Comparing Modern LLM Quantization Methods Across Natural Languages

Maksym Iakovenko and Stéphane Dupont

University of Mons - MAIA Unit
Mons - Belgium

Abstract. Weight quantization has become a key tool for democratizing access to large language models (LLMs). Despite the technique's growing popularity and potential to aid speakers of diverse languages worldwide, new LLM quantization methods are predominantly validated in monolingual English contexts. This study explores ways to consistently evaluate the multilingual performance of a variety of LLaMA-based models under different quantization configurations. We identify links between the multilingual performance of widely adopted LLM quantization methods and multiple factors such as language's prevalence in the training set and similarity to model's dominant language.

1 Introduction

In recent years, the rise of general purpose capabilities in transformer-based large language models (LLMs) has lead to a wide and rapid adoption of these systems for diverse applications. Among various optimization techniques employed to enable the deployment of LLMs on resource-constrained devices, model compression via weight quantization has risen as an especially prominent approach.

While the outreach of LLMs gradually grows beyond the English-centered contexts and into more multilingual settings, the popular quantization techniques contributing to the proliferation of these models are still mainly validated in terms of their performance on monolingual English-centered evaluations, e.g., measuring the model's perplexity on a generic English-only dataset such as *wikitext2*. Thus, the study of potential biases introduced by these compression techniques remains a severely underrepresented research topic.

This work aims to explore methods for evaluating the performance of quantized LLMs from the LLaMA family of models, in a manner comparable across languages.[1] The sections that follow will first present the relevant context regarding LLM quantization, then explain the experimental setup used for the evaluation, analyze the evaluation results and finally conclude this work by summarizing our findings and discussing possible future avenues of research.

[1]The evaluation code alongside extended evaluation results are available at `https://github.com/MaksymIakovenko/llm_quant_across_langs`.

2 Background and Related Works

2.1 Weight Quantization

Weight quantization aims to convert model parameters into lower precision formats in order to reduce its overall memory footprint. The simplest linear quantization approach can be defined as follows: $Q(x) = \lfloor \frac{x}{S} \rceil - Z$, where Q is the quantization operator, x is the floating-point input, $\lfloor . \rceil$ is the rounding operation which maps the input to the nearest integer value, S is the scaling factor and Z is the zero-point used to re-center the values. To further decrease the reconstruction error, S and Z are often individually attributed to smaller blocks of fixed size within a weight tensor rather than the entire tensor itself.

2.2 LLM Quantization

The high compute requirements of LLMs have lead to the development of dedicated quantization approaches that leverage various properties of LLMs to better preserve the compressed model without costly computations. `LLM.int8()` [1] uses mixed precision in order to protect critical outlier channels within an LLM. GPTQ [2] optimally quantizes the columns of a weight tensor in an iterative manner using forward pass information from a calibration dataset. AWQ [3] optimizes over the scalings of outlier channel weights to improve the reconstruction error of the output embeddings rather than the weights themselves.

2.3 LLM Quantization in Multilingual Settings

To the best of our knowledge, the only other recent work exploring the intersection of LLMs, quantization and multilingual performance is [4]. While this work shares many of the same motivations, it focuses on different families of models, tests different quantization techniques, and prioritizes extensive general evaluation in multilingual contexts rather than trying to isolate the language mastery itself as the main factor to evaluate and compare across model configurations.

3 Methodology

3.1 Quantization methods

To maximize the relevance of the results we focus on assessing widely adopted quantization approaches integrated into the popular Hugging Face `transformers` library, namely GPTQ, AWQ and the `LLM.int8()`-inspired quantization method implemented in `bitsandbytes` (bnb) library. As baselines, these methods will be compared to the full precision models as well as a naive round-to-nearest (rtn) linear quantization. To ensure that no method benefits from a higher bit allocation per individual weight, all of the above quantization methods use 4-bit precision, block size of 128 for the scaling constants, and no zero-point.

