilingual dictionary was employed for the Health domain. The Spanish—Galician and Portuguese—Galician pairs each use a single transfer rules file per direction, while the English—Galician pair uses three transfer rules files for each direction.

(Lu et al., 2024).[42]

Non-instructed for MT

EuroLLM is a family of open multilingual LLMs designed for EU languages (Martins et al., 2024).[43]

Salamandra-7B-Instruct is a 7B parameter model instruction-tuned for multilingual tasks (Gonzalez-Agirre et al., 2025).[44]

Llama-3.1-Carballo and **Llama-3.1-8B-Instruct-Galician** are LLaMA-based models tuned specifically for Galician (Gamallo et al., 2024; Bao, Pérez, and Parapar, 2024).[45]

DeepSeek-R1-Distill-Llama-70B is a distilled LLaMA model optimised for reasoning tasks (DeepSeek-AI, 2025).[46]

3.3 Evaluation Metrics

We assessed translation quality using lexical-based and embedding-based metrics, following (Lee et al., 2023). Lexical metrics—BLEU (Papineni et al., 2002), chrF (Popović, 2015), and TER (Snover et al., 2006)—measure surface-level overlap, while embedding-based metrics as COMET (Rei et al., 2022) evaluate semantic similarity using pre-trained models.

We used SacreBLEU[47] to compute BLEU, chrF, and TER with standardised preprocessing (Post, 2018). Additionally, we used COMET (wmt22-comet-da)[48] for reference-based, semantic evaluation.

4 Experiments

This section outlines the various aspects considered in the MT evaluation process. First, it explains how the models evaluation was conducted (Section 4.1), then describes how variability across different datasets was measured (Section 4.2), and finally discusses how computational performance was taken into account (Section 4.3).

4.1 Models Evaluation

The evaluation of machine translation models typically involves the utilisation of a singular dataset or benchmark, such as Flores in a generalist domain or TaCon for a specific domain. However, as previously mentioned by (Xiang et al., 2022), this can result in a biased view of the performance of the models and potentially lead to the variability of their performance going unnoticed. This oversight can result in the failure to recognise the potential inconsistency and lack of robustness of MT models.

To address this limitation, a comprehensive evaluation of 33, 32, and 30 models described in section 3.2, has been conducted to identify the most suitable translation model for the English–Galician, Spanish–Galician, and Portuguese–Galician pairs, respectively. Finally, the average performance of each model across the different datasets was calculated for the four metrics.

4.2 Models Variability

Following the acquisition of the results and means for each model, the variability of the models in the evaluation process should be

[42] https://huggingface.co/LLaMAX/LLaMAX3-8B-Alpaca, https://huggingface.co/LLaMAX/LLaMAX2-7B-Alpaca

[43] https://huggingface.co/utter-project/EuroLLM-1.7B-Instruct, https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[44] https://huggingface.co/BSC-LT/salamandra-7b-instruct

[45] https://huggingface.co/proxectonos/Llama-3.1-Carballo, https://huggingface.co/irlab-udc/Llama-3.1-8B-Instruct-Galician

[46] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B

[47] https://github.com/mjpost/sacrebleu

[48] https://github.com/Unbabel/COMET/blob/master/MODELS.md

considered. To this end, the coefficient of variation (CV) was selected as a measure of dispersion, in contrast to the disagreement variant utilised by (Xiang et al., 2022). The CV, being a dimensionless measure that relates the standard deviation to the arithmetic mean and is expressed as a percentage, allows the relative dispersion between datasets with different scales or units to be compared, thereby eliminating possible distortions arising from the absolute magnitudes of the means. This approach offers an objective and consistent method for assessing model consistency and variability. In contrast to the error number, the CV furnishes a relative perspective on variability, which can be more informative when comparing results between different models, datasets and pairs.

This evaluation has been carried out with BLEU and TER. The selection of these metrics is predicated on their capacity to encompass diverse evaluation typologies and their distinctive characteristics. BLEU is a metric that compares n-grams in machine translations with reference tests. Conversely, TER quantifies the necessity for edits to transform a machine translation into the reference, thereby offering a complementary perspective that focuses on errors and correctness.

Although chrF and COMET were also employed in the evaluation process explained in 4.1, chrF shares similarities with BLEU as a lexical-based metric, while COMET, being a deep learning-based metric, tends to show little variation between results, which makes it less suitable for capturing significant differences in model variability. Therefore, the two metrics that best facilitated the visualization of variability were selected.[49]

4.3 Computational Performance

A further salient aspect that is seldom considered during MT evaluation pertains to their size and computational cost. In the context of SMEs both size and cost become relevant when evaluating and selecting a model.

For neural-based models, the size of each model was obtained from the configuration files available on Hugging Face, and, for those not explicitly quantified, estimated based on their suitability for execution in our experimental environment. The size of Apertium was calculated by measuring the storage space occupied by each folder.[50]

5 Results

The results are organised into four tables. Tables 2 and 3 report the overall performance of the models based on the mean of BLEU and TER metrics, for translations into Galician (Table 2) and from Galician (Table 3). Each table includes the minimum, mean, and maximum scores across models for each language pair and direction, along with the average coefficient of variation (CV). Additionally, Spearman's correlation coefficient (ρ) is employed to examine the relationship between translation quality (mean BLEU/TER scores) and performance variability. This statistical measure evaluates the strength and direction of a monotonic association between two variables, without requiring assumptions of linearity or normality. Furthermore, the associated p-value is computed to determine the statistical significance of the observed correlations. Tables 4 and 5 present the best-performing translation models according to evaluation metrics (BLEU and TER), variability and model size, into Galician (Table 4) and from Galician (Table 5).

BLEU						
Pair	**Min.**	**Mean**	**Max.**	**CV**	ρ	p**-value**
en-gl	27.83	33.51	43.01	16.08	-0.38	0.02
es-gl	20.87	50.95	74.86	41.07	0.65	2.29×10^{-5}
pt-gl	26.54	33.41	51.92	36.90	-0.01	0.95
TER						
Pair	**Min.**	**Mean**	**Max.**	**CV**	ρ	p**-value**
en-gl	39.10	54.77	72.62	15.85	-0.10	0.56
es-gl	14.69	38.09	73.82	60.34	-0.87	1.26×10^{-11}
pt-gl	34.42	52.66	60.75	23.86	-0.46	0.0095

Table 2: Minimum, Mean, Maximum, Correlation Variability (CV), and Spearman Correlation (ρ) Between Mean and CV for BLEU and TER Across English–Galician, Spanish-Galician, and Portuguese–Galician models.

[49]Finally, due to constraints on length, only the mean of models results for BLEU and TER metrics have been presented in this paper for model evaluation. Nevertheless, the results obtained by each of the models on each evaluation test for the four metrics (BLEU, chrF, TER, and COMET) can be found in the following Zenodo's page: https://zenodo.org/records/15547513

[50]The Apertium version used was 3.7.2, which consists of three packages (folders) per language pair: two monolingual and one bilingual package.

