Evaluating Small Language Models for English-to-Portuguese Translation: Impact of Model Scale and Quantization

Gustavo Lopes Tamiosso[1], Rafael Oleques Nunes[1] and Dennis Giovani Balreira[1]

[1]Institute of Informatics, Federal University of Rio Grande do Sul
Porto Alegre, Brazil

{gltamiosso, ronunes, dgbalreira}@inf.ufrgs.br

Abstract

Small language models (SLMs) are increasingly adopted for machine translation due to their lower computational and deployment costs, yet a focused and systematic evaluation for English-to-Portuguese remains limited. We benchmarked dozens of SLMs (135M–20B parameters) across multiple architectures and quantization schemes (FP16, Q8_0, Q4_K_M) on two datasets: FLORES-101 (Portuguese subset, 1,012 sentences) and the multidomain OPUS-100 dataset ($\sim$10k sentences). We computed lexical and semantic metrics (BLEU, chrF, and BERTScore) and assessed statistical differences using non-parametric Friedman tests over paired sentence-level scores, followed by Wilcoxon signed-rank post-hoc comparisons with Holm correction. Normality assumptions are evaluated using the Shapiro–Wilk test. Our results strongly suggest that 8-bit quantization (Q8_0) preserves semantic quality with negligible average loss, while 4-bit quantization (Q4_K_M) reaches statistical significance in roughly half of model configurations, paired effect sizes (Cliff's δ) remain negligible to small in magnitude, with measurable degradation concentrated in lower-capacity models. Model scale exhibits only a weak correlation with translation quality: medium-sized models can match or outperform larger ones depending on model family and pretraining. These findings highlight trade-offs between efficiency and quality and inform the design of practical English–to-Portuguese translation pipelines based on SLMs.

1 Introduction

Small and medium-scale language models (SLMs) have recently achieved competitive performance on a wide range of natural language processing tasks, challenging the assumption that high-quality machine translation requires very large, proprietary systems. At the same time, practical deployment constraints—such as memory footprint, latency, and inference cost—have increased interest in efficient inference techniques, particularly low-bit quantization.

Despite extensive multilingual evaluation efforts, English-to-Portuguese (En-Pt) translation remains comparatively underexplored in this context. Portuguese exhibits rich morphology and significant regional variation, and high-quality translation is critical for real-world applications in Brazil and Portugal. Understanding how different model families and efficiency-oriented configurations behave on this language pair is therefore important for both research and deployment.

In this work, we present a large-scale evaluation of En-Pt translation using a diverse set of SLMs spanning multiple model families (Gemma, Qwen, LLaMA, Mistral, OLMo, GPT, and others), parameter scales (from 135M to 20B), and inference-time quantization strategies (FP16, Q8_0, and Q4_K_M). We evaluate performance on two complementary datasets: FLORES-101, representing clean and formal text, and the OPUS-100 multidomain dataset, capturing a diverse range of linguistic registers.

Our analysis focuses on three questions: (i) how quantization affects semantic and lexical translation quality across model families, (ii) how model scale relates to translation performance under realistic deployment settings, and (iii) whether strong open-weight models can match or surpass proprietary systems in practice. By combining aggregate metrics, statistical testing, and effect-size analysis, we aim to provide empirically grounded guidance for selecting and deploying SLMs for En-Pt machine translation.

2 Related Work

Recent work on neural machine translation has increasingly explored the trade-offs between model size, numerical precision, and translation quality,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 943–952

motivated by the need for more efficient yet accurate systems. In particular, quantization has emerged as a central technique to reduce memory footprint and computational cost while preserving performance. Prior studies have investigated quantization at different bit widths, its interaction with model capacity, and its impact across architectures and evaluation benchmarks.

Quinn and Ballesteros (2018) are pioneers in applying 8-bit quantization to neural machine translation without quality loss. They demonstrate that 8-bit precision matched or exceeded the 32-bit baseline's BLEU scores, confirming that reduced precision can preserve translation accuracy and adequacy.

Prato et al. (2020) propose FullyQT, a fully quantized Transformer for NMT. Testing on WMT14 (English–to-French/German), they observe that all components can be quantized to 8 bits without degrading quality; the quantized models attained BLEU scores comparable to or surpassing their full-precision baselines. This work demonstrates that aggressive quantization can preserve translation quality.

Behnke et al. (2021) evaluated smaller models for the efficient translation task (English-to-German), focusing on simplified RNNs, distillation, pruning, and quantization targeting lower latency. They report that reduced precision impacts smaller models more strongly—the smaller the original model, the larger the quality drop after 8-bit quantization. In contrast, fine-tuning the quantized model helps recover part of the lost quality. Thus, they empirically confirm that small models suffer greater degradation under quantization.

