A Strategy for Evaluating Machine Translation Models of Different Paradigms

Una estrategia para evaluar modelos de traducción automática de diferentes paradigmas

Pablo Gamallo,[1] **Daniel Bardanca,**[1] **Raquel Vázquez,**[1]
Lúa Santamarina,[1] **Saúl Buján**[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Univ. de Santiago de Compostela
{pablo.gamallo,lua.santamaria.montesinos}@usc.gal
{febratec,rvazqg,saulbujans}@gmail.com

Abstract: Currently, translation models based on different paradigms coexist, which can yield mixed results with no clear hierarchy yet established. This poses a challenge for automatic evaluation metrics, which appear to be more reliable when comparing models within the same paradigm, but not when comparing models of very different types. Using a Spanish–Galician test suite of 335 sentences and three representative machine translation systems (from seq2seq to LLMs), we measure the correlation of eight automatic metrics with binary human correctness judgments and find that BLASER aligns best with human evaluation, though the correlation remains moderate for all metrics. To address the residual bias, we propose **BLASER$_{\mathbf{PPL}}$**, a perplexity-weighted extension of BLASER that modulates the semantic similarity score with perplexity-based weight derived from a target-language model. With a weight parameter $\alpha \in [0.15, 0.20]$, the proposed metric recovers the human ranking for all three systems, rewarding translations that are both semantically faithful and linguistically natural without penalizing well-formed paraphrases, although not strictly literal.
Keywords: Machine translation, sequence-to-sequence models, generative language models, fine-tuning, evaluation metrics

Resumen: Actualmente, coexisten modelos de traducción basados en diferentes paradigmas que pueden dar resultados dispares sin que haya todavía una jerarquía clara entre ellos. Esto supone un desafío para las métricas de evaluación automática, que parecen ser más fiables cuando comparan modelos dentro del mismo paradigma, pero no cuando comparan modelos de tipos muy distintos. Utilizando un conjunto de prueba español-gallego de 335 oraciones y tres sistemas representativos de traducción automática (desde seq2seq hasta LLM), medimos la correlación de ocho métricas automáticas con juicios humanos binarios de corrección y encontramos que BLASER es la que mejor se alinea con la evaluación humana, aunque la correlación sigue siendo moderada para todas las métricas. Para abordar el sesgo residual, proponemos **BLASER$_{\mathbf{PPL}}$**, una extensión de BLASER ponderada por perplejidad que modula la puntuación de similitud semántica con un peso basado en la perplejidad derivada de un modelo de la lengua meta. Con un parámetro de peso $\alpha \in [0.15, 0.20]$, la métrica propuesta recupera el ranking humano para los tres sistemas, recompensando traducciones que son tanto semánticamente fieles a sus fuentes como lingüísticamente naturales sin penalizar las paráfrasis bien formadas aunque alejadas de la literalidad.
Palabras clave: Traducción automática, modelos sequence-to-sequence, modelos del lenguaje generativos, ajuste fino, métricas de evaluación

1 Introduction

Machine translation (MT) has undergone a profound transformation over the past decade. The rise of neural sequence-to-sequence (seq2seq) architectures (Sutskever, Vinyals, and Le, 2014) replaced statistical phrase-based approaches, achieving unprecedented translation quality on standard benchmarks. More recently, large language models (LLMs) based on the decoder-only transformer architecture (Vaswani et al., 2017) have further shifted the translation paradigm, demonstrating high-quality multilingual translation capabilities, in particular by means of fine-tuning and instruction techniques, even if the fine-tuned or instructed LLMs produce less literal translations (Stap et al., 2024)

Despite these advances, the evaluation of MT systems remains a largely unsolved challenge. As we have observed in the experiments we will describe later, MT systems of different architectures produce outputs of quite different linguistic nature, which makes it difficult to compare them with automated metrics. Seq2seq models, trained to maximize token-level likelihood against reference translations, tend to produce literal, word-order-preserving outputs that are well aligned with standard references (the gold standards). In contrast, instruction-tuned LLMs often produce more fluent, paraphrastic translations that rearrange constituents and choose near-synonymous expressions not present in the gold standard, generally obtaining lower perplexity (Hendy et al., 2023). This stylistic divergence causes automatic metric scores to vary across MT paradigms, not because one system translates better, but because the metrics tend to be sensitive to superficial differences between output styles.

This phenomenon poses a serious methodological problem: if we evaluate seq2seq models and instruction-tuned LLMs with the same automatic metric under the same conditions, we risk drawing conclusions that are misleading by output style rather than translation quality. The objective of this paper is to investigate evaluation strategies that are robust to this variability arising from the internal characteristics of MT systems.