BLEU						
Pair	**Min.**	**Mean**	**Max.**	**CV**	ρ	p**-value**
gl-en	30.6	39.35	54.09	26.43	-0.58	0.0008
gl-es	23.26	54.88	77.71	39.02	0.62	0.0001
gl-pt	26.62	35.27	55.73	39.06	-0.23	0.21
TER						
Pair	**Min.**	**Mean**	**Max.**	**CV**	ρ	p**-value**
gl-en	30.67	100.69	169.87	24.90	0.012	0.94
gl-es	12.58	34.27	67.56	63.21	-0.78	7.7×10^{-8}
gl-pt	31.80	51.78	60.84	26.34	-0.37	0.04

Table 3: Minimum, Mean, Maximum, Correlation Variability (CV), and Spearman Correlation (ρ) Between Mean and CV for BLEU and TER Across Galician–English, Galician-Spanish, and Galician–Portuguese models.

6 Discussion

In the following subsections, the results are analysed in terms of model variability by translation pair (Section 6.1), model performance for each pair and direction (Section 6.2), and finally, the Portuguese–Galician results in particular (Section 6.3).[51]

6.1 Variability Analysis

As shown in Tables 2 and 3, some correlations can be observed between the results and the Variability Coefficient depending on the translation direction and language pair. For Spanish–Galician, a significant relationship emerges: the better the model's performance, the greater the dispersion in both metrics, regardless of direction. In contrast, for English–Galician, a weaker and inverse relationship is observed for the BLEU metric, where improved results are associated with lower variability. However, Spearman's coefficient indicates no clear relationship between the mean and variability in this pair for TER metric. Particularly noteworthy in this context are the high TER scores in the Galician–English direction, in some cases even exceeding a score of 100. This outcome is largely due to the poor performance of certain LLMs, such as EuroLLM-1.7B-Instruct, which in this direction failed to translate several sentences across several evaluation datasets. These isolated cases significantly contributed to the overall degradation of TER scores. Future work will focus on a more detailed analysis of the issues identified in these models.

Lastly, for Portuguese–Galician, a weak correlation is evident only in the TER metric, where higher model performance corresponds to lower variability, mirroring the behavior observed in English–Galician.

The results obtained for Spanish–Galician, which show greater consistency in both directions compared to the other pairs when analysing the relationship between performance and variability, suggest that higher average quality (approximately 20 points in both metrics relative to the other language pairs) also leads to increased variability across datasets within each model. This may occur because better-performing models tend to achieve higher maximum scores, thereby widening the gap between minimum and maximum and increasing overall variability. Conversely, in pairs with lower overall performance, the difference between minimum and maximum values is less pronounced, resulting in a weaker statistical relationship. In such cases, better-performing models exhibit greater consistency, as the range of scores remains narrower.

Moreover, while a direct correlation between performance and variability is not consistently observed in the other pairs, variability remains significant. These findings, as noted in (Xiang et al., 2022) paper, suggest that the results are heavily influenced by the evaluation dataset, especially in cases where translation models achieve favorable results and exhibit high proficiency. The selection of the evaluation corpus in machine translation is a critical factor that must not be treated as trivial. Relying on a single dataset, regardless of the domain or language pair being evaluated, can lead to biased or unrepresentative conclusions. A well-rounded evaluation requires diverse corpora that reflect the linguistic, cultural, and contextual challenges inherent in real-world translation tasks.

Especially for the Spanish–Galician language pair, further research is needed to clarify where this significant variability is occurring and why. It is essential to determine whether this variability stems from the poor performance of certain models overall or from the models' low performance on specific datasets. This is of particular importance in determining whether an average of higher score is sufficient to select a satisfactory model, or whether the dispersion presented in the evaluations may act as a disadvantage or an aspect to be considered during the implementation stage. To facilitate the visualisation of this variabil-

[51] Owing to space constraints, a qualitative analysis of the results presented in this paper is deferred to future work.

Pair	Model	Avg BLEU	Avg TER	CV BLEU	CV TER	Prec.	Size
en-gl	madlad-7b	**43.41**	**42.72**	12.88	13.36	float32	38.90
	EuroLLM-9B	40.16	47.29	**8.65**	10.72	bfloat16	18.30
	Apertium	19.33	65.30	22.03	**8.15**	float32	**0.027**
es-gl	Nos-es-gl	**59.90**	30.87	46.54	80.52	fp16	0.50
	Apertium	58.99	**30.29**	44.28	80.25	float32	**0.097**
	salamandra-7b	47.70	42.42	**31.74**	46.14	bfloat16	15.54
	nllb-600M	44.12	40.60	32.26	**39.37**	float32	2.46
pt-gl	nllb-moe-54b	**39.55**	**46.28**	36.20	29.52	int8	51.74
	m2m100-12B-last	31.73	55.18	**13.02**	**5.90**	int8	5.99
	Apertium	32.13	53.28	53.53	31.76	float32	**0.072**

Table 4: Average BLEU and TER Scores, Coefficient of Variation, Inference Precision, and Model Size for English–Galician (en-gl), Spanish–Galician (es-gl) and Portuguese–Galician (pt-gl) models.

Pair	Model	Avg BLEU	Avg TER	CV BLEU	CV TER	Prec.	Size
gl-en	madlad-7b	**50.31**	**37.11**	16.36	21.41	float32	37.90
	nllb-54b	50.21	38.28	**15.32**	20.85	int8	51.74
	mbart	30.87	56.98	16.74	**10.52**	float32	2.44
	Apertium	17.03	72.48	19.84	12.48	float32	**0.027**
gl-es	Nos-gl-es	**63.75**	27.52	43.60	86.11	fp16	0.50
	madlad-7b	63.60	**27.22**	41.69	84.75	float32	38.90
	madlad-10b	63.50	**27.22**	41.73	84.38	int8	9.49
	salamandra-7b	56.40	32.43	**30.42**	50.00	bfloat16	15.54
	LLaMAX2	39.69	50.85	35.19	**26.42**	bfloat16	26.96
	Apertium	59.2	29.97	42.11	75.36	float32	**0.097**
gl-pt	nllb-54b	**41.60**	**45.88**	37.80	31.05	int8	51.74
	m2m100-12B-avg10	35.08	52.30	**24.57**	15.13	int8	5.99
	mbart	20.25	65.40	38.43	**14.63**	int8	2.44
	Apertium	31.90	54.05	54.61	32.02	float32	**0.072**

Table 5: Average BLEU and TER Scores, Coefficient of Variation, Inference Precision, and Model Size for Galician–English (gl-en), Galician–Spanish (gl-es) and Galician–Portuguese (gl-pt) models.

ity, refer to Figure 1, which presents a box plot illustrating the mean, median and dispersion for each model in the Spanish–Galician direction and BLEU metric.[52]

6.2 MT Systems Analysis

Given that no single model has consistently demonstrated superior performance across all datasets, metrics and parameters, tables 4 and 5 present the models that have exhibited the highest performance in at least one of the averages identified, namely mean BLEU, TER, CV BLEU, CV TER and size. With the exception of BLEU, it is understood that lower values are preferable for the remaining metrics.

The tables offer insights into various phenomena. Firstly, in both the Spanish–Galician and Galician–Spanish pairs, both Apertium and the bilingual sequence-to-sequence model created by the Nos Project demonstrate competitiveness in terms of results and size when compared to larger multilingual models and LLMs. The present findings are consistent with those reported in other studies (García, 2024; González and Claramunt, 2024) that indicate that, contingent on the prevailing context, models characterised by low computational cost can indeed yield optimal performance. This prompts the question of whether the expense associated with a more substantial model is justifiable in terms of the results achieved, or whether the enhancement of these

[52]For the graphs of the other translation pairs, directions and TER metric, please refer to the Appendix B.