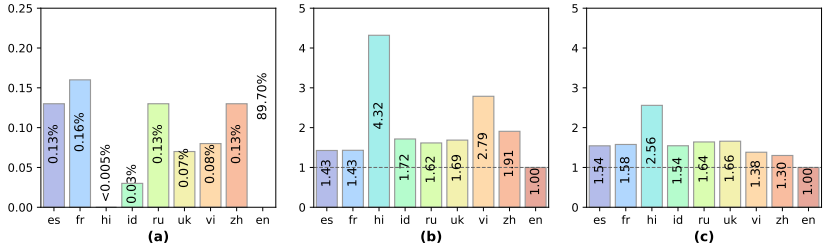

Fig. 1: **(a)** Share of the languages in the pretraining dataset of LLaMA 2 models as given in [5]. **(b)** Tokenization density inefficiency of LLaMA 2 tokenizer and **(c)** LLaMA 3 tokenizer, estimated by tokenizing the corresponding FLORES+ corpus and normalizing resulting token counts by the token count for English.

3.2 Models

We chose to evaluate various iterations of the LLaMA family of models, commonly used both in LLM research as well as in the industry as a whole. More specifically, we will employ `LLaMA 2 7B Chat` from the LLaMA 2 family of models [5] as well as `LLaMA 3.1 8B Instruct` and `LLaMA 3.2 3B Instruct` from the LLaMA 3 family of models [6].

3.3 Evaluation

The core of the proposed evaluation process relies on the FLORES+ [7] translation benchmark dataset, composed of over 2000 English sentences alongside their corresponding high quality human-made translations into over 100 languages. For the practical considerations linked to the available computation time the evaluation was limited to a representative set of 9 languages: English (en), French (fr), Russian (ru), Ukrainian (uk), Spanish (es), Vietnamese (vi), Indonesian (id), Simplified Chinese (zh) and Hindi (hi).

The translations were produced by prompting the models to translate the samples from English to a target language and vice-versa with temperature set to 0. The models' performance across languages was measured on the `devtest` split of FLORES+ using the SacreBLEU[8] implementation of the BLEU score with the FLORES101 tokenizer, evaluated on a per-sentence level.

Additionally, a modified perplexity formulation, aimed at better comparability across languages, was evaluated over the `dev` split of FLORES+. Regular perplexity may face challenges when applied to tokenization schemes where different languages have varying token densities per unit of information, such as in the case of LLaMA models as shown in Fig. 1. In languages that are tokenized less densely, a given context window contains less semantic information compared to languages with denser tokenization. This can lead to artificially reduced perplexity for less densely tokenized languages, as fixed-size sequences of this kind contain less concrete information and thus are easier to model. To address this, we employ the following modified perplexity evaluation formula:

	Lang.	es	fr	hi	id	ru	uk	vi	zh	en
LLaMA 2 7B Chat	awq	1.65%	1.90%	36.15%	8.04%	8.43%	9.59%	12.31%	8.25%	1.98%
	gptq	4.98%	1.65%	36.00%	11.53%	10.22%	22.58%	22.29%	9.95%	1.51%
	bnb	7.97%	6.57%	76.16%	17.87%	12.59%	17.14%	24.08%	17.30%	3.08%
	rtn	4.35%	2.81%	59.86%	11.65%	9.19%	11.93%	22.15%	14.47%	2.51%
LLaMA 3.1 8B Instruct	awq	2.73%	2.32%	12.91%	9.21%	9.97%	9.49%	2.83%	7.79%	0.52%
	gptq	10.24%	11.09%	44.08%	17.24%	26.35%	30.14%	11.98%	17.74%	10.30%
	bnb	9.77%	11.18%	26.17%	18.33%	17.43%	24.80%	19.12%	16.17%	4.84%
	rtn	4.77%	4.09%	17.95%	9.73%	11.26%	15.71%	7.30%	12.71%	2.95%
LLaMA 3.2 3B Instruct	awq	5.70%	3.90%	13.22%	7.63%	15.66%	28.13%	7.07%	8.90%	0.81%
	gptq	12.19%	14.23%	41.02%	16.57%	39.02%	73.65%	15.13%	18.97%	1.89%
	bnb	14.67%	19.46%	27.16%	22.76%	33.93%	76.09%	18.80%	25.22%	4.07%
	rtn	10.86%	11.95%	20.22%	14.85%	22.58%	49.94%	14.64%	16.01%	3.50%

Table 1: Relative increase in perplexity per language and model configuration.

$$ppl_l(W_{lang}) = \frac{1}{N}\sum_{n=1}^{N} exp\left(-\frac{1}{|W_{en}^n|}\sum_{i}^{pos(W_{lang}^n)} log(p(w_i|w_{i-C:i-1}))\right)$$

Here W_{lang}^n represents the n^{th} sample in the dataset for a language $lang$, with $|W_{lang}^n|$ denoting its token length. N stands for the total number of samples, while w_i is the i^{th} token in the concatenated sequence of all samples. The function $pos(W_{lang}^n)$ returns the list of position indices for all tokens in the given sample within the whole sequence, and C represents the remaining number of free tokens within the context window.