To achieve this objective, we adopt the following methodology. First, we select a challenging evaluation dataset with Spanish-Galician sentence pairs, and use it to compare three representative MT paradigms: (*i*) a seq2seq encoder-decoder model, (*ii*) a fine-tuned decoder-only LLM, and (*iii*) a sophisticated instruction-following LLM. We evaluate the outputs of all three systems with a battery of different types of automatic metrics: shallow lexical metrics, such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006), and semantic embedding-based metrics, such as COMET (Rei et al., 2020), or BLASER (Chen et al., 2023). Second, we complement this quantitative evaluation with a structured human evaluation and measure the correlation between each automatic metric and human judgments across MT paradigms. Third, on the basis of this correlation analysis, we identify the metric that best tracks human evaluation and investigate how it can be further improved to account for the stylistic variation we observe among different MT paradigms.

Our analysis reveals that BLASER correlates most closely with human judgments. Nevertheless, even BLASER struggles to penalize adequately those translations produced by instruction-tuned LLMs that, although linguistically very different from the reference, exhibit low perplexity with respect to the target language model, which is a sign that the translations are well formed. Building on this finding, we propose a new **perplexity-weighted BLASER** score that modulates the BLASER similarity with the perplexity of the translated sentence. By incorporating a language model prior (i.e., perplexity), our metric rewards translations that are not only semantically close to the reference but also linguistically natural, and it is able to capture quality differences that arise when MT systems from different paradigms are compared.

The remainder of this paper is structured as follows. Section 2 reviews related work on MT evaluation metrics and the comparison of MT paradigms. Section 3 describes the evaluation dataset, the three MT systems representing the three paradigms under study, and the evaluations conducted with the different types of automatic metrics. Section 4 presents the human evaluation protocol and the correlation analysis

between automatic and human scores. Section 5 provides a qualitative linguistic analysis of the differences in output style across paradigms, giving rise to the proposed perplexity-weighted metric, and reports experimental results demonstrating its improved alignment with human judgments. Finally, Section 6 presents conclusions, limitations of the current approach, and directions for future work.

2 Related Work

2.1 Automatic Metrics for Machine Translation Evaluation

The automatic evaluation of machine translation has been studied for decades, giving rise to a diverse landscape of metrics that can be broadly categorized into two families: shallow similarity-based metrics and semantics-based neural metrics.

Shallow similarity-based metrics such as BLEU (Papineni et al., 2002), chrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006) compare system output against one or more reference translations using lexical overlap, or even character overlap, like in the case of chrF. More recent neural metrics such as COMET (Rei et al., 2020), BLEURT (Sellam, Das, and Parikh, 2020), BERTScore (Zhang et al., 2020), and BLASER (Chen et al., 2023) rely on more abstract representations to capture semantic similarity beyond surface form. We will use these eight metrics (four shallow and four semantic) in our experiments.

2.2 Correlation Between Automatic Metrics and Human Evaluation

The WMT shared tasks on metrics evaluation (Freitag et al., 2022; Kocmi et al., 2023) have provided a systematic framework for measuring how well automatic metrics correlate with human judgments. A consistent finding across years is that lexical metrics such as BLEU and TER correlate poorly with direct assessment scores (Ma, Bojar, and Graham, 2018). Neural metrics consistently outperform lexical ones, with COMET-based metrics dominating recent WMT shared tasks (Freitag et al., 2022; Kocmi et al., 2023).

However, a complementary line of work has investigated the failures of semantic-based automatic metrics. A more recent study (Buján et al., 2025), following a systematic comparison, shows that BLEU correlates better with human evaluations than neural metrics, such as COMET, in language pairs including the Galician language. Mathur et al. (2020) showed that many reported correlations are inflated by easy system comparisons, and that metrics diverge substantially when discriminating among high-quality systems. This is in accordance with the findings of our work.

2.3 Evaluation Across MT Paradigms

As mentioned earlier, neural MT research systems are based on two main translation paradigms: neural seq2seq models and, most recently, LLMs.

Encoder-decoder or seq2seq architectures based on parallel corpora (and thus fully supervised) became the dominant paradigm for MT, in particular, when large amounts of parallel data is available. Libraries such as OpenNMT (Klein et al., 2017), Marian (Junczys-Dowmunt et al., 2018), and FairSeq (Ott et al., 2019) made pre-trained seq2seq models very popular. A key characteristic of seq2seq models is their tendency to produce translations that closely mirror the structure of the source sentence, yielding outputs that score well on lexical metrics precisely because they stay close to reference expressions (Freitag, Grangier, and Caswell, 2020).