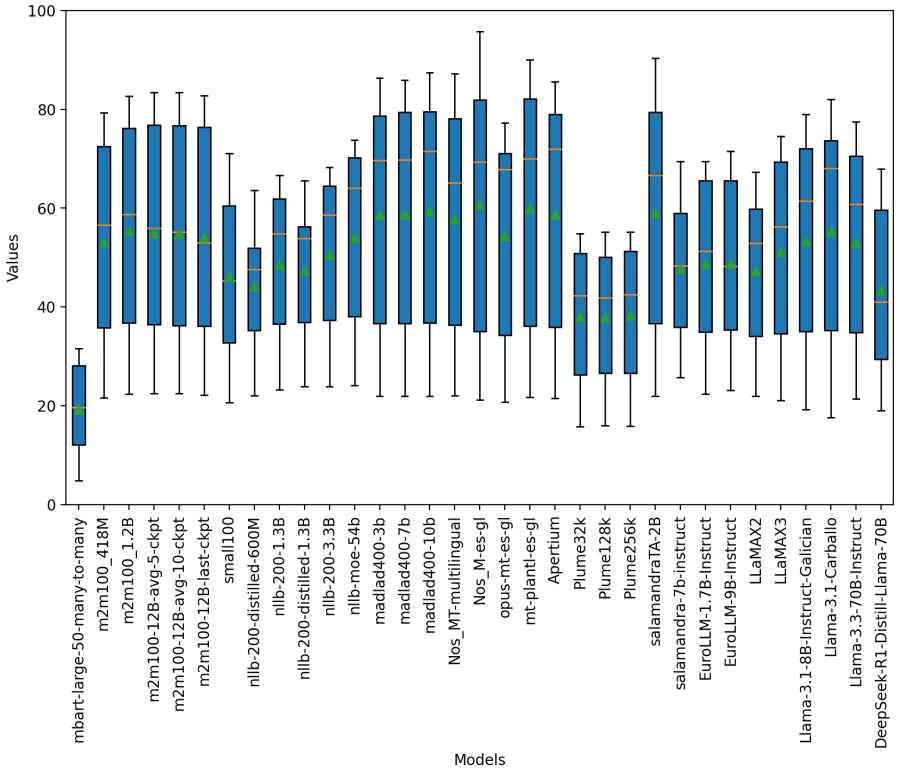

Figure 1: Box Plot of Spanish—Galician BLEU Scores for All Evaluated Models that shows the variability across test datasets for each model.

efficient models would prove to be a sufficient substitute.

Conversely, in a more distant language pair, this reality undergoes a transformation. In such cases, multilingual sequence-to-sequence models have been shown to achieve the most optimal average results. However, it is noteworthy that multilingual models trained with a larger number of languages and fewer parameters have consistently outperformed larger models such as nllb-54b in terms of results. In this particular instance, and in both directions, the madlad models appear to demonstrate the most optimal performance in English–Galician pair.

Finally, another conclusion drawn from these results is that, as highlighted in the recent WMT24 findings, for low-resource language pairs, LLMs still fall short of the performance achieved by sequence-to-sequence models—including those specifically trained for machine translation, such as SalamandraTA or LlaMAX (Kocmi et al., 2024).

6.3 Portuguese–Galician Analysis

The results for the Portuguese–Galician pair are more noteworthy. These languages are closely related, with a common ancestor, *Galaico-Português*, from which they have evolved over six centuries as distinct languages while maintaining common characteristics. The orthography of Galician has evolved to become more similar to Spanish orthography since its standardisation in the 1980s. Recent studies have indicated that this has led to Galician's increased proximity to Spanish compared to Portuguese. Nevertheless, these same studies demonstrate that a spelling change in Galician, such as the change of the spelling *ñ* for *nh*, or *ll* for *lh*, brings this written language closer to Portuguese than to Spanish (Campos et al., 2020). Consequently, the substandard results observed in the experiments conducted during this study are noteworthy, as the mean values approximate those attained in English–Galician more closely than in Spanish–Galician, as would be anticipated.

The findings can be attributed to a number of factors. The paucity of research conducted on this particular language pair in the area of machine translation, coupled with the dearth of available resources, is a salient one. The datasets and models employed in this paper are generic, multilingual benchmarks that do not focus on the causatives of this

particular language pair, in contrast to the evaluations of the other two pairs analysed in this paper. Galician, on the other hand, is more closely related to the European variant of Portuguese than to its Brazilian variant. As noted in other studies, the Brazilian variant is the predominant one in open-source resources for Portuguese (Sanches, Ribeiro, and Coheur, 2024). Consequently, the creation of specific assessment resources that facilitate an efficient and focused evaluation of this language pair is imperative for future research. Finally, these results are similar to those obtained when trying to translate variants of Brazilian and European Portuguese, where machine translation models seem to have issues when translating two variants of the same language (Costa-jussà, Zampieri, and Pal, 2018; Sanches, Ribeiro, and Coheur, 2024). This is a research area that will be addressed in the future.

Currently, NLLB-54B—a large-scale, computationally demanding multilingual model—consistently delivers the best performance for this pair across both translation directions.

7 Conclusions & Future Work

This study aims to establish the most effective MT models for English–Galician, Spanish–Galician, and Portuguese–Galician in the general domain. To achieve this, an evaluation of all open-source models and systems across the existing benchmarks has been conducted. The evaluation not only examines model quality but also accounts for variability across datasets, as well as the size and resource consumption of the models. These factors are often overlooked in traditional MT evaluations.

The results demonstrate variability depending on the translation pair. For Spanish–Galician, models with lower computational costs still achieve the best overall results, while for English–Galician and Portuguese–Galician, multilingual sequence-to-sequence models outperform the others. In all cases, LLMs fall behind in terms of quality.

The contributions of this study are as follows:

1. An analysis of the Spanish–Galician, English–Galician, and Portuguese–Galician pairs in both translation directions, considering additional factors beyond traditional metrics, such as variability, model size, and resource consumption.

2. The first quantitative analysis of the Portuguese–Galician pair in machine translation.

3. A comparison of various systems (RBMT, sequence-to-sequence, and LLMs) for three language pairs, including one involving a low-resource language.

4. A Spanish–Galician evaluation dataset in the health domain.

Future research will address variability across different datasets, aiming to create more effective and robust evaluation datasets for all three language pairs. This effort should involve both domain-specific and general datasets. Additionally, further work is required on the Portuguese–Galician pair, with the goal of developing a more efficient translation system that surpasses the current models.

Lastly, future studies will integrate a qualitative analysis to provide a deeper understanding of the numerical results, offering more clarity and insight into model performance.

References

Artetxe, M. and H. Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Bao, E., A. Pérez, and J. Parapar. 2024. Adapting Large Language Models for Underrepresented Languages. In *VII Congreso XoveTIC: impulsando el talento científico*. Universidade da Coruña, Servizo de Publicacións.

Barrault, L., O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Campos, J., P. Gamallo, I. Alegria, and M. Neves. 2020. A Methodology to Measure the Diachronic Language Distance between Three Languages Based on Perplexity. *Journal of Quantitative Linguistics*, 28:1–31, 03.

Carrino, C. P., J. Armengol-Estapé, O. d. G. Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, and M. Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for Spanish biomedical language models. *arXiv preprint arXiv:2109.07765*.

