Lost in Quantization: Activation Outliers Explain Language-Specific FP8 Sensitivity in Llama-3

Guilherme Silva[1], Pedro Silva[2], Matheus Peixoto[1], Gladston Moreira[2] and Eduardo Luz[2]
[1]Postgraduate Program in Computer Science, Federal University of Ouro Preto, Brazil
[2]Computing Department, Federal University of Ouro Preto, Brazil
{guilherme.lopes,matheus.peixoto}@aluno.ufop.edu.br, {silvap,gladston,eduluz}@ufop.edu.br

Abstract

Quantization is key for efficient LLM inference, but its language-specific effects are understudied. We compare INT8 and FP8 (E4M3) quantization for Meta-Llama-3-8B on English and Brazilian Portuguese (PT-BR). INT8 with outlier handling preserves perplexity in both languages, while naive FP8 casting degrades English far more than PT-BR (+18% vs. +3.9%). Activation analysis shows rarer, larger English spikes (> 35) that are more prone to saturation under unscaled E4M3, whereas PT-BR activations are more concentrated. Our FP8 results reflect a naive casting stress test (no calibration/scaling), not an optimized FP8 recipe.

1 Introduction

Large Language Models (LLMs) based on the Transformer architecture (Vaswani et al., 2017) have scaled to billions of parameters, necessitating substantial computational resources that often impede deployment on consumer-grade hardware. To address this, aggressive numerical quantization has become a standard optimization technique (Gholami et al., 2022). While 8-bit integer (*INT8*) quantization significantly reduces memory footprint and latency with minimal performance degradation on standard benchmarks (Dettmers et al., 2022), these results are largely reported in English-centric evaluations (Joshi et al., 2020). Such studies often underrepresent linguistic variability; here we focus on a controlled English vs. Brazilian Portuguese (PT-BR) comparison.

PT-BR is characterized by complex verbal inflection and gender-number agreement, in contrast to the analytic structure of English. These properties not only result in different tokenization efficiencies (Rust et al., 2021), but may also alter the model's internal representation. Transformer-based LLMs exhibit "emergent outliers", that is, dimensions with extreme activation magnitudes that are vital for model performance (Dettmers et al., 2022; Kovaleva et al., 2019). We hypothesize that the distribution of these activation outliers can be language- and domain-dependent. Accordingly, we compare an *outlier-aware* INT8 recipe (LLM.int8()-style handling) against a deliberately *naive* FP8 (E4M3) casting setup to isolate how outlier statistics relate to degradation patterns across languages.

In practice, low-precision inference quality is strongly shaped by (i) how outliers are handled (e.g., decomposition/mixed precision) and (ii) whether FP8 uses calibration and scaling factors to prevent saturation. Production-grade FP8 implementations typically employ scaling (often per-tensor/per-channel or blockwise) to mitigate clipping; in contrast, *pure casting* provides a useful stress test that exposes sensitivity to rare extreme activations (Micikevicius et al., 2022).

The central research question of this paper is: *Do low-precision inference regimes (INT8 with outlier handling vs. naive FP8 E4M3 casting) exhibit different degradation patterns across English and PT-BR, and can these differences be explained by activation-outlier statistics?* We address this by evaluating controlled perplexity on deterministic segments and performing an activation-outlier analysis using forward hooks. We do not report targeted syntactic stress tests in this paper; we discuss them as future work.

The remainder of this work is organized as follows. Section 2 details the modular pipeline designed for deterministic data processing and metric evaluation. Section 3 describes the experimental setup, including the specific quantization regimes applied to the Meta-Llama-3-8B model. Section 4 presents the comparative analysis of perplexity and activation distributions, highlighting the unexpected resilience of PT-BR under FP8. Finally, Section 5 summarizes our findings and discusses the implications of outlier management for multilingual inference.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1044–1048

2 Methodology

This section describes the full methodology pipeline for evaluating the linguistic sensitivity of quantization.

