How Does Quantization Affect Multilingual LLMs?

Anonymous ACL submission

Abstract

Quantization techniques are widely used to improve inference speed and deployment of large language models. While a wide body of work examines the impact of quantized LLMs on English tasks, none have examined the effect of quantization across languages. We conduct a thorough analysis of quantized multilingual LLMs, focusing on their performance across languages and at varying scales. We use automatic benchmarks, LLM-as-a-Judge methods, and human evaluation, finding that (1) harmful effects of quantization are apparent in human evaluation, and automatic metrics severely underestimate the detriment: a 1.7% average drop in Japanese across automatic tasks corresponds to a 16.0% drop reported by human evaluators on realistic prompts; (2) languages are disparately affected by quantization, with non-Latin script languages impacted worst; and (3) challenging tasks such as mathematical reasoning degrade fastest. As the ability to serve low-compute models is critical for wide global adoption of NLP technologies, our results urge consideration of multilingual performance as a key evaluation criterion for efficient models.

1 Introduction

Multilingual large language models (LLMs) have the power to bring modern language technology to the world, but only if they are cheap and reliable. Known as the *low-resource double bind*, underserved languages and severe compute constraints often geographically co-occur (Ahia et al., 2021), meaning that for wide adoption, multilingual LLMs must be highly-performant *and* lightweight.

With the shift towards large models, quantization is a widely adopted technique to reduce cost, improve inference speed, and enable wider deployment of LLMs. Work on quantization, however, is by-and-large evaluated in English only (e.g. Xiao et al., 2023; Ahmadian et al., 2024; Frantar et al., 2022). No works to our knowledge have characterized the impact of quantization on the multilingual generation capabilities expected from modern LLMs. Ubiquitous use of compression techniques in the real world drives urgency to the question *how are multilingual models impacted?*

Our question is timely, given recent work showing that compression techniques such as quantization and sparsity amplify disparate treatment of long-tail features, which may have implications for under-represented languages in multilingual LLMs (Hooker et al., 2019, 2020; Ahia et al., 2021; Ogueji et al., 2022). Indeed, many model designs choices implicitly overfit to a handful of resource rich languages: from tokenizer choice, to weighting of training data, and to widely-used quantization techniques. Focusing on a small subset of high-resource languages in design degrades model performance for overlooked languages (Schwartz et al., 2022; Kotek et al., 2023; Khandelwal et al., 2023; Vashishtha et al., 2023; Khondaker et al., 2023; Pozzobon et al., 2024), introduces security vulnerabilities (Yong et al., 2023; Nasr et al., 2023; Li et al., 2023a; Lukas et al., 2023; Deng et al., 2023), and unfairly passes high costs to non-English users faced with high latency (Held et al., 2023; Durmus et al., 2023; Nicholas and Bhatia, 2023; Ojo et al., 2023; Ahia et al., 2023).

We analyze four state-of-the-art multilingual LLMs across 3 different sizes ranging from 8 to 103 billion parameters and covering up to 23 languages, under various quantization techniques. Critically, it is vital that we move beyond automatic evaluation and gather *real human feedback on performance cost*. We thus perform multilingual human evaluation on challenging real-world prompts in addition to LLM-as-a-Judge and evaluation on standard automatic benchmarks such as multilingual MMLU (Hendrycks et al., 2020), MGSM (Shi et al., 2023), and FLORES-200 (Costa-jussà et al., 2022a). Across experimental set-ups we find that:

1. **Automatic metrics severely underestimate damage from quantization.** Automatic evaluations estimate performance deterioration relative to FP16 across tasks at -0.3% (French) and -1.7% (Japanese) vs -16.6% and -16.0% reported by human evaluators.
2. **Quantization affects languages differently.** Non-Latin script languages are more greatly harmed on average. Across tasks, Latin-script languages scored -0.7% relative to FP16 for a 103B parameter model while non-Latin scripts scored -1.9%. For a smaller 8-billion parameter model, drops were -3.0% vs. -3.7%.
3. **Challenging tasks degrade fastest.** Mathematical reasoning (-13.1%), performance on real-world challenging prompts judged by humans (-10.5%), and LLM-as-a-Judge (-25.9%) are severely reduced.
4. **Occasionally, quantization brings benefits.** Similar to Ahia et al. (2021) and Ogueji et al. (2022) on sparsity, we find that quantization *benefits* model performance in some cases: e.g., an average 1.3% boost across tasks for a 35B model quantized with W8A8.

As the first to broadly study the impact of quantization on multilingual LLMs, our work is part of a wider body of literature that considers the impact of model design choices on downstream performance. Our results urge attention to multilingual performance at all stages of system design.

2 Background

Quantization compresses the weights and potentially activations of a neural network to lower-bit representations. Compression can be done by training the model at lower precision, known as Quantization Aware Training (QAT), or performed on the final model weights, known as Post Training Quantization (PTQ). Given the difficulties in training LLMs especially at precision lower than 16-bits floating point, PTQ methods which perform the quantization single-shot without needing gradient updates are highly desirable. Training is completed at higher precision, then weights/activations are quantized without further training. In this work, we focus on *post-training quantization* because of its simplicity and applicability at scale. PTQ of LLMs can be further categorized into:

Weight-Only Quantization Weight matrices are quantized offline and the compressed matrices are loaded from memory during inference. Quantized weight matrices have a smaller memory footprint compared to FP16 ($2\times$ smaller for 8-bit and almost $4\times$ smaller for 4-bit), enabling inference with less compute. In memory-bound scenarios, it also enables faster inference due to fewer bytes transferred from GPU memory to the compute units.