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Costa-jussà, M. R., M. Zampieri, and S. Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

de Dios-Flores, I., C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. García, P. Gamallo, E. Fernández Rei, A. Bugarín-Diz, M. González González, S. Barro, and X. L. Regueira. 2022. The nós project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, Marseille, France, June. European Language Resources Association.

de Gibert Bonet, O., K. Kharitonova, B. Calvo Figueras, J. Armengol-Estapé, and M. Melero. 2022. Quality versus quantity: Building Catalan-English MT resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69, Marseille, France, June. European Language Resources Association.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Fan, A., S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Federmann, C., T. Kocmi, and Y. Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, November. Association for Computational Linguistics.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Gamallo, P., P. Rodríguez, I. de Dios-Flores, S. Sotelo, S. Paniagua, D. Bardanca, J. R. Pichel, and M. Garcia. 2024. Open generative large language models for galician.

García, S. 2024. Enhaced apertium system: Translation into low-resource languages of Spain Spanish–Asturian. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 878–884, Miami, Florida, USA, November. Association for Computational Linguistics.

García-Ferrero, I., R. Agerri, and G. Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

García-Mateo, C. and M. Arza. 2012. *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9. Georg Rehm and Hans Uszkoreit (series editors).

Gilabert, J. G., C. Escolano, A. S. Savall, F. D. L. Fornaciari, A. Mash, X. Liao, and M. Melero. 2024. Investigating the translation capabilities of large language models trained on parallel data only.

González, S. G. and G. R. Claramunt. 2024. Study of the State of the Art Galician Machine Translation: English-Galician and Spanish-Galician models. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of*

Portuguese - Vol. 1, pages 411–421, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Gonzalez-Agirre, A., M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, I. Pikabea, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, O. Pareras, V. Ruiz-Fernández, and M. Villegas. 2025. Salamandra Technical Report.

Goyal, N., C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kocmi, T., E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, B. Marie, C. Monz, K. Murray, M. Nagata, M. Popel, M. Popović, M. Shmatova, S. Steingrímsson, and V. Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November. Association for Computational Linguistics.

Kudugunta, S., I. Caswell, B. Zhang, X. Garcia, C. A. Choquette-Choo, K. Lee, D. Xin, A. Kusupati, R. Stella, A. Bapna, and O. Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset.

Kwon, W., Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lample, G., A. Conneau, L. Denoyer, and M. Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lee, S., J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Lu, Y., W. Zhu, L. Li, Y. Qiao, and F. Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages.

Martins, P. H., P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, and A. F. T. Martins. 2024. Eurollm: Multilingual language models for europe.

Mohammadshahi, A., V. Nikoulina, A. Berard, C. Brun, J. Henderson, and L. Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

OpenAI, J. Achiam, S. Adler, et al. 2024. GPT-4 Technical Report.

Outeirinho, D. B., P. G. Otero, I. de Dios-Flores, and J. R. P. Campos. 2024. Exploring the effects of vocabulary size in neural machine translation: Galician as a target language. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 600–604, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Papineni, K., S. Roukos, T. Ward, and W.-J. 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.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, M. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation:*

Research Papers, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Rei, R., J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. 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, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Sanches, J., R. Ribeiro, and L. Coheur. 2024. From brazilian portuguese to european portuguese.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks.

Sánchez, J. M. R. and C. G. Mateo. 2022. Deliverable D1.15 Report on the Galician Language. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Tang, Y., C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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

Tiedemann, J. 2020b. The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT. *arXiv preprint arXiv:2010.06354*.

Tiedemann, J. and S. Thottingal. 2020. OPUS-MT–Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. 2023. Llama: Open and efficient foundation language models.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2023. Attention is all you need.

Xiang, J., H. Li, Y. Liu, L. Liu, G. Huang, D. Lian, and S. Shi. 2022. Investigating data variance in evaluations of automatic machine translation metrics. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 150–157, Dublin, Ireland, May. Association for Computational Linguistics.

Zoph, B., D. Yuret, J. May, and K. Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November. Association for Computational Linguistics.

A Methodology Details

In Section A.1, Table 6 provides the details of the datasets' features. In Section A.2, Table 7 presents the models' sizes, precision, and VRAM requirements.

A.1 Datasets Details

A.2 Models Details

A.3 MT Prompts by Model Type

```
f"[{src_lang_code}] {text} \n[{
    ↪ tgt_lang_code}]"
```

Listing 1: SalamandraTA-2B prompt. Source: https://huggingface.co/BSC-LT/salamandraTA-2B

#	Dataset	Pairs	Sentences	Domain	Source
1	Nós_MT_Gold	English–Galician Spanish–Galician	1777 1998	General	Public
2	Test Suite	English–Galician Spanish–Galician	364 334	General	Public
3	Flores200 devtest	English–Galician Spanish–Galician Portuguese–Galician	1012	General	Public
4	Flores200 dev	English–Galician Spanish–Galician Portuguese–Galician	997	General	Public
5	Tatoeba	English–Galician Spanish–Galician Portuguese–Galician	1022 3133 437	General	Public
6	NTREX–128	English–Galician Spanish–Galician Portuguese–Galician	1997	General	Public
7	Covid19-Health-Wikipedia	English–Galician	957	Health	Public
8	TaCON	English–Galician Spanish–Galician	1100	Legal	Public
9	LEGA	Spanish–Galician	1100	Legal	Public
10	GalicianHealthMT	Spanish–Galician	959	Health	Custom

Table 6: Overview of the datasets used in this paper: Number of sentences and domain.

```
"<s> [{}] {} \n[{}]'.format(
    ↪ src_lang_code, text,
    ↪ tgt_lang_code)"
```

Listing 2: Plume LLM models prompt. Source: https://huggingface.co/projecte-aina/Plume32k

```
f"<|im_start|>system\n<|im_end|>\
    ↪ n<|im_start|>user\
    ↪ nTranslate the following {
    ↪ src_lang} source text to {
    ↪ tgt_lang}:\n{src_lang}: {
    ↪ text}\n{tgt_lang}: <|im_end
    ↪ |>\n<|im_start|>assistant\n
    ↪ "
```

Listing 3: Eurollm models prompt. Source: https://huggingface.co/projecte-aina/Plume32k

B Analysis Details

B.1 BLEU & TER box plots

Refer to Figures 2, 3, 4, 5, and 6 for the box plots of BLEU results across the different language pairs and directions. For the TER results, see Figures 7, 8, 9, 10, 11, and 12.

```
f"""Translate the following
    ↪ sentence from {src_lang} to
    ↪  {tgt_lang}.
        Respond with ONLY the
            ↪ translated sentence
            ↪ , without
            ↪ explanations or
            ↪ extra text.
        Do NOT leave the
            ↪ translation blank,
            ↪ even if uncertain,
            ↪ always provide a
            ↪ translation.

        {src_lang}: {text}
        {tgt_lang}:"""
```

Listing 4: Author designed MT prompt.