The emergent abilities of LLMs give them strong translation capabilities (Zhu et al., 2024), even if these models were not pre-trained with parallel corpora. This led researchers to explore post-training strategies for further enhancing this performance and better aligning model outputs with human translations (Alves et al., 2024). A particularly effective approach involves continual pre-training on a mixture of monolingual and parallel corpora, followed by supervised fine-tuning on high-quality translation data (Garcia Gilabert et al., 2025).

A recurring observation in the studies that compare the two paradigms is that LLM outputs differ stylistically from those of seq2seq systems: they tend to be more

paraphrastic, more idiomatic, and more variable in structure (Hendy et al., 2023; Kocmi et al., 2023). This stylistic divergence has direct consequences for automatic evaluation. Several studies have noted that BLEU scores underestimate LLM translation quality compared to human judgments, because the metric penalizes valid paraphrases not present in the reference (Freitag et al., 2022; Buján et al., 2025). Neural metrics are less affected but still exhibit paradigm-dependent behavior, as shown in the WMT23 findings (Kocmi et al., 2023). The present paper extends these findings by proposing a corrective approach based on language model perplexity.

3 Data, Models, and Metrics

3.1 Evaluation Dataset

The evaluation dataset used in this work is a Spanish–Galician test suite developed and released by Proxecto Nós (de Dios-Flores et al., 2022). The suite consists of 335 sentence pairs specifically designed to challenge MT systems across a range of linguistic phenomena, including lexical ambiguities, syntactic divergences between the two languages, morphological variation, and other translation difficulties that are characteristic of closely related Iberian Romance languages. By explicitly targeting these phenomena, the test suite provides a more diagnostic evaluation than general-purpose test sets, making it particularly appropriate for distinguishing the translation behavior of systems built on different paradigms.

3.2 Machine Translation Systems

We compare three MT systems, each representative of a different translation paradigm.

Seq2seq model The first system is TradutorNós,[1] a neural encoder-decoder model for Spanish–Galician translation developed within Proxecto Nós by using more than 30M training sentence pairs. It follows the standard seq2seq transformer architecture. This model is representative of the dominant supervised neural MT paradigm, whose outputs tend to closely mirror source structure and use expressions well aligned with standard reference translations (Buján et al., 2025; Outeirinho et al., 2024).

Fine-tuned LLM: The second system is a decoder-only LLM adapted for Spanish–Galician translation we have trained for the current experiments using parameter-efficient fine-tuning. Specifically, we applied LoRA (Hu et al., 2022) via the Unsloth framework to the Galician-Portuguese Llama-Carvalho 8B model (Rodríguez et al., 2025), a multilingual Llama-based model oriented towards Galician and Portuguese. Fine-tuning was performed on a Spanish–Galician subset of approximately 30,000 sentence pairs extracted from the OPUS corpus.[2] This system represents the *fine-tuned LLM* paradigm: a large generative model specialized for the translation task through targeted supervision, but with substantially fewer updated parameters than with full fine-tuning.

Instruction-following LLM: The third system is SalamandraTA (Garcia Gilabert et al., 2025), a 7B-parameter model built upon the SALAMANDRA architecture and developed at the Barcelona Supercomputing Center (BSC). SalamandraTA was trained in multiple phases: an initial stage of continual pretraining on large amounts of parallel and monolingual corpora, followed by a supervised fine-tuning stage on a smaller, curated set of parallel data. The model, which is fully fine-tuned, is guided at inference time through natural language instructions, placing it in the *instruction-following LLM* paradigm. Its outputs are accordingly more paraphrastic and stylistically natural than those of seq2seq or LoRA-fine-tuned systems.

3.3 Automatic Evaluation

We evaluate the three systems with a broad battery of automatic metrics. As noted above, we distinguish two families: *shallow* metrics, which rely on surface-form overlap with a reference translation, and *neural semantic* metrics, which leverage pre-trained multilingual representations to assess meaning equivalence.

Table 1 reports the scores obtained by the three systems across all metrics. For TER, lower values indicate better translation quality; for all other metrics, higher values

[1] `https://tradutor.nos.gal`

[2] `https://opus.nlpl.eu/`

Metric	Systems		
	Seq2seq (A)	**LLM-LoRA (B)**	**LLM-FFT (C)**
Shallow metrics			
BLEU	**74.15**	67.34	53.50
chrF	**86.86**	83.22	73.42
TER‡	**15.52**↓	19.91↓	33.53↓
METEOR	**0.882**	0.844	0.760
Neural semantic metrics			
COMET	**0.934**	0.916	0.902
BLEURT	**0.872**	0.849	0.808
BERTScore	**0.893**	0.859	0.783
BLASER	**4.610**	4.520	4.450

Table 1: Automatic evaluation results for the three MT systems on the Spanish–Galician test suite. Bold indicates best score per metric. ‡TER is an error rate (lower is better). Pattern ranking: A > B > C with no overlap. FFT stands for full fine-tune.

indicate better quality.