Jin et al. (2024) analyzed LLMs of various sizes (from a few billion to tens of billions of parameters). Their findings suggest that while 4-bit quantized LLMs often retain performance parity with high-precision versions, model scale plays a crucial role: larger quantized models frequently outperform smaller non-quantized ones, indicating that capacity can effectively offset the loss in numerical precision.

While prior studies have established the general viability of quantization, they predominantly focus on high-resource pairs like English-German or rely on findings from earlier architectures and massive LLMs. To the best of our knowledge, our work differentiates itself by conducting a systematic evaluation specifically for English-to-Portuguese translation within the emerging ecosystem of Small Language Models (SLMs). We extend the existing literature by applying rigorous non-parametric statistical testing to quantify the precise trade-offs between model scale (135M–20B), aggressive quantization (down to 4-bit), and semantic quality, filling a critical gap for efficient deployment in Portuguese-centric applications.

3 Datasets

We evaluate models on two complementary datasets to capture different translation conditions.

FLORES-101. We use the Portuguese subset (por_Latn) of FLORES-101 with 1,012 sentences from the official test split (Goyal et al., 2022). This dataset contains high-quality, professionally curated sentences with a relatively formal style, making it suitable for controlled evaluation of core machine translation performance.

OPUS-100. We use the English-to-Portuguese subset of the OPUS-100 corpus (Zhang et al., 2020), a large-scale multilingual corpus covering 100 languages. Specifically, we sampled 9,271 parallel sentence pairs from the test split to evaluate performance on a diverse range of domains. OPUS-100 is constructed by sampling from the broader OPUS collection (Tiedemann, 2012), which incorporates diverse sources such as movie subtitles, parliamentary proceedings, software documentation, web crawls, and religious texts, without curation or domain balancing (Zhang et al., 2020). This dataset complements the Wikimedia-based FLORES-101 by providing a more representative sample of general-domain language and testing model robustness across different linguistic registers.

4 Models and Prompting

We evaluated a large set of translation models. All models were evaluated under a controlled zero-shot setting using a fixed user-level instruction prompt, while system-level prompts were either neutral or model-specific when required by the interface.

The user prompt explicitly specifies the translation task and output format, ensuring consistent behavior across models:

```
Translate the following text from English
to Portuguese.

Return ONLY the translated text with no
explanations.
```

```
Text: {text}

Translation:
```

We ensured that the `source` sentence content was identical across all models, using the same cleaned input text, so that each translation corresponds to a strict one-to-one pairing. This setup enables reliable per-sentence statistical comparisons while isolating model effects from prompting variations.

We evaluate a diverse set of small and medium-scale language models covering multiple families, architectural choices, and training paradigms, as summarized in Table 1. The evaluated models span open-weight systems, instruction-tuned variants, and proprietary API-based models, with parameter counts ranging from 135M to 20B. This selection was designed to reflect the current ecosystem of Small Language Models (SLMs), emphasizing practical deployment scenarios where efficiency and quantization are critical.

The benchmark includes distilled reasoning-oriented models from the DeepSeek R1 family, instruction-tuned variants of Gemma 3 and Gemma 3n, and multiple generations of the LLaMA family (3, 3.1, and 3.2), covering both base and instruction-tuned configurations. We further include Mistral and OLMo instruction-tuned models, which are commonly used as strong open-weight baselines, as well as Qwen 3 base and instruction variants spanning sub-billion to mid-sized models. To capture the behavior of very compact instruction-following systems, we also evaluate SmolLM2 models at different parameter scales. Finally, GPT-4o and GPT-4o-mini are included as reference systems accessed via API, for which internal numerical precision is not disclosed.

Across all open-weight models, we evaluate a consistent set of inference-time quantization formats, including FP16, Q4_K_M, and Q8_0. This setup allows for controlled comparisons across families and model sizes, isolating the impact of quantization on translation quality while maintaining architectural diversity.

5 Metrics and Evaluation

To provide a comprehensive and balanced assessment of translation quality, we employ a set of complementary automatic evaluation metrics that capture different aspects of system performance. These metrics jointly evaluate surface-level lexical overlap, robustness to morphological variation, structural similarity, and semantic adequacy, allowing for a more nuanced comparison across models. All metrics are computed under identical input conditions to ensure fair and directly comparable results. For each evaluated translation, we computed the following automatic metrics:

- **BLEU** (Papineni et al., 2002), computed at the corpus level, with sentence-level approximations used for fine-grained analysis.
- **chrF** (Popović, 2015), a character n-gram F-score metric that is particularly robust for morphologically rich languages.
- **BERTScore** (Zhang* et al., 2020), an embedding-based metric designed to capture semantic similarity beyond surface-level n-gram overlap.