Model	Size (GB)	Precision	VRAM (GB)
mbart-large-50-many-to-many-mmt	2.44	float32	2.44
opus-mt-en-ROMANCE	312	float32	312
m2m100_418M	1.94	float32	1.94
m2m100_1.2B	4.96	float32	4.96
m2m100-12B-avg-5-ckpt	47.2	int4	5.99
m2m100-12B-avg-10-ckpt	47.2	int4	5.99
m2m100-12B-last-ckpt	47.2	int4	5.99
small100	1.33	float32	1.33
nllb-200-distilled-600M	2.46	float32	2.46
nllb-200-1.3B	5.48	float32	5.48
nllb-200-distilled-1.3B	5.48	float32	5.48
nllb-200-3.3B	24	float32	24
nllb-moe-54b	228.62	int8	51.74
madlad400-3b-mt	3.88	float32	3.88
madlad400-7b-mt	38.9	float32	38.9
madlad400-10b-mt	49.89	int8	9.49
Nos_MT-OpenNMT-multilingual	0.472	fp16	0.472
Nos_MT-OpenNMT-en-gl	0.497	fp16	0.497
Nos_MT-OpenNMT-gl-en	0.497	fp16	0.497
Nos_MT-OpenNMT-es-gl	0.497	fp16	0.497
Nos_MT-OpenNMT-gl-es	1.56	fp16	1.56
opus-mt-en-gl	0.293	float32	0.293
opus-mt-gl-en	0.223	float32	0.223
translate-en-gl-v1.0-hplt_opus	0.22	float32	0.22
opus-mt-es-gl	0.191	float32	0.191
opus-mt-gl-es	0.191	float32	0.191
opus-mt-pt-gl	0.191	float32	0.191
oput-mt-gl-pt	0.191	float32	0.191
Apertium-en-gl	0.027	float32	0.027
Apertium-es-gl	0.097	float32	0.097
Apertium-pt-gl	0.072	float32	0.072
plantl-es-gl	1.83	float32	1.83
Plume32k	8.19	float32	8.19
Plume128k	8.98	float32	8.98
Plume 256k	10.2	float32	10.2
salamandraTA-2B	4.51	bfloat16	4.51
salamandra-7B-instruct	15.54	bfloat16	15.54
EuroLLM-1.7B-Instruct	3.31	bfloat16	3.31
EuroLLM-9B-Instruct	18.3	bfloat16	18.3
LLaMAX2-7B-Alpaca	26.96	bfloat16	26.96
LLaMAX3-8B-Alpaca	32.12	bfloat16	32.12
Llama-3.1-8B-Instruct-Galician	15.04	bfloat16	15.04
Llama-3.1-Carballo	16.07	bfloat16	16.07
Llama-3.3-70B-Instruct	112.8	int4	14.1
DeepSeek-R1-Distill-Llama-70B	141.07	int4	17.63375

Table 7: Overview of Evaluated Models in all the translation pairs: Model Size, Inference Precision, and VRAM Required.

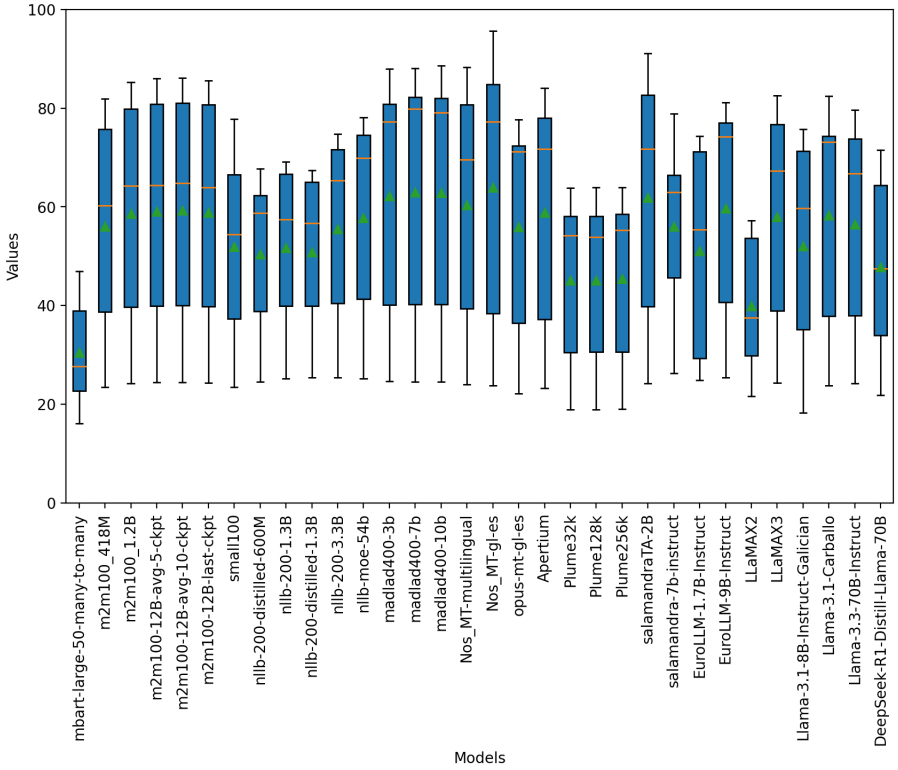

Figure 2: Box Plot of Galician–Spanish BLEU Scores for All Evaluated Models that shows the variability across test datasets for each model.

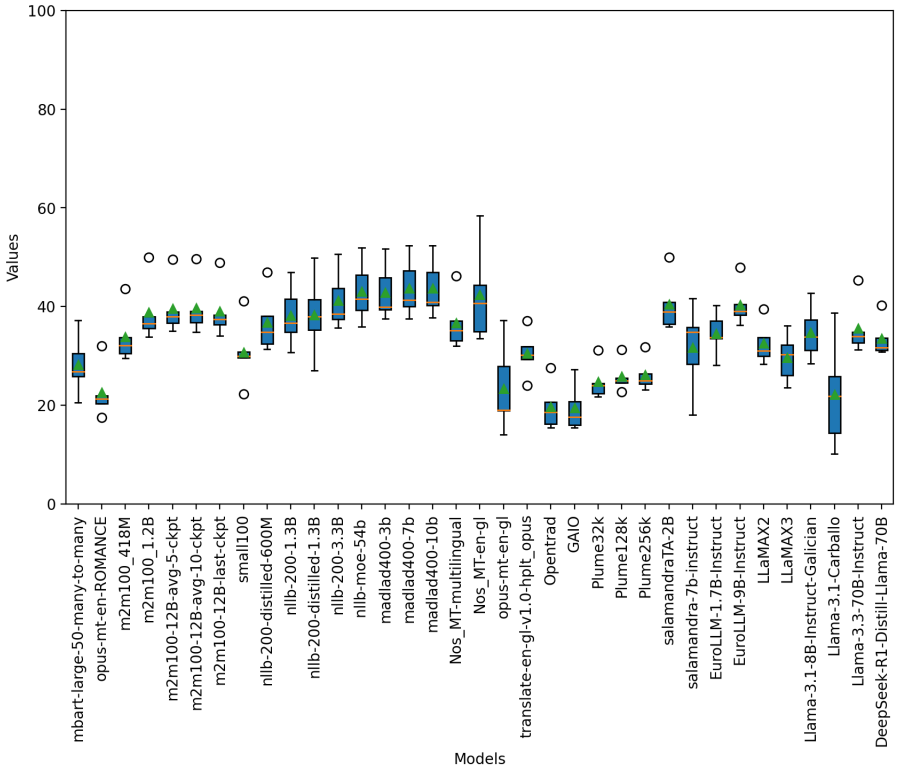

Figure 3: Box Plot of English—Galician BLEU Scores for All Evaluated Models that shows the variability across test datasets for each model.