2.1 Model Selection and Configuration

We select the Meta-Llama-3-8B (Dubey et al., 2024) as our primary language model. This choice represents a state-of-the-art, open-weight Transformer of medium scale. The model is initialized using the `Hugging Face Transformers` library with `device_map="auto"` to optimize GPU memory allocation. We evaluate the model under three distinct inference configurations:

- **BF16 (Baseline):** The model is loaded in BFloat16 precision, serving as the high-fidelity reference. This matches the native training precision of Llama-3.
- **INT8:** We utilize the `bitsandbytes` library with the `load_in_8bit` configuration. This applies 8-bit quantization to the weights while dynamically managing activation precision, a standard industry approach for reducing memory.
- **FP8 (E4M3):** We explicitly cast model weights to the 8-bit Floating Point format (E4M3). Inference is performed under `torch.autocast`; importantly, this is a *naive casting* setup without dynamic scaling/calibration factors (commonly used in production FP8), and thus serves as a stress test rather than an optimized FP8 recipe.

2.2 Data Acquisition and Processing

To isolate language-specific degradation, we process raw text data from Project Gutenberg for both English and Brazilian Portuguese (PT-BR). We designed a multistage data pipeline that generates persistent artifacts (`raw`, `processed`, and `segments`) at each step. The workflow proceeds as follows:

1. **Ingestion and Normalization:** The pipeline downloads source texts via specific URLs, automatically stripping metadata headers and footers. To ensure a balanced evaluation between languages, we normalize whitespace and truncate each dataset to a strict budget of 200,000 characters.
2. **Tokenization:** Following validation of minimum length requirements, the normalized text is encoded using the native Meta-Llama-3 tokenizer.
3. **Deterministic Segment Construction:** The tokenized sequence is sliced into fixed-length windows of 2048 tokens. We utilize a fixed random seed ($s = 42$) to extract 64 evenly spaced segments from the available pool. These sequences are serialized to JSON files (`{lang}_segments.json`), ensuring that each quantization regime evaluates the same input data.

3 Experiments

In this section, we detail the experiments to evaluate the impact of numerical precision on linguistic performance. The code and resources associated with this article are publicly available at `https://github.com/silvagal/llama3_fp8_language_sensitivity.git`.

3.1 Datasets and Preprocessing

To ensure high-quality linguistic density, we selected representative literary works from Project Gutenberg (We explicitly treat this as a *literary-domain* case study; see limitations in Section 5.) For Brazilian Portuguese (PT-BR), we utilize *Dom Casmurro* by Machado de Assis (Source ID: 55752)[1]. For English, we utilize *Hamlet* by William Shakespeare (Source ID: 2265)[2].

Following the methodology described in Section 2, the raw text is cleaned and truncated to a strict budget of 200,000 characters per language. Tokenization is performed using the Meta-Llama-3-8B tokenizer, followed by a deterministic split into 64 segments of 2048 tokens using a fixed random seed ($s = 42$). This ensures identical token budgets and segment identities across quantization regimes; however, absolute PPL values are not compared across languages, and we focus on relative degradation trends.

3.2 Evaluation Metrics

In this section, we describe the metrics employed to assess the impact of quantization on model performance and stability.

[1] `https://www.gutenberg.org/cache/epub/55752/pg55752.txt`

[2] `https://www.gutenberg.org/cache/epub/2265/pg2265.txt`

3.2.1 Perplexity Evaluation

Perplexity (PPL) is computed as the primary quantitative metric. For each pre-generated segment of 2048 tokens, we perform an autoregressive forward pass and calculate the negative log-likelihood loss. The final perplexity is reported as the exponential of the mean loss across all segments. Results are persisted to JSON files for subsequent analysis.