Across both metric families, a consistent ranking emerges: the seq2seq model (A) outperforms the LoRA fine-tuned LLM (B), which in turn outperforms the fully instruction-following LLM (C), yielding the ordering A > B > C for all eight metrics. This result is somewhat counterintuitive from the perspective of model size and general-purpose capability, as SalamandraTA is a substantially larger and more powerful language model than TradutorNós, but it is consistent with the well-documented tendency of shallow metrics to favor outputs that closely mirror reference wording.

4 Human Evaluation

4.1 Evaluation Protocol

Human evaluation was carried out by a trained linguist who assessed each translated sentence produced by the three MT systems against the corresponding source sentence. A translation was judged *correct* if it satisfied two independent criteria simultaneously: (i) it was *grammatically well-formed* in the target language (formal criterion), and (ii) it conveyed the *same state of affairs* as the source sentence (semantic criterion). Any translation that failed either criterion was marked as incorrect. This binary correctness judgment was applied uniformly across all three systems and the full 335-sentence test suite.

We deliberately adopt a coarse-grained protocol rather than a fine-grained error taxonomy such as the Multidimensional Quality Metrics framework (Freitag et al., 2021). Our objective is not to produce a detailed diagnostic of error types but to obtain a reliable human quality reference for ranking the three systems, which is what we need to compare against automatic metric rankings. The annotation protocol is deliberately binary and coarse-grained, which minimizes the subjectivity that typically motivates multi-annotator designs. Since the goal of the human evaluation is not to produce a fine-grained diagnostic annotation, but merely to establish a reliable system-level quality ranking for metric correlation analysis, we minimize the subjectivity that typically motivates multi-annotator designs. In this context, the binary judgments of a single expert linguist provide a sufficient and internally consistent gold reference.

4.2 Human Evaluation Results and Metric Alignment

Table 2 reports the percentage of sentences judged correct for each system. The human ranking is A > C > B. This ordering diverges from the A > B > C ranking returned by all automatic metrics (Section 3): the instruction-following model (C) is ranked above the LoRA fine-tuned model (B) by human judges, whereas all automatic metrics place B above C. This inversion reveals a systematic bias in automatic metrics: they seem to over-penalize the more paraphrastic output of the instruction-following LLM compared to what human judges actually perceive as translation quality. This will be

examined in the qualitative error analysis below.

System	Correctness
Seq2seq (A)	0.89
LLM-LoRA (B)	0.71
LLM-FFT (C)	0.78

Table 2: Human evaluation results: proportion of sentences judged correct (grammatically well-formed and semantically faithful) for each MT system.

To quantify how well each automatic metric aligns with the human ranking, the *gap asymmetry* $(A-B)-(B-C)$ is measured after normalizing all metric scores to the $[0, 1]$ range. This is an ad hoc, basic metric we have designed to compare the two rankings (or ordering patterns) by taking into account the distances between the models that the metrics reflect. According to the gap asymmetry, a positive value indicates that the metric assigns a larger gap between A and B than between B and C, which is consistent with the human-ordered pattern found in the current dataset: A > C > B (where B is clearly the weakest system). A negative value indicates the opposite: the metric overestimates the difference between B and C compared to the gap between A and B, misrepresenting the human ranking. Table 3 reports the gap asymmetry for all eight automatic metrics.

Metric	$(A-B)-(B-C)$
Shallow metrics	
BLEU	−0.3404
chrF	−0.4583
TER	−0.5125
METEOR	−0.3770
Neural semantic metrics	
COMET	+0.1250
BLEURT	−0.2812
BERTScore	−0.3820
BLASER	+0.1250

Table 3: Gap asymmetry $(A-B)-(B-C)$ for each automatic metric after score normalization. Positive values, in bold, indicate better alignment with the human ranking A > C > B.

Two metrics stand out with positive gap asymmetry: COMET and BLASER (both +0.1250). All shallow metrics show negative asymmetry, but BLEU has the least negative value (−0.3404), meaning it diverges less from the human ranking than its shallow counterparts. On the basis of these results, we select BLEU (best shallow metric), COMET, and BLASER for a deeper correlation study.

4.3 Correlation Between Human Evaluation and Automatic Metrics

We compute Pearson (r) and Spearman (ρ) correlations between the sentence-level automatic metric scores and the binary human correctness judgments for each system separately, as well as a ranking similarity score. Table 4 reports all values.