This formulation leverages the uniformity of FLORES+ in terms of real information contents per sample across languages, thus enabling the normalization of perplexity with respect to a common quantity per sample, here the token lengths of each corresponding English sample were chosen as baseline.

4 Results

We primarily focus on comparing the quantized models' performance across languages, as such, we will mainly analyze the results expressed in terms of how

Lang.	es	fr	hi	id	ru	uk	vi	zh	mean
Translating from English to target language									
awq	-2.00%	-3.96%	-13.24%	-5.29%	-6.28%	-10.12%	-9.32%	-2.86%	-6.63%
gptq	-3.37%	-3.12%	-14.76%	-9.85%	-9.88%	-10.01%	-8.11%	-6.35%	-8.18%
bnb	-1.72%	-4.12%	-21.15%	-10.15%	-3.95%	-11.95%	-6.73%	-5.28%	-8.13%
rtn	-4.43%	-4.88%	-16.36%	-8.02%	-5.20%	-12.22%	-9.15%	-4.16%	-8.05%
mean	-2.88%	-4.02%	-16.38%	-8.33%	-6.33%	-11.07%	-8.33%	-4.66%	-7.75%
Translating from target language to English									
awq	-1.12%	-1.57%	-12.73%	-4.10%	-0.23%	-3.19%	-3.97%	-6.22%	-4.14%
gptq	-0.91%	0.03%	-19.17%	-0.18%	-1.24%	-3.49%	-4.84%	-5.73%	-4.44%
bnb	-1.08%	-0.10%	-22.02%	-3.00%	-0.41%	-2.33%	-6.26%	-2.54%	-4.72%
rtn	-3.13%	-1.88%	-19.92%	-7.50%	-2.49%	-5.28%	-7.83%	-4.58%	-6.58%
mean	-1.56%	-0.88%	-18.46%	-3.70%	-1.09%	-3.57%	-5.73%	-4.77%	-4.97%
overall	-2.22%	-2.45%	-17.42%	-6.01%	-3.71%	-7.32%	-7.03%	-4.72%	-6.36%

Table 2: Relative drop in BLEU scores for FLORES+ translation for `LLaMA 2 7B Chat`.

Lang.	es	fr	hi	id	ru	uk	vi	zh	mean
Translating from English to target language									
awq	-2.63%	-1.90%	-4.81%	-4.35%	-4.16%	-0.15%	-1.51%	-5.08%	-3.07%
gptq	-2.08%	-4.41%	-9.93%	-9.42%	-7.70%	-8.37%	-6.03%	-21.55%	-8.69%
bnb	-3.49%	-4.67%	-7.81%	-4.44%	-5.68%	-3.09%	-4.54%	-16.10%	-6.23%
rtn	-2.03%	-3.08%	-5.87%	-4.31%	-3.50%	-6.45%	-3.85%	0.29%	-3.60%
mean	-2.56%	-3.52%	-7.10%	-5.63%	-5.26%	-4.51%	-3.98%	-10.61%	-5.40%
Translating from target language to English									
awq	-3.93%	-2.24%	-5.05%	-2.76%	-3.76%	-3.66%	-2.90%	-2.08%	-3.30%
gptq	-4.23%	-3.13%	-8.91%	-4.38%	-3.36%	-4.84%	-5.62%	-2.34%	-4.60%
bnb	-3.01%	-2.39%	-7.50%	-3.42%	-2.83%	-3.91%	-3.86%	-2.43%	-3.67%
rtn	-2.13%	-2.59%	-4.32%	-2.76%	-2.49%	-0.68%	-2.47%	-1.27%	-2.34%
mean	-3.32%	-2.59%	-6.44%	-3.33%	-3.11%	-3.27%	-3.71%	-2.03%	-3.48%
overall	-2.94%	-3.05%	-6.77%	-4.48%	-4.19%	-3.89%	-3.85%	-6.32%	-4.44%

Table 3: Relative drop in BLEU scores for FLORES+ translation for `LLaMA 3.1 8B Instruct`.

much the score of a quantized model differs from results of the original model for a given language. All the showcased results will be expressed as the offset between the quantized and original model score, normalized by the original model score, shown as percentages for clarity. This evens out the scores that vary greatly in absolute terms from one language to another.