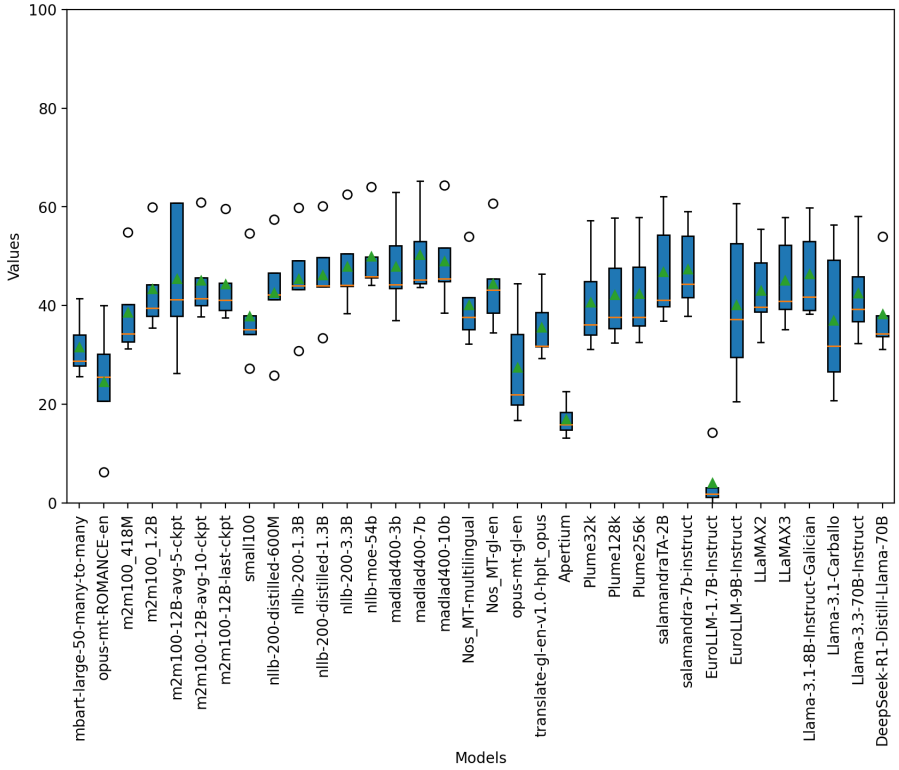

Figure 4: Box Plot of Galician–English BLEU Scores for All Evaluated Models that shows the variability across test datasets for each model.

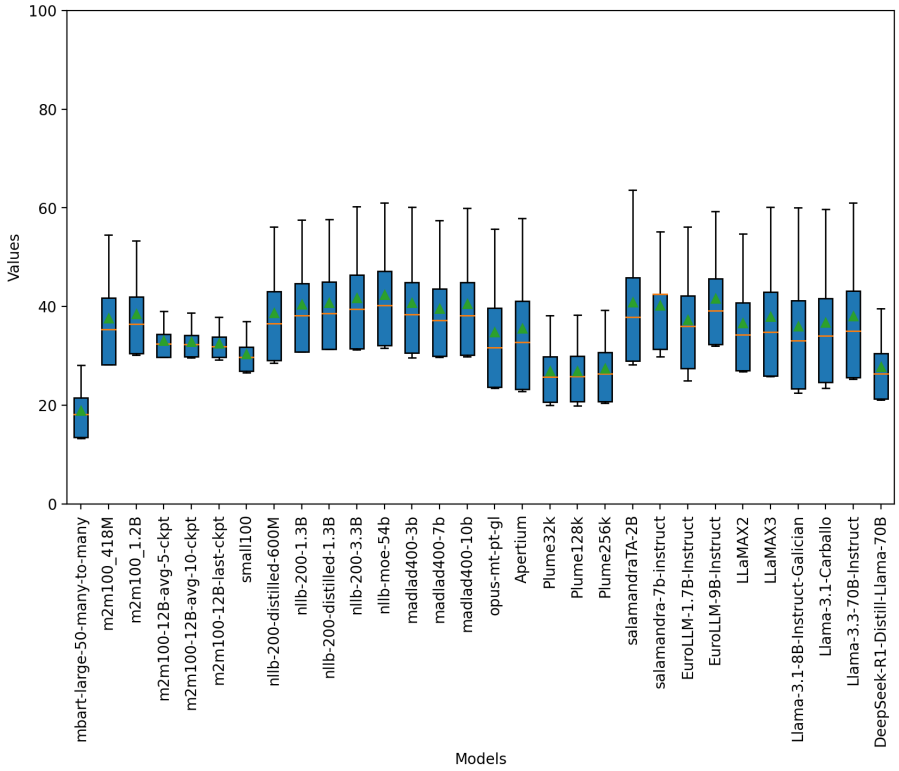

Figure 5: Box Plot of Portuguese—Galician BLEU Scores for All Evaluated Models that shows the variability across test datasets for each model.

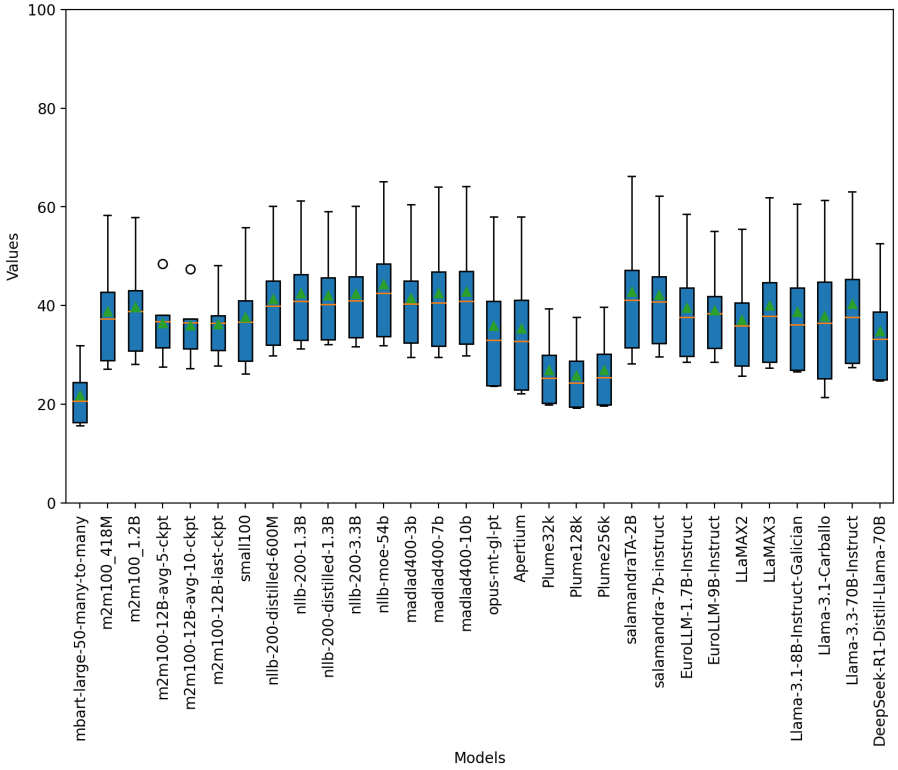

Figure 6: Box Plot of Galician–Portuguese BLEU Scores for All Evaluated Models that shows the variability across test datasets for each model.