	BLEU	COMET	BLASER
r-A	0.29	0.19	0.37
ρ-A	0.30	0.28	0.29
r-B	0.40	0.33	0.38
ρ-B	0.39	0.46	0.39
r-C	0.29	0.43	0.33
ρ-C	0.29	0.42	0.40
Similarity	0.86	0.87	0.87
Total	**0.403**	**0.426**	**0.433**

Table 4: Pearson (r) and Spearman (ρ) correlations between automatic metric scores and human judgments, computed per system (A, B, C), together with a ranking similarity score and an overall average. All correlations are statistically significant ($p \leq 0.05$).

All correlations fall in the weak-to-moderate positive range, which is consistent with the known difficulty of correlating segment-level automatic metrics with binary human judgments (Mathur, Baldwin, and Cohn, 2020). The similarity score in the last row measures how closely the ranking induced by each automatic metric across the three systems matches the human ranking by comparing the normalized score vectors (of human and metric scores), sentence by sentence. More precisely, given a sentence, the similarity between the automatic evaluation and the human evaluation is measured as follows: the three values of a metric (for example, BLEU) for each of the three translations (the three systems' outputs) are encoded into a three-dimensional vector, as are the three values of the human evaluators' judgments. The cosine between these two vectors is then

computed, and the final similarity value is the mean across the 335 translated sentences. The Total score reported in Table 4 is the unweighted average of all six correlation coefficients (Pearson and Spearman for each of the three systems) plus the Similarity score, for a total of seven values. We acknowledge this is an ad hoc aggregation, but it is intended only as a convenient summary statistic rather than a formally derived measure.

Looking at the overall averages, the differences are small but consistent: BLASER is a slightly better predictor of human judgment than COMET, and both neural metrics outperform BLEU. These results confirm that, across paradigms, BLASER is the automatic metric most closely aligned with human evaluation. Nevertheless, the overall correlations remain moderate, and no metric fully captures the human A > C > B ordering. This limitation motivates a deeper analysis, which will be carried out in the next section, in order to search for a more appropriate automatic metric.

5 A New Metric

5.1 Qualitative Analysis of Cross-Paradigm Differences

Before introducing the new metric, we examine three representative sentences from the test suite that can be used to illustrate the different behavior of each model. Table 5 shows the Spanish source, the three system outputs, and a binary human correctness judgment for each.

These three examples illustrate the different behavior of each translation paradigm. System C (SalamandraTA) tends to produce non-literal translations with restructured syntax and varied word order: for instance, it uses a passive periphrasis (*foi expulsado* - was sent off) instead of a reflexive impersonal in the third example. This stylistic divergence from the reference is precisely what causes all automatic metrics to penalize system C more heavily than system B: all the metrics consistently give system C a lower score in these three examples, even though the answer is correct in the first and third examples, where the restructured output preserves identical truth conditions. In the second example, by contrast, system C makes a genuine semantic error –replacing *eye strain* with *blindness*–, which is a case of hallucination rather than mere paraphrase.

A complementary observation concerns system B. Its errors are of a different nature: they are morpho-syntactic (missing clitic in example 2, wrong personal verbal phrase instead of impersonal in example 3) rather than semantic. Since the output, even with errors, stays structurally close to the source and to the reference, all automatic metrics rate it above system C, while human judges correctly identify and penalize these errors. This explains the human ranking A > C > B: system C produces bolder paraphrases but fewer grammatical errors, while system B produces more literal output but with recurring grammatical mistakes.

5.2 Perplexity-Weighted BLASER

The qualitative analysis above suggests that a good metric should reward semantic correctness without penalizing well-formed paraphrases. As BLASER already captures semantic fidelity, we propose to augment it with a *perplexity* weight computed from a target-language model. This perplexity-based weight measures how linguistically natural the translated sentence is independently of its meaning.

BLASER and perplexity operate in very different measurement spaces, which makes their combination non-trivial. BLASER produces bounded scores in $[1, 5]$, where higher values indicate better translation quality. Perplexity, by contrast, is unbounded on $[1, \infty)$, where lower values indicate more fluent output. In order to make perplexity a reliable weight, we normalize it by mapping each perplexity value to weight $w \in [0, 1]$ as follows:

$$w = \exp\left(-\frac{\log \mathrm{PPL} - \log \mathrm{PPL}_{\min}}{\log \mathrm{PPL}_{\max} - \log \mathrm{PPL}_{\min}}\right) \quad (1)$$

This transformation yields $w \to 1$ for low-perplexity (fluent) outputs and $w \to 0$ for high-perplexity (disfluent) outputs. Finally, we define the metric via a weighted geometric combination controlled by w in this way:

$$\mathrm{BLASER}_{\mathrm{PPL}} = B \cdot w^{\alpha} \quad (2)$$

where $B \in [1, 5]$ is the BLASER score, $w \in [0, 1]$ is the normalized perplexity weight