Post-hoc Analysis: To compare impacts within each language we analytically derive the degradation (Δ_{Quant}) using the recorded PPL values, as presented in Equation 1:

$$\Delta_{\text{Quant}} = \text{PPL}_{\text{Quant}} - \text{PPL}_{\text{BF16}} \tag{1}$$

Since absolute PPL levels differ by language and domain, cross-language comparisons are reported as *relative increases* (percentage) computed from the same per-language baseline.

Larger within-language Δ_{Quant} (and relative increase) indicates stronger sensitivity to precision reduction under that regime.

3.2.2 KL Divergence for Distributional Drift

To capture deviations in prediction confidence, the pipeline optionally computes the Kullback-Leibler (KL) Divergence between the logits of the reference model (BF16) and the quantized model. This metric is computed per batch and averaged over the evaluation set, thereby isolating distributional shifts that may not be fully captured by the aggregate perplexity score.

3.3 Activation Outlier Analysis

To investigate the root cause of potential degradation, we employ forward hooks on attention layers to record the maximum absolute activation value per batch. These values are aggregated, saved as JSON, and visualized as histograms.

3.4 Experimental Setup

Experiments are conducted using the Meta-Llama-3-8B model. To enable native FP8 acceleration, we utilize NVIDIA H200-class GPUs. The model is loaded in inference-only mode; strictly no fine-tuning, parameter updates, or gradient calculations are performed.

We compare three specific quantization conditions:

- **BF16 (Reference):** High-precision baseline.
- **INT8:** Post-training quantization via `BitsAndBytes`.
- **FP8:** E4M3 format cast (naive casting; no dynamic scaling).

3.5 Evaluation Protocols

We assess the impact of quantization using a multi-metric approach:

1. **Perplexity (PPL):** Serves as the primary measure of language modeling capability, calculated as the exponential of the mean negative log-likelihood loss across the evaluation segments.
2. **KL Divergence:** Quantifies the distributional drift between the logits of the quantized model and the BF16 baseline, isolating precision-induced deviations.
3. **Outlier Analysis:** We generate histograms of maximum activation values in attention layers to empirically verify the preservation (or truncation) of outlier features described in the literature.

4 Results

In this section, we present an evaluation of quantization effects on Brazilian Portuguese (PT-BR) and English, focusing on Perplexity (PPL) stability and activation-outlier distributions.

4.1 Perplexity and Quantization Degradation

Table 1 summarizes the perplexity scores across the three experimental regimes: BF16 (baseline), INT8 (LLM.int8() with outlier separation), and FP8 (E4M3 cast).

Table 1: Perplexity (PPL) comparison. Δ denotes the absolute degradation relative to the BF16 baseline. INT8 inference maintained strict parity, while FP8 introduced measurable degradation.

Method	**English**		**PT-BR**	
	PPL	Δ	PPL	Δ
BF16 (Ref)	3.45	-	13.25	-
INT8	3.45	+0.00	13.25	+0.00
FP8 (E4M3)	4.08	+0.63	13.77	+0.52

Contrary to the common intuition that PT-BR would be inherently more fragile, we observed that the INT8 implementation (utilizing mixed-precision decomposition for outliers) achieved lossless inference for both languages ($\Delta = 0.00$). This confirms the efficacy of explicit outlier-separation

strategies in maintaining performance under compression.

In contrast, the naive FP8 casting (E4M3) degraded performance in both languages. Notably, English exhibited a slightly larger absolute degradation ($\Delta = +0.63$) compared to PT-BR ($\Delta = +0.52$). In relative terms, English perplexity increased by approximately 18%, whereas PT-BR increased by only 3.9%. Because absolute PPL differs across languages and domains, we emphasize these *relative* increases when discussing cross-language sensitivity. This suggests that, in the absence of outlier separation (as in our pure FP8 implementation), the model's performance on English text was paradoxically more sensitive to the reduction in precision than on PT-BR.

4.2 Activation Outlier Analysis

To investigate the source of the FP8 sensitivity, we analyzed the distribution of activation outliers in the attention mechanisms. Figure 1 presents the histogram of maximum activation values per batch.