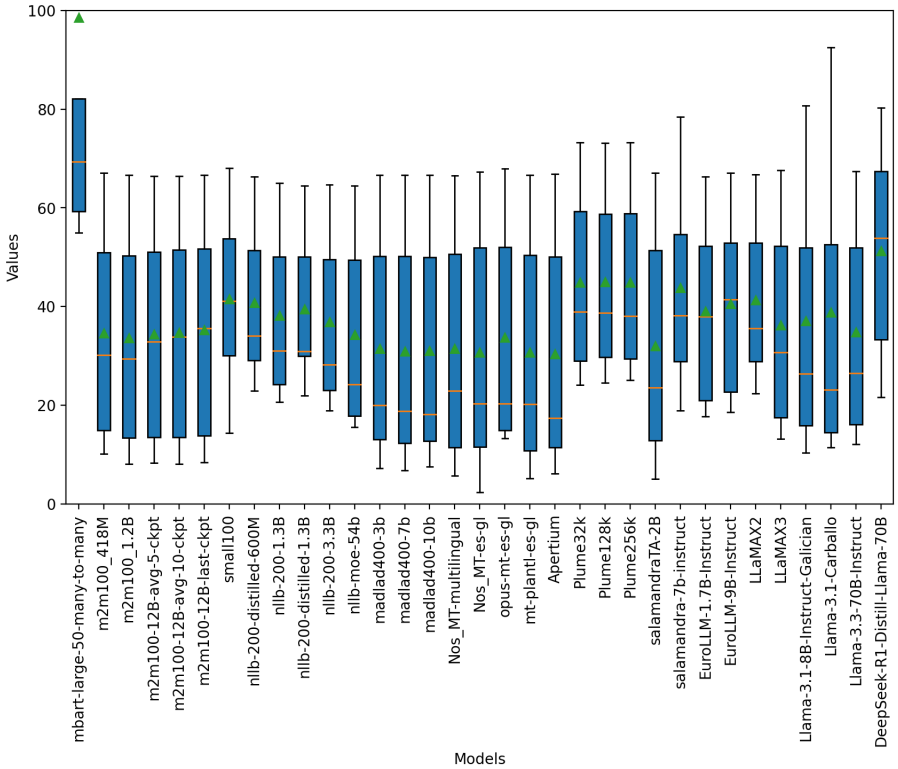

Figure 7: Box Plot of Spanish–Galician TER Scores for All Evaluated Models that shows the variability across test datasets for each model.

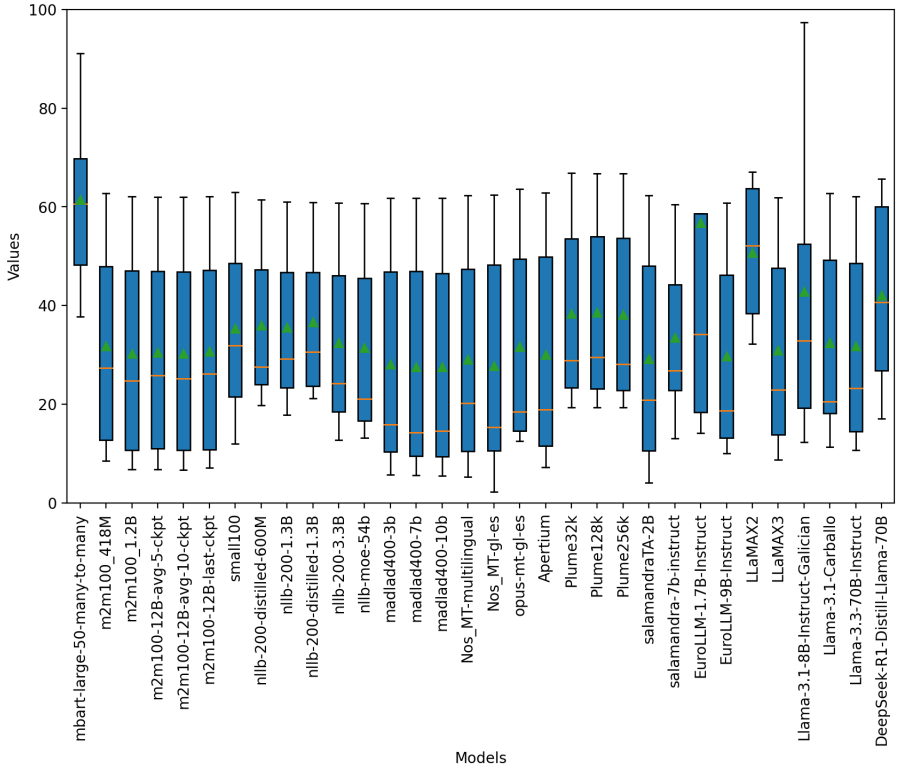

Figure 8: Box Plot of Galician–Spanish TER Scores for All Evaluated Models that shows the variability across test datasets for each model.

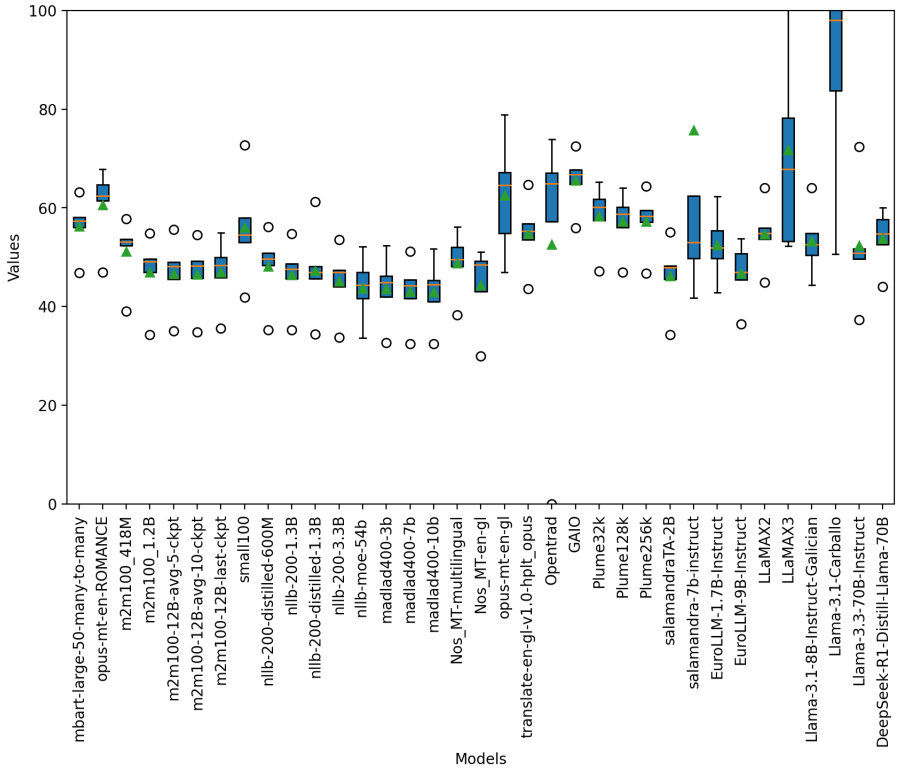

Figure 9: Box Plot of English—Galician TER Scores for All Evaluated Models that shows the variability across test datasets for each model.