Source (ES)	Seq2seq (A)	LLM-LoRA (B)	LLM-FFT (C)
El debate me pareció muy interesante. (I found the debate very interesting.)	*O debate pareceume moi interesante.* ✓	*O debate pareceume moi interesante.* ✓	*Considerei o debate moi interesante.* ✓
Esa luz tan blanca le puede producir molestias en la vista. (That bright white light may cause eye strain.)	*Esa luz tan branca pódelle producir molestias na vista.* ✓	*Esa luz tan branca pode producir molestias na vista.* ×	*A luz branca pode cegarlle.* ×
Se expulsó al jugador del campo con tarjeta roja. (The player was sent off with a red card.)	*Expulsouse ao xogador do campo con tarxeta vermella.* ✓	*Expulsou ao xogador do campo con tarxeta vermella.* ×	*O xogador foi expulsado do campo con tarxeta vermella.* ✓

Table 5: Three illustrative translation examples from the Spanish–Galician test suite. Correctness (✓ = correct, × = incorrect) is based on the criteria described in Section 4.

defined in Equation (1), and $\alpha \in [0, 1]$ is a tunable hyperparameter controlling the degree of trust placed in the perplexity value: when $\alpha = 0$, the metric behaves as just BLASER, and when $\alpha = 1$, perplexity fully modulates the score.

5.3 Experimental Results

To compute the perplexity value of the Galician sentences, we use Carballo-Bloom 1.3B (Gamallo et al., 2024), a Galician language model trained independently of all three MT systems under evaluation. Using a model that is unrelated to any of the three systems is essential to avoid contamination. Measuring the perplexity of the translated outputs with Carballo-Bloom 1.3B gives rise to a fluency ranking across MT models from the most to the least fluent: C > B > A, that is, SalamandraTA (LLM-FFT) produces the most fluent Galician according to the language model, followed by Llama-Carvalho (LLM-LoRA), with TradutorNós (seq2seq) scoring lowest. This is, in fact, the expected outcome: LLM-based systems, having been trained or fine-tuned on large amounts of natural text, tend to produce more fluent language than seq2seq models. Crucially, this fluency ordering (C > B > A) is the inverse of that given by the automatic metrics. Given this observation, it seems clear that automatic metrics (e.g., BLASER) and perplexity should be combined to refine the final ranking and recover the human ordering, with C positioned in the middle.

We evaluate $\text{BLASER}_{\text{PPL}}$ on the Spanish–Galician test suite for nine values of α ranging from 0 to 0.40 in steps of 0.05. Figure 1 plots the score for the three systems as a function of α.

Table 6 complements the figure with the exact scores and the induced system ranking at each α value.

α	A	B	C	Ranking
0.00	3.7816	3.6015	3.4991	A > B > C
0.05	3.8182	3.7308	3.4997	A > B > C
0.10	3.8474	3.7900	3.7381	A > B > C
0.15	3.8618	3.8236	3.8240	A > C ≈ B
0.20	**3.8670**	**3.8418**	**3.8647**	**A > C > B**
0.25	3.8661	3.8503	3.8867	C > A > B
0.30	3.8610	3.8520	3.8969	C > A > B
0.35	3.8529	3.8491	3.8994	C > A > B
0.40	3.8426	3.8428	3.8965	C > A ≈ B

Table 6: $\text{BLASER}_{\text{PPL}}$ scores (scale 1–5) and induced ranking for each value of α.

At $\alpha = 0$, the metric is equivalent to plain BLASER and yields the same A > B > C ordering as all other automatic metrics. As α increases, it gradually brings model C, which has better linguistic fluency (though a lower automatic score), closer to the other two. The crossover occurs at $\alpha \approx 0.15$–0.20: at $\alpha = 0.15$ the scores for B and C are virtually tied (3.8236 vs. 3.8240), and at $\alpha = 0.20$ the correct human ranking A > C > B is recovered for the first time. Beyond $\alpha = 0.25$, the perplexity weight inflates the score of system C disproportionately and produces a C > A > B ordering that no longer reflects human judgment.