All semantic metrics are computed at sentence level and then aggregated by simple averaging. BLEU and chrF are additionally reported at corpus level. For pairwise model comparisons, we compute the paired difference per sentence:

$$\Delta = \text{Score}_A - \text{Score}_B$$

and analyze the distribution of Δ values across the test set.

5.1 Statistical testing

Sentence-level machine translation metrics are bounded and often non-normally distributed, with strong ceiling effects observed for semantic similarity measures such as BERTScore. We therefore adopt a non-parametric statistical framework.

For a given experimental factor (e.g., quantization format or model size), we construct sentence-aligned score matrices where each row corresponds to a sentence and each column to a model configuration. Overall differences are assessed using the Friedman test over paired sentence-level scores. To investigate intra-family stability, we further apply the Friedman test independently within each model family, treating its various versions and quantization levels as experimental conditions.

When the Friedman test indicates significant differences, we perform pairwise post-hoc comparisons using the Wilcoxon signed-rank test. To control the family-wise error rate across multiple comparisons, p-values are adjusted using Holm's correction. Additionally, as p-values are highly

Family	Model Specifications	Quantizations
DeepSeek	R1-Distill-Qwen (1.5B)	fp16, q4_k_m, q8_0
Gemma	3 Instruct (1B, 4B), 3n Instruct (2B, 4B)	fp16, q4_k_m, q8_0
GPT	4o, 4o-mini, OSS (20B)	N/A (API/Proprietary)
Llama 3	Base (8B), Inst. (8B)	fp16, q4_k_m, q8_0
Llama 3.1	Base (8B), Inst. (8B)	fp16, q4_k_m, q8_0
Llama 3.2	Inst. (1B)	fp16, q4_k_m, q8_0
Mistral	Inst. v0.1 (7B), Inst. v0.3 (7B)	fp16, q4_k_m, q8_0
OLMo	Inst. 2-1124 (7B), Inst. 3 (7B)	fp16, q4_k_m, q8_0
Qwen	3 Base (600M, 1.7B), 3 Inst. (4B)	fp16, q4_k_m, q8_0
SmolLM2	Inst. (135M, 360M, 1.7B)	fp16, q4_k_m, q8_0

Table 1: Overview of evaluated model families, variants, and supported quantization formats.

sensitive to large sample sizes (N), we report the paired Cliff's Delta (δ) to quantify the magnitude of differences and assess their practical significance. This non-parametric effect-size measure allows for a more nuanced interpretation of performance gaps that reach statistical significance despite being marginal in absolute terms.

Normality is evaluated using the Shapiro–Wilk test on sentence-level scores (with subsampling when necessary due to large sample sizes) and is used only as a diagnostic; all reported significance tests are non-parametric.

6 Implementation details and reproducibility

All experiments were executed with identical prompts and post-processing of outputs. A script was used to sanitize the `source` field (removing the prompt wrappers) and to recompute metrics consistently. BERTScore was computed with a portuguese encoder (`neuralmind/bert-large-portuguese-cased`) (Souza et al., 2020) and rescaled with baseline as recommended.

7 Results

7.1 Overview

All reported tests use complete paired sentence sets after preprocessing: FLORES-101 ($N = 1{,}012$) and OPUS-100 ($N = 9{,}271$). Table 2 summarizes per-family performance. For lexical metrics (BLEU, chrF), values represent the mean and standard deviation across the different model variants and quantization levels within each family. For BERTScore (F1), we report the mean and standard deviation computed over all individual sentence-level scores across all models in the family, thus capturing both model-induced variance and sentence-level difficulty. These aggregated values provide a compact view of family-level behavior; subsequent paragraphs interpret these aggregates together with the per-family Friedman tests and the paid-versus-open comparisons.

7.2 Aggregated observations

Overall dataset contrast. Scores on FLORES-101 are consistently and substantially higher than on OPUS-100 across all families, reflecting the difference between a curated, formal benchmark (FLORES) and the multidomain sources of OPUS-100. BLEU and chrF show the largest absolute declines; BERTScore decreases are smaller in absolute terms, indicating that semantic similarity is better preserved than surface overlap when models translate multidomain input. This pattern is explicitly visualized in the scatter plots of Figure 1, where the cluster of models shifts leftward (lower BLEU) on OPUS-100 while maintaining a relatively high position on the y-axis (BERTScore). The family-level averages confirming these declines across all metrics are detailed in Table 2.