The relative degradation of the performance is not the same across all languages. The majority of results favor English compared to other languages as seen in the perplexity in Table 1 as well as the results of translating from target language to English being better than the other way around seen in Tables 2, 3 and 4. Furthermore, Latin script-based European languages such as French and Spanish (which share a larger proportion of typical tokens with English) also perform significantly better than other languages.

The prevalence of a language in the pretraining corpus, as well as the tokenization density for the language also seem to play a role. In the case of `LLaMA 2 7B`, Hindi experiences extreme performance drops while both being severely underrepresented in the pretraining corpus and having very low tokenization density for the model. However, this is not the only explaining factor as for example the performance drop for Spanish is lesser than for Chinese despite both

Lang.	es	fr	hi	id	ru	uk	vi	zh	mean
Translating from English to target language									
awq	-2.32%	-1.46%	-6.33%	-3.56%	-8.41%	-9.18%	-3.16%	-6.24%	-5.08%
gptq	-3.35%	-5.84%	-12.52%	-9.46%	-11.86%	-27.45%	-7.92%	-13.30%	-11.46%
bnb	-2.67%	-5.98%	-9.66%	-8.47%	-13.42%	-19.81%	-7.01%	-12.31%	-9.92%
rtn	-1.65%	-3.03%	-5.94%	-5.34%	-6.36%	-11.53%	-4.53%	-14.91%	-6.66%
mean	-2.50%	-4.08%	-8.61%	-6.71%	-10.01%	-16.99%	-5.66%	-11.69%	-8.28%
Translating from target language to English									
awq	-1.05%	-0.75%	-4.70%	-3.01%	-1.52%	-4.70%	-1.29%	-3.38%	-2.55%
gptq	-3.91%	-3.77%	-10.72%	-6.52%	-7.49%	-10.12%	-5.68%	-7.56%	-6.97%
bnb	-0.85%	-2.05%	-9.21%	-6.05%	-3.60%	-7.38%	-5.77%	-5.39%	-5.04%
rtn	-3.18%	-1.27%	-4.99%	-2.82%	-3.46%	-5.10%	-3.24%	-3.35%	-3.43%
mean	-2.25%	-1.96%	-7.40%	-4.60%	-4.02%	-6.82%	-3.99%	-4.92%	-4.50%
overall	-2.37%	-3.02%	-8.01%	-5.65%	-7.02%	-11.91%	-4.82%	-8.31%	-6.39%

Table 4: Relative drop in BLEU scores for FLORES+ translation for `LLaMA 3.2 3B Instruct`.

languages being present in the same proportion in the pretraining dataset.

Among the evaluated quantization methods AWQ performs the best overall. The gap is more pronounced for the adjusted perplexity results, yet it is still present when comparing the BLEU scores. Surprisingly, the round-to-nearest quantization results stay competitive with the dedicated approaches, particularly when examining the scores for the more recent LLaMA 3 family of models.

5 Conclusion

In summary, we perform an adjusted multilingual evaluation of a series of modern LLMs under different popular quantization configurations. Through these evaluations, our work has further validated the hypothesis that current quantization methods favor the LLM's dominant language, alongside languages close to it. Furthermore we have identified potential additional factors influencing multilingual performance degradation, namely the share of the language in the pretraining corpus as well as tokenization density. Our results have also shown that AWQ has the least amount of negative impact on multilingual performance.

Our results are limited in terms of the model architecture coverage as well as language representation. As such future work could expand upon the existing observations by extending the evaluation to more models and more languages. Furthermore, it may prove interesting to explore metrics centered around the grammatical fluency of the generated outputs, as qualitative evaluation of model outputs has shown that quantization may lead to degradation of the text style which is harder to capture with metrics used in this study.

References

[1] Tim Dettmers et al. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[2] Elias Frantar et al. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[3] Ji Lin et al. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa, editors, *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100, 2024.

[4] Kelly Marchisio et al. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv preprint: `arXiv:2307.09288`.

[6] Aaron Grattafiori et al. The llama 3 herd of models, 2024. arXiv preprint: `arXiv:2407.21783`.

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

[8] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.