These results confirm that $\text{BLASER}_{\text{PPL}}$ with $\alpha \in [0.15, 0.20]$ successfully corrects

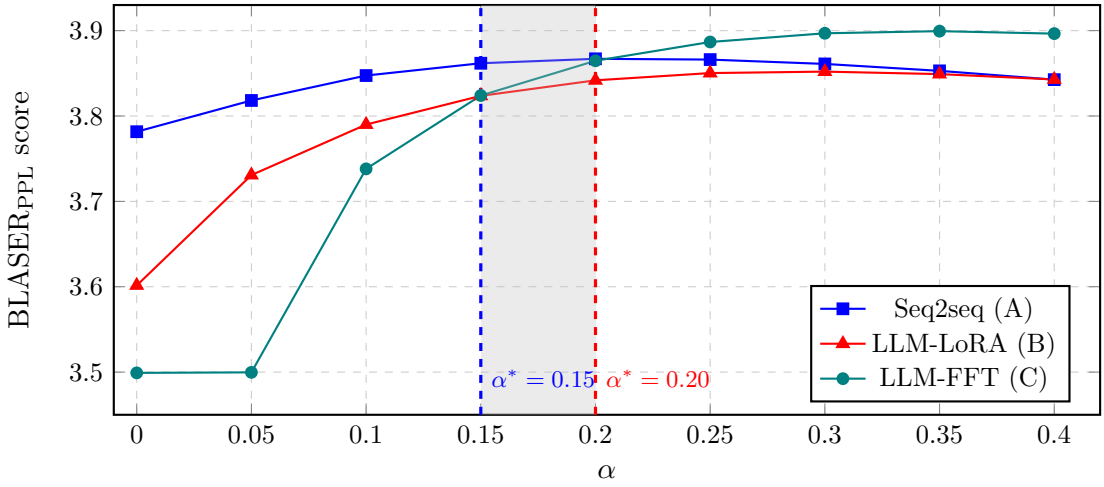

Figure 1: BLASER$_{PPL}$ scores for the three MT systems as a function of the perplexity weight α. The dashed vertical lines marks the space between $\alpha^* = 0.15$ and $\alpha^* = 0.20$, the smallest values of α at which the metric recovers the human ranking A > C > B.

the systematic bias of standard automatic metrics when comparing MT systems across different types of models. The improvement comes from a very simple, interpretable modification: incorporating a language model that rewards outputs that are linguistically natural in the target language. Obviously, this technique can be applied to any other MT metric. The weight of the perplexity is independent of the metric (BLASER or another) that it modifies.

6 Conclusions

6.1 General Conclusions

This paper examined the reliability of automatic MT evaluation metrics when comparing different types of translation systems. Our central finding is that all standard metrics –both shallow (BLEU, chrF, TER, METEOR) and neural semantic (COMET, BLEURT, BERTScore, BLASER)– show a consistent bias across different system types: They all rank a LoRA-tuned LLM higher than an instruction-following LLM, even though human evaluators consistently prefer the opposite order. This bias does not come from any single metric's design. Instead, it reflects a broader issue with reference-based evaluation: metrics reward outputs that stay close to the reference in form or embedding space, and instruction-following LLMs produce paraphrastic, syntactically varied outputs that diverge from the reference wording even when the translation is semantically correct. Therefore, automatic metrics should be used with caution when comparing systems from different paradigms, and cross-paradigm benchmarks should always include human evaluation alongside automatic metrics.

6.2 Specific Conclusions

The following specific conclusions can be drawn from the experiments reported in this paper.

First, the human evaluation reveals that the instruction-following LLM (SalamandraTA, system C) is preferred by a linguist evaluator over the LoRA fine-tuned LLM (Llama-Carvalho, system B), yielding the ranking A > C > B. The errors of system B are predominantly morpho-syntactic, whereas the errors of system C are rarer but more severe, involving occasional hallucinations of new states of affairs. The greater frequency of grammatical errors in system B is what ultimately places it below system C in human judgment.

Second, among the eight automatic metrics evaluated, BLASER achieves the highest overall alignment with human judgment (total score 0.433), followed closely by COMET (0.426). These two metrics are also the only ones with a positive gap asymmetry, meaning they partially capture the weaker discrimination between B and C that characterizes the human ranking. All shallow metrics show a pronounced negative gap asymmetry, with BLEU performing best.

Third, a perplexity analysis reveals that instruction-following LLMs produce more fluent Galician than seq2seq systems (fluency ranking: C > B > A), confirming that the metric bias seems to be rooted in the tendency of these LLMs to generate paraphrases with great structural variety rather than in translation errors per se.

Fourth, the proposed $\text{BLASER}_{\text{PPL}}$ metric, which modulates the BLASER score with perplexity, successfully recovers the human ranking A > C > B at $\alpha \in [0.15, 0.20]$. The improvement requires only a single tunable parameter and a target-language model, making the approach lightweight and applicable to any language pair for which a language model is available. However, since the parameter was adjusted using the evaluation test, we must be very cautious about the results, which are merely exploratory.