Top-performing families. On FLORES-101 the GPT and Gemma families lead in all metrics (BERTScore ≈ 0.90 for Gemma3n and ≈ 0.91 for GPT; BLEU above 40–50), clustering tightly in the top-right quadrant of Figure 1(a). Gemma3 also performs strongly, with Gemma3 typically being the single most stable open-weight family (high mean and low variance, as seen in Table 2). OLMo and Mistral form a middle tier. SmolLM and DeepSeek form the lower tier, with substantially lower absolute scores in both metrics.

Family	Dataset	BLEU	BERTScore (F1)	chrF
DeepSeek	FLORES-101	5.81 ± 0.33	0.66 ± 0.07	31.39 ± 0.28
	OPUS-100	2.72 ± 0.35	0.66 ± 0.08	21.06 ± 0.88
Gemma3	FLORES-101	41.13 ± 5.75	0.88 ± 0.06	65.63 ± 4.26
	OPUS-100	21.17 ± 3.96	0.81 ± 0.09	47.61 ± 3.41
Gemma3n	FLORES-101	47.25 ± 0.82	0.90 ± 0.05	70.00 ± 0.45
	OPUS-100	23.64 ± 4.14	0.83 ± 0.09	50.83 ± 2.28
GPT	FLORES-101	50.55 ± 2.34	0.91 ± 0.06	71.98 ± 1.23
	OPUS-100	26.22 ± 0.96	0.83 ± 0.10	52.04 ± 0.29
LLaMA	FLORES-101	22.73 ± 16.42	0.85 ± 0.10	47.87 ± 18.01
	OPUS-100	11.06 ± 9.62	0.78 ± 0.13	30.30 ± 18.18
Mistral	FLORES-101	26.40 ± 7.04	0.84 ± 0.08	57.46 ± 3.87
	OPUS-100	13.86 ± 2.38	0.79 ± 0.10	41.71 ± 1.61
OLMo	FLORES-101	33.36 ± 1.42	0.85 ± 0.06	60.95 ± 0.90
	OPUS-100	18.25 ± 0.29	0.80 ± 0.09	44.75 ± 0.23
Qwen	FLORES-101	28.58 ± 14.52	0.84 ± 0.09	55.22 ± 13.49
	OPUS-100	16.00 ± 7.10	0.78 ± 0.11	40.25 ± 10.98
SmolLM	FLORES-101	4.20 ± 5.18	0.65 ± 0.09	24.23 ± 13.96
	OPUS-100	2.44 ± 3.40	0.65 ± 0.10	14.70 ± 11.97

Table 2: Translation performance by model family (mean $\pm$ std). For BLEU and chrF, statistics reflect variation across different models within each family. For BERTScore, statistics are computed over all individual sentence-level scores across all models in the family, capturing both model and sentence-level variance.

Stability and variance. LLaMA and Qwen exhibit large standard deviations (noticeable on FLORES-101), signaling high sensitivity to variant choice and quantization; families such as Gemma3n, GPT and OLMo show comparatively low variance, indicating stable behavior across tested variants and quantization formats. See Table 2 for aggregated dispersions and Figures 2) for family-level quantization stability.

7.3 Statistical tests and interpretation

Intra-family variation (Friedman). Per-family Friedman tests (performed per dataset) indicate significant intra-family differences for several families: Qwen, LLaMA, Mistral, Gemma3, SmolLM, OLMo and GPT show statistically significant heterogeneity across their variants (FLORES-101 and OPUS-100; $p \ll 0.05$ in the reported tests). Gemma3n shows no significant intra-family variation in either dataset (FLORES-101: $p = 0.074$; OPUS-100: $p = 0.868$), and DeepSeek is not significant on FLORES-101 ($p = 0.536$). These results demonstrate that choosing a variant (or quantization) within a family can materially affect translation quality for some families but not for others. The per-family differences align with the family trajectories shown in Figure 4.

Paid models vs. best open models (Wilcoxon). Using an adversarial baseline composed of the best open-weight performance per sentence, paired Wilcoxon signed-rank tests indicate that GPT-4o and GPT-4o-mini are, on average, slightly outperformed by the best open alternative. The open-weight ensemble for this comparison includes the top-ranking configurations in our benchmark: Gemma 3n (it), Qwen 3 (it), and GPT-OSS (20B), covering multiple quantization levels (FP16, Q8_0, and Q4_K_M) where applicable.

- FLORES-101: GPT-4o $\bar{\Delta} = -0.0188$ (Wilcoxon $p \approx 8.1 \times 10^{-63}$); GPT-4o-mini $\bar{\Delta} = -0.0196$ ($p \approx 6.7 \times 10^{-75}$).
- OPUS-100: GPT-4o $\bar{\Delta} = -0.0489$ ($p \ll 0.001$); GPT-4o-mini $\bar{\Delta} = -0.0494$ ($p \ll 0.001$).