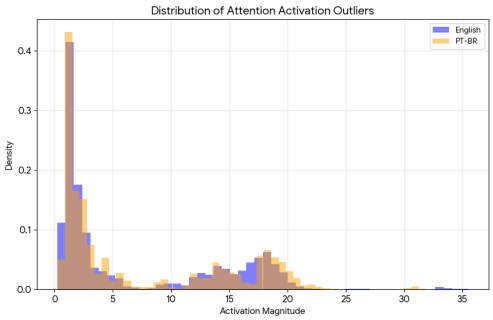

Figure 1: Distribution of maximum attention activations. English (Blue) shows a longer tail reaching extreme values (> 35), while PT-BR (Orange) is more concentrated in the 15–20 range.

As shown in Figure 1, the distribution is heavily unbalanced and bimodal. While the vast majority of activation magnitudes are concentrated in the lower range (peaks around 1.0–3.0), a distinct secondary cluster of "outlier" features appears in the 15.0–20.0 range. These large-magnitude activations dictate the scaling factors in quantization, thereby disproportionately affecting precision. Comparing the languages, English (Blue) exhibits a sparse tail that extends to extreme values (exceeding 35), whereas PT-BR (Orange) shows a higher density accumulation in the 15–20 range, indicating a more consistent presence of high-magnitude features rather than isolated extreme spikes.

Table 2 details the distributional statistics.

Table 2: Statistics of attention activation outliers. While PT-BR has a slightly heavier tail on average (higher P95/P99), English exhibits the highest absolute peaks (Max).

Language	Mean	P95	P99	Max
English	6.97	19.12	21.88	**35.50**
PT-BR	7.23	19.75	22.38	31.88

The analysis reveals distinct distributional characteristics. PT-BR exhibits a heavier average tail, indicated by higher Mean (7.23 vs 6.97) and P99 (22.38 vs 21.88) values. This supports the notion that PT-BR consistently engages high-magnitude features. However, English produced the most extreme individual outliers, with a maximum recorded value of 35.50 compared to 31.88 for PT-BR.

This finding offers a plausible explanation for the higher degradation observed in English under FP8. Under *naive* E4M3 casting (without dynamic scaling), rare extreme spikes are more likely to saturate and incur severe quantization noise, which aligns with the observed 18% perplexity increase in English. PT-BR, despite having more frequent "moderately high" activations, stayed within a slightly safer range, resulting in greater resilience to the E4M3 cast.

5 Conclusion

This study evaluated the impact of INT8 and FP8 quantization on Llama-3-8B across English and Brazilian Portuguese. Our findings yield two primary conclusions. First, INT8 quantization with outlier separation ensures lossless performance regardless of linguistic morphological complexity. Second, naive FP8 (E4M3) quantization disproportionately degrades English compared to PT-BR, with perplexity increases of 18% and 3.9%, respectively.

The analysis confirms that this disparity is driven by the distribution of attention-activation outliers. While PT-BR exhibits a denser concentration of moderately high activations, English relies on sparse, extreme outliers (> 35) that are more likely to saturate under naive E4M3 casting without dynamic scaling, leading to severe quantization noise. These results demonstrate that quantization sensitivity is primarily a function of activation

numerical ranges rather than linguistic morphology. **Limitations.** This paper is restricted to one model (Llama-3-8B), two languages, and a literary-domain dataset with non-matched style/era across languages. Moreover, our FP8 setup is a deliberate *pure casting* configuration (no calibration/scaling factors). Future research should focus on outlier-aware and language-/domain-robust scaling strategies (e.g., calibrated FP8) to preserve rare extreme activations in low-precision formats.

Acknowledgment

This work was supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq - grants 308400/2022-4, 307151/2022-0), *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES - grant 001), and *Fundacão de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, grants APQ-01768-24). We also thank the UFOP/PPGCC for their support.

References

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. 2022. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, and 1 others. 2022. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.