6.3 Limitations

Several limitations of this work should be acknowledged. First, the evaluation is conducted on a single language pair (Spanish–Galician) with a single test suite of 335 sentences. Spanish and Galician are closely related Romance languages, which may limit the generalization of the findings to more typologically distant pairs where the translation challenge is qualitatively different. Second, the optimal value of α for $\text{BLASER}_{\text{PPL}}$ was determined by grid search on the same evaluation set used for reporting, without an external development set; the reported threshold should therefore be treated as indicative rather than definitively calibrated. And finally, we evaluate only one system per paradigm, and the conclusions about paradigm-level differences may be sensitive to the specific models chosen.

6.4 Future Work

Several directions are open for future research. The most immediate priority is to replicate the analysis on additional language pairs, including both closely related languages and more distant pairs, to assess the robustness of the cross-paradigm bias and the optimal α range of $\text{BLASER}_{\text{PPL}}$. Thus, a natural extension of this work is to apply our approach to multilingual publicly available data, which includes human annotations; doing so would broaden the scope of our evaluation and yield a more substantial and generalizable contribution. A second direction is to replace the binary evaluation protocol with a fine-grained error taxonomy, which would allow a more precise characterization of the error types given by each paradigm and a more sensitive correlation analysis. Third, the perplexity weight could be extended to incorporate additional target-language information, such as grammaticality checkers or acceptability judgments. For instance, the final metric could also consider the edit distance between the system output and the corrected version produced by a language checker: the shorter the distance, the higher the translation quality. Finally, it would be interesting to investigate whether the α hyperparameter can be automatically estimated or trained from a small sample of human annotations, rather than tuned via search on the evaluation data.

The annotated dataset and the tool, with our metric, to reproduce the results are available under a free license.[3]

Acknowledgments

This paper was funded by MCIU/AEI (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, and AIA2025-163322-C62), by the Galician Government (Research Center of Galicia accreditation 2024-2027 ED431G-2023/04 and GPCE ED431B 2025/16), by QUANTUM_IBER_IA (ERDF/POCTEP), and by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU — NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*.

References

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

[3] `https://github.com/proxectonos/MT_PL`

Buján, S., D. Bardanca, P. Gamallo, I. de Dios-Flores, and J. R. Pichel. 2025. Machine translation for low-resource languages: Performance trade-offs between seq2seq and generative approaches. *Procesamiento del Lenguaje Natural*, 75(0):297–315.

Chen, M., P.-A. Duquenne, P. Andrews, J. Kao, A. Mourachko, H. Schwenk, and M. R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada, July. 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 I. Aldabe, B. Altuna, A. Farwell, and G. Rigau, editors, *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.

Freitag, M., G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, M., D. Grangier, and I. Caswell. 2020. BLEU might be guilty but references are not innocent. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online, November. Association for Computational Linguistics.

Freitag, M., R. Rei, N. Mathur, C. kiu Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and O. Bojar. 2022. Results of the WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the 7th Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.

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. *Procesamiento del Lenguaje Natural*, 73(0):259–270.

Garcia Gilabert, J., X. Liao, S. Da Dalt, E. Bohman, A. Mash, F. De Luca Fornaciari, I. Baucells, J. Llop, M. Claramunt, C. Escolano, and M. Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 614–637, Suzhou, China, November. Association for Computational Linguistics.

Hendy, A., M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint*, arXiv:2302.09210.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch. 2018. Marian: Fast neural machine translation in C++. In F. Liu and T. Solorio, editors, *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In

M. Bansal and H. Ji, editors, *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Kocmi, T., E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, P. Koehn, B. Marie, C. Monz, M. Morishita, K. Murray, M. Nagata, T. Nakazawa, M. Popel, M. Popović, M. Shmatova, and J. Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

Ma, Q., O. Bojar, and Y. Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels, October. Association for Computational Linguistics.

Mathur, N., T. Baldwin, and T. Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July. Association for Computational Linguistics.

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In W. Ammar, A. Louis, and N. Mostafazadeh, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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 (ACL)*, pages 311–318. Association for Computational Linguistics.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Rei, R., C. Stewart, A. C. Farinha, and A. Lavie. 2020. COMET: A neural framework for MT evaluation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rodríguez, P., S. P. Suárez, P. Gamallo, and S. S. Docio. 2025. Continued pretraining and interpretability-based evaluation for low-resource languages: A Galician case study. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL*

2025, pages 4622–4637, Vienna, Austria, July. Association for Computational Linguistics.

Sellam, T., D. Das, and A. P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7881–7892. Association for Computational Linguistics.

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 (AMTA)*, pages 223–231.

Stap, D., E. Hasler, B. Byrne, C. Monz, and K. Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206, Bangkok, Thailand, August. Association for Computational Linguistics.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.