The negative mean differences indicate systematic advantages for the best open-weight oracle. However, it is critical to distinguish between this ensemble-based oracle and individual model capabilities, as shown in Table 3. Pairwise 1v1 comparisons reveal that while GPT-4o holds a consistent statistical edge, the absolute semantic gap remains marginal. Remarkably, the Gemma 3n family achieves near-parity with GPT-4o, trailing by only a marginal semantic gap of 0.0095 average BERTScore F1 on FLORES-101 (Win Rate: 68.8%), and further shrinking to a negligible 0.0013 on OPUS-100 (Win Rate: 53.5%).

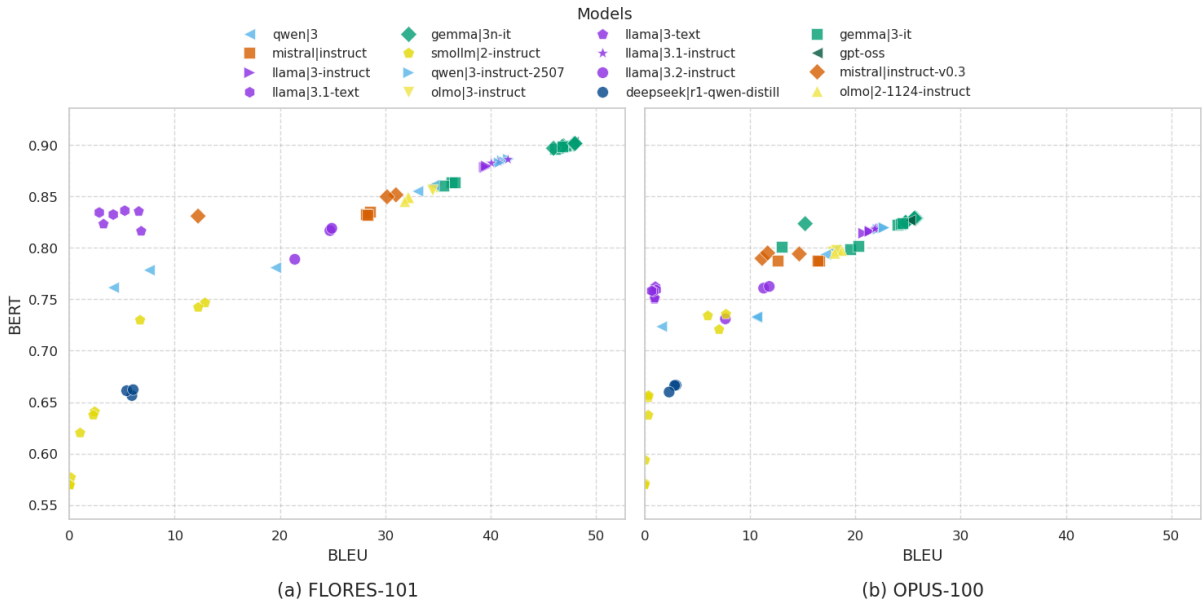


Figure 1: Correlation between lexical overlap (BLEU) and semantic quality (BERTScore F1) across (a) FLORES-101 and (b) OPUS-100 datasets. Note how BERTScore remains relatively high on OPUS-100 even as BLEU drops significantly compared to FLORES-101.

This demonstrates that mid-sized SLMs can now achieve semantic scores that are numerically close to state-of-the-art proprietary systems, although proprietary models still maintain a consistent statistical advantage in per-sentence win rates.

Effect-size and practical relevance. Because p-values are strongly influenced by large N, we report paired Cliff's delta (δ) as a non-parametric measure of practical significance (Cliff, 1993). Individual pairwise comparisons reveal a more nuanced picture than the oracle baseline: GPT-4o versus Gemma 3n (FP16), the strongest single open-weight model, yields a medium effect on FLORES-101 ($\delta = 0.39$) but a negligible effect on OPUS-100 ($\delta = 0.14$). GPT-4o versus GPT-OSS shows a small effect on FLORES-101 ($\delta = 0.28$) and negligible on OPUS-100 ($\delta = 0.13$). These results confirm that while GPT-4o maintains a consistent statistical advantage in direct 1v1 comparisons, the absolute gap remains practically marginal across both datasets. For quantization comparisons, Cliff's δ was negligible ($|\delta| < 0.147$) in 41 of 45 intra-family pairs on FLORES-101 and 44 of 45 on OPUS-100. The massive impact of instruction tuning is further evidenced by the performance of the LLaMA 3 family, where the Instruct variant outperforms the Base variant by a substantial ~ 0.065 BERTScore F1 gap on both datasets.