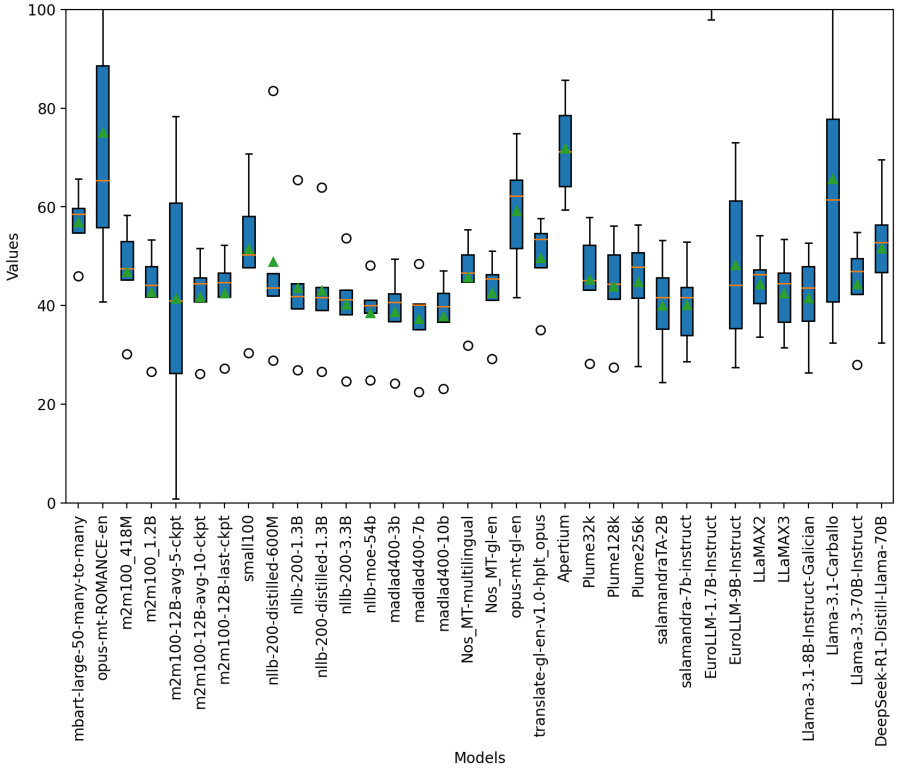

Figure 10: Box Plot of Galician–English TER Scores for All Evaluated Models that shows the variability across test datasets for each model.

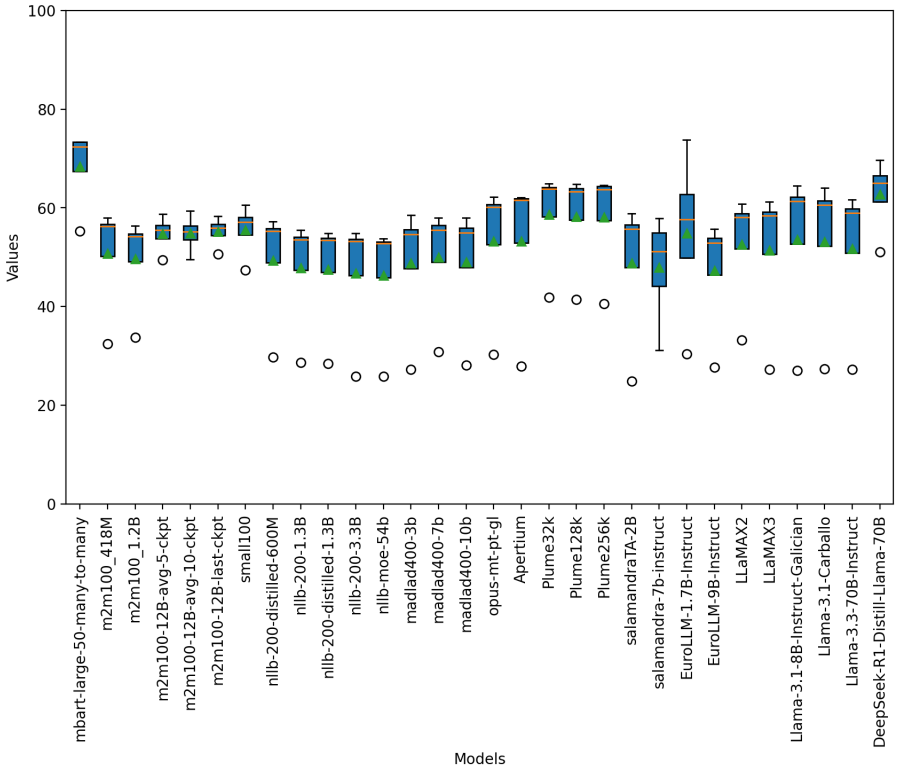

Figure 11: Box Plot of Portuguese-Galician TER Scores for All Evaluated Models that shows the variability across test datasets for each model.

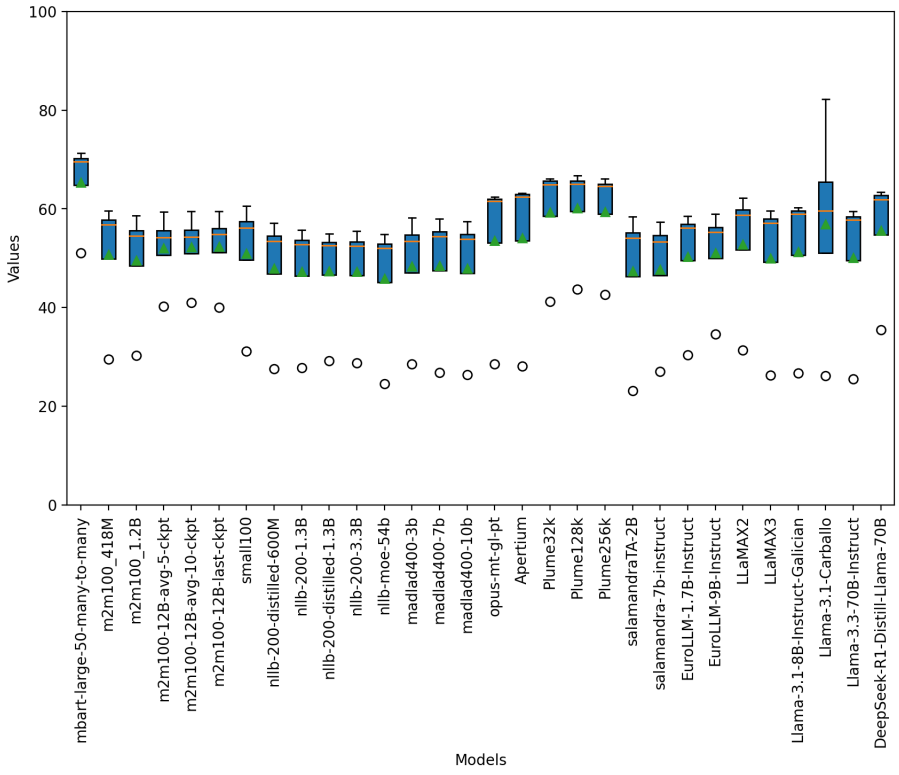

Figure 12: Box Plot of Galician–Portuguese TER Scores for All Evaluated Models that shows the variability across test datasets for each model.