Qualitative insights. A qualitative review of translations from high-performing models (GPT-4o, Gemma 3n) and low-capacity systems (SmolLM2-135M) reveals sharp behavioral divides. Stronger models consistently maintain syntactic fluency and proper terminology in Portuguese, even when their stylistic choices differ. In contrast, the smallest models frequently fail to follow instructions, often including prompt meta-talk or producing incoherent, mixed-language outputs. These failures are particularly pronounced in formal technical domains, where low parameter counts appear insufficient to capture complex grammatical and semantic relationships.

8 Discussion

The results highlight that model scale correlates only weakly with translation quality ($r \approx 0.3$). Our findings confirm that mid-sized models in the 2–4B range often match or exceed 20B-class counterparts. This demonstrates that architectural choices, such as those in the Gemma 3n family, outweigh raw parameter count for En-Pt translation.

Aggregated analysis (visualized in Figure 2, 3) confirms that Q8_0 quantization preserves performance close to FP16: across all 15 model configurations, the maximum observed semantic impact was only 0.0069 BERTScore F1, and all paired Cliff's δ values were negligible ($|\delta| < 0.06$). Q4_K_M reached significance more frequently

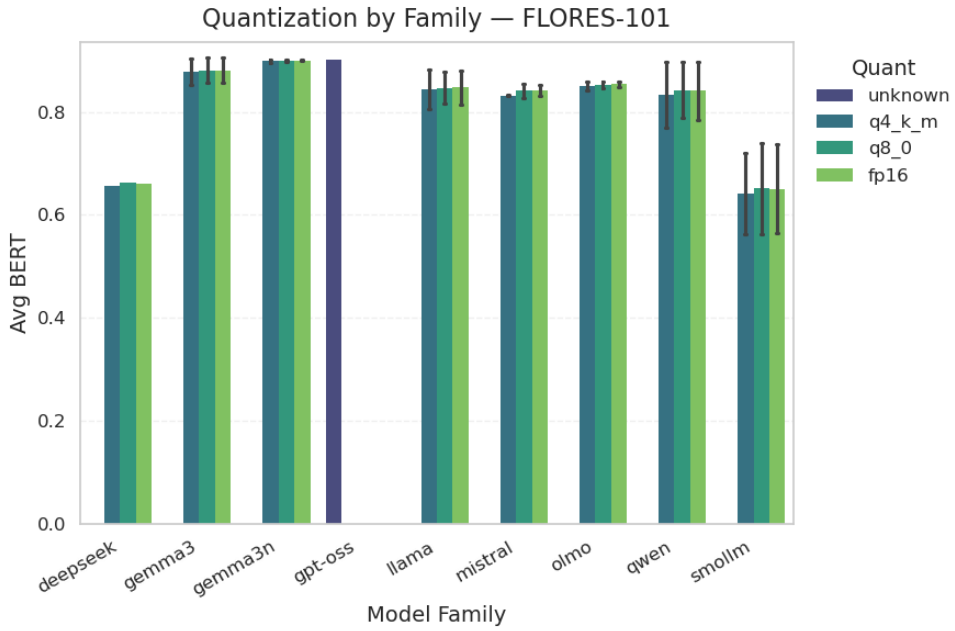


Figure 2: Comparison of quantization stability across model families on FLORES-101 (BERT score).

Comparison	Dataset	Δ BERT	Win %	Cliff's δ
GPT-4o vs Gemma 3n	FLORES	+0.009	68.8%	0.39 (Medium)
GPT-4o vs Gemma 3n	OPUS	+0.001	53.5%	0.14 (Negligible)
GPT-4o vs GPT-OSS	FLORES	+0.007	64.2%	0.28 (Small)
GPT-4o vs GPT-OSS	OPUS	+0.001	52.8%	0.13 (Negligible)

Table 3: Compact win-rate summary comparing GPT-4o against top-ranking open models (Gemma 3n 4B FP16 and GPT-OSS 20B). Metrics: average BERTScore F1 difference, GPT-4o win rate, and Cliff's delta (Negligible, Small, Medium).

(6/15 on FLORES-101, 9/15 on OPUS-100), but paired Cliff's δ remained negligible for the majority of comparisons.

The near-lossless performance of Q8_0 and the resilience of Q4_K_M quantization are critical for **edge deployment**. Since 4B models can maintain high semantic quality even under 4-bit compression, professional-grade translation is now feasible on consumer-grade hardware, eliminating the need for costly cloud-based APIs.

Only LLaMA 3.2 (1B) showed a consistent small effect across datasets ($\delta \approx 0.18$–0.23), confirming that 4-bit degradation is concentrated in lower-capacity models. This observation is consistent with the trends reported by Behnke et al. (2021), suggesting that models with lower parameter counts possess less redundant capacity to absorb the quantization errors introduced by 4-bit precision. The impact of quantization remains model-dependent: families such as Gemma3n are resilient across all tested formats, while LLaMA and Qwen exhibit higher sensitivity.

Model scale and trajectories. Model size correlates weakly with BERTScore (observed Pearson $r \approx 0.3$ in our size-vs-quality plots; see Figure 4). Several mid-sized models (2–4B Gemma variants) match or outperform certain 20B-class models, showing that architecture and training regime can outweigh raw parameter count. Notably, the Gemma 3n family displays atypical trajectories, where smaller variants maintain near-parity with larger counterparts, a behavior consistent with its nested architectural design (see Section 9). Effect-size analysis confirms that size differences produce large effects only at extreme scale gaps (e.g., SmolLM 135M vs. 1.7B, Cliff's $\delta = 0.99$; Qwen 600M vs. 4B, $\delta = 0.96$), while models within the same order of magnitude show negligible to small differences.

9 Architectural Considerations and Quantization Robustness

Our results highlight how specific architectural choices influence translation quality and robustness

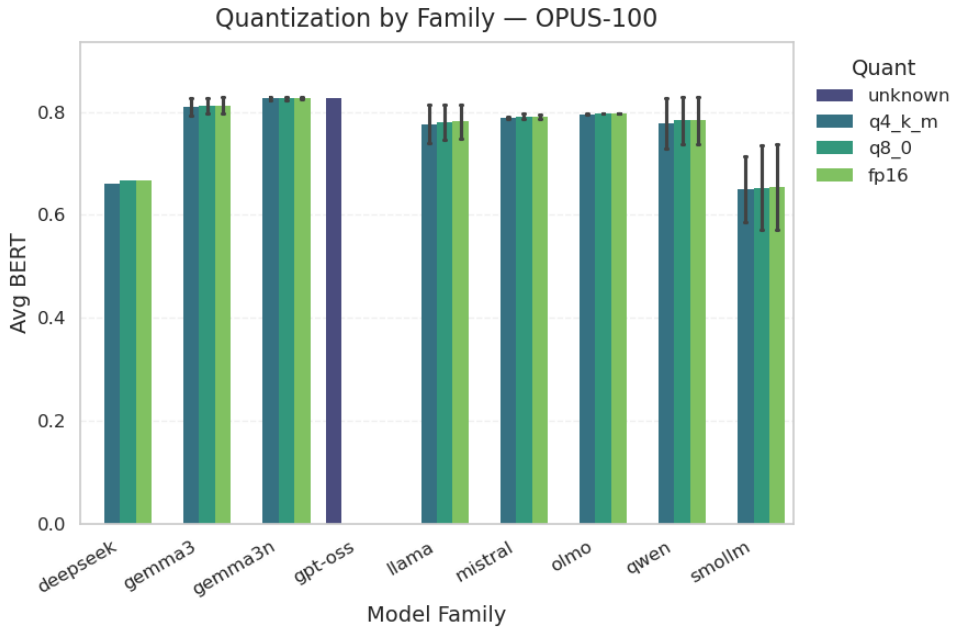


Figure 3: Comparison of quantization stability across model families on OPUS-100 (BERT score).

to quantization. In particular, the anomalies observed in Gemma 3n's performance—specifically the near-zero variance across different quantization levels and sizes (see Figure 2)—can be explained by its specialized MatFormer and Per-Layer Embedding (PLE) architecture (Gemma Team, 2025). While standard dense transformers like Llama or Mistral exhibit significant performance drops under aggressive quantization, the "nested" weight structures of MatFormer are optimized for consistency across scales and precision formats. This architectural robustness results in the overlapping performance clusters seen in our figures, which may appear atypical when compared to traditional model families but represent a key efficiency feature of the Gemma 3 generation. Furthermore, the strong multilingual performance of families like Qwen 3 likely stems from their broad pre-training on 119 languages and dialects (Yang et al., 2025), including Portuguese, which contrasts with English-centric models like SmolLM2.

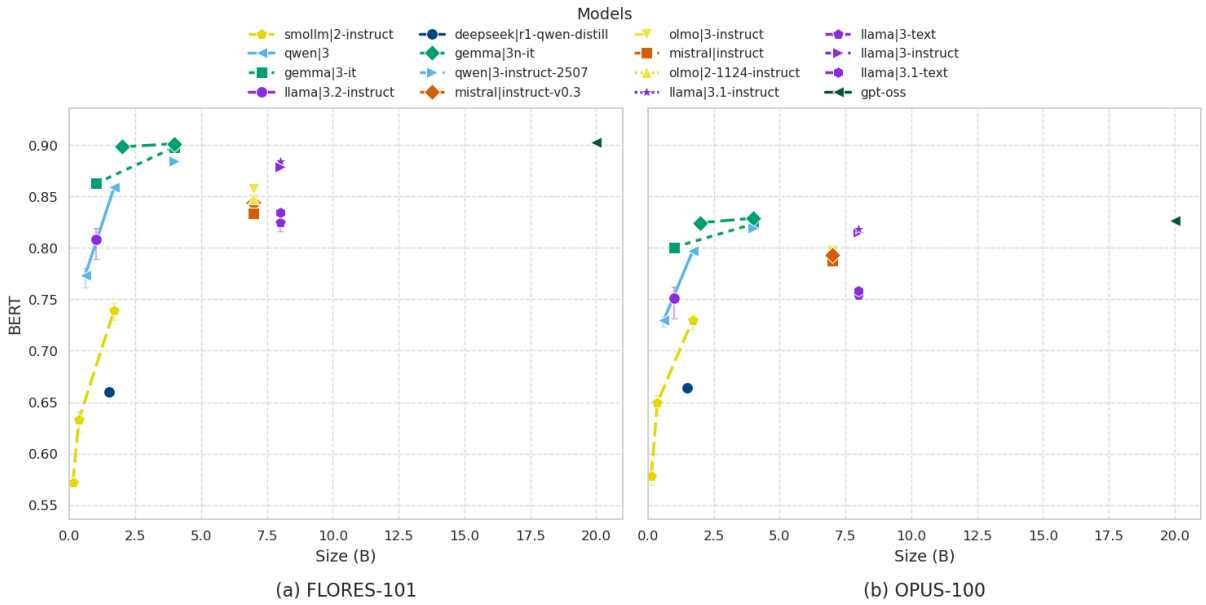
10 Limitations and Future Work

While our evaluation employs a broad suite of automatic metrics and rigorous statistical testing, it is subject to several limitations. First, the lack of human evaluation is a known constraint; while BERTScore captures semantic adequacy, small absolute differences require further qualitative validation to confirm their practical impact. Second, our zero-shot instruction setting may naturally favor models optimized for instruction-following over those primarily trained for raw translation tasks.

Additionally, our study uses datasets whose Portuguese portions are labeled generically as `pt`, without explicit annotation of regional variants. In practice, the distributions are likely skewed towards Brazilian Portuguese, but this has not been quantified. Future work should explicitly assess performance across regional variants (such as European and African Portuguese), extend this analysis to additional language pairs, include human evaluation, and integrate latency and energy measurements to better align MT evaluation with real deployment constraints.

11 Conclusion

In conclusion, the current generation of SLMs, particularly the 4B Gemma 3n, has reached a level of quality where deployment constraints—such as latency, cost, and local execution—become the primary decision factors. Our study reinforces that localized, quantized SLMs provide a transparent and efficient alternative to proprietary systems for the En-Pt pair. By demonstrating that a 4B parameter model like Gemma 3n can maintain high semantic quality even under 4-bit quantization, our findings suggest that "proprietary-class" translation performance is now feasible on consumer-grade local hardware, significantly lowering the barrier to professional-quality translation tools.



(a) FLORES-101 (b) OPUS-100

Figure 4: Relationship between model size (billions of parameters) and semantic quality (BERTScore F1) across (a) FLORES-101 and (b) OPUS-100 datasets.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We also acknowledge the financial support from the Brazilian funding agencies CNPq and FAPERGS, and Petrobras. Some experiments in this work used the PCAD infrastructure (http://gppd-hpc.inf.ufrgs.br) at INF/UFRGS. Parts of this manuscript were written with the support of a generative AI tool (ChatGPT); all content was reviewed and validated by the authors.

References

Maximiliana Behnke, Nikolay Bogoychev, Alham Fikri Aji, Kenneth Heafield, Graeme Nail, Qianqian Zhu, Svetlana Tchistiakova, Jelmer van der Linde, Pinzhen Chen, Sidharth Kashyap, and Roman Grundkiewicz. 2021. Efficient machine translation with model pruning and quantization. In *Proceedings of the Sixth Conference on Machine Translation*, pages 775–780, Online. Association for Computational Linguistics.

Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509.

Gemma Team. 2025. Gemma 3n. Accessed November 2025.

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

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.

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

Maja Popović. 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. Association for Computational Linguistics.

Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. 2020. Fully quantized transformer for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1–14, Online. Association for Computational Linguistics.

Jerry Quinn and Miguel Ballesteros. 2018. Pieces of eight: 8-bit neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

3 (Industry Papers), pages 114–120, New Orleans - Louisiana. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.