For a weight matrix $\mathbf{W} \in \mathbb{R}^{d_{in} \times d_{out}}$ and input $\mathbf{X} \in \mathbb{R}^{seq \times d_{in}}$, if only a single scaling factor is used for naive quantization (per-tensor), then the quantized weights are given by:

$$\mathbf{W}_Q = \Delta \cdot \left\lfloor \frac{\mathbf{W}}{\Delta} \right\rceil, \quad \Delta = \frac{\max(|\mathbf{W}|)}{2^{N-1}} \quad (1)$$

where $\Delta \in \mathbb{R}$ denotes the scale, N the bit precision, $|.|$ the absolute value over each element in $\mathbf{W}$ and $\lfloor . \rceil$ rounding to the nearest integer.

A single scaling factor might not be enough if the distribution of parameters in the weight matrix has high variance; thus one could increase the granularity of quantization by using a scale for each output dimension (per-column), i.e., $\Delta \in \mathbb{R}^{d_{out}}$. However, when N is aggressively lowered to 4 bits or lower, even per-column granularity might be insufficient to cover the range of values in a column. The granularity can be further increased by using a shared scale for a subset of input dimensions called groups (g), thus the scale $\Delta \in \mathbb{R}^{\frac{d_{in}}{g} \times d_{out}}$. A commonly used group size is 128.

Equation 1 gives the simplest way to quantize the weights. For $N \leq 4$ bits, using more advanced Weight-Only Quantization methods like GPTQ (Frantar et al., 2022) or AWQ (Lin et al., 2024) leads to better downstream performance.

Weight-and-Activation Quantization As the name suggests, Weight-and-Activation Quantization quantizes the model activations alongside the weights. Unlike Weight-Only Quantization where weights can be quantized offline, quantization of activations happens at runtime. One could compute the quantization scales for various activations by using a small slice of training or validation data (static scaling) but this method typically has large degradation (Xiao et al., 2023). For minimal degradation, it is preferred to calculate the quantization scaling factor dynamically (dynamic scaling) for each input on-the-fly. While quantizing activations is more difficult, reducing the precision of the activations alongside the weights enables the usage of specialized low-precision matrix multiplication hardware in modern GPUs leading to up to 2× improvement in throughput. For a weight matrix $\mathbf{W} \in \mathbb{R}^{d_{in} \times d_{out}}$ and input $\mathbf{X} \in \mathbb{R}^{seq \times d_{in}}$, naive Weight-and-Activation Quantization with per-token input granularity and per-column weight granularity can be done by:

$$\mathbf{W}_{Q_{:,j}} = \left\lfloor \frac{\mathbf{W}_{:,j}}{\Delta^{W}_{:,j}} \right\rceil, \Delta^{W}_{:,j} = \frac{\max(|\mathbf{W}_{:,j}|)}{2^{N-1}} \quad (2)$$

$$\mathbf{X}_{Q_{i,:}} = \left\lfloor \frac{\mathbf{X}_{i,:}}{\Delta^{X}_{i,:}} \right\rceil, \Delta^{X}_{i,:} = \frac{\max(|\mathbf{X}_{i,:}|)}{2^{N-1}} \quad (3)$$

$$Y = \Delta^{X} \odot (\mathbf{X}_Q \mathbf{W}_Q) \odot \Delta^{W} \quad (4)$$

where, $\Delta^W \in \mathbb{R}^{d_{out}}$, $\Delta^X \in \mathbb{R}^{seq}$ and $\odot$ denotes element-wise multiplication by broadcasting the elements to match the shape of the operands.

3 Experiment Set-up

Models We evaluate Command R+[1], Command R[2], and Aya 23 models (Aryabumi et al., 2024) as representatives of state-of-the-art multilingual LLMs. Command models are 103 and 35 billion parameters, and Aya 23 models are 35 and 8 billion parameters, respectively. We quantize the models using the open weights on HuggingFace.

Quantization For Command R and R+, we evaluate both **weight-only quantization** at 8-bit (**W8** with per-column scaling) and 4-bit (**W4-g** with group-wise scaling using GPTQ (Frantar et al., 2022)), as well as **weight-and-activation quantization** at 8-bit (**W8A8** with per-column scaling for weights and per-token scaling for activations).

Ahmadian et al. (2024) show that if the model is trained with the right hyper-parameters, naive Weight-and-Activation Quantization has minimal degradation. Otherwise, one may leverage SmoothQuant (Xiao et al., 2023) to smoothen the distribution of activations making them more amenable to quantization. We therefore also explore **W8A8-SmoothQuant** (a W8A8 variant with SmoothQuant) for Command R+ as well as a 4-bit weight-only quantized variant with column-wise scaling (**W4**) to understand the impact of scaling granularity at extremely low-bit precision. Following (Frantar et al., 2022; Xiao et al., 2023), we use 128 English samples for calibration for SmoothQuant and GPTQ.

For Aya 23 8B and 35B, we use bitsandbytes[3] to obtain 8-bit and 4-bit quantized models. Bitsandbytes uses LLM.int8() (Dettmers et al., 2022)—similar to W8A8 described above except that it performs certain computations in FP16. Bitsandbytes 4-bit uses the NF4 datatype (Dettmers et al., 2023) to perform Quantile Quantization which limits degradation at the expense of inference speedups.

3.1 Automatic Evaluation

We evaluate in 10 primary languages: *Arabic*, *French*, *German*, *English*, *Spanish*, *Italian*, *Portuguese*, *Korean*, *Japanese*, and *Chinese*. Quantized models are compared to the original **FP16** versions, and we primarily report results as **relative degradation** compared to this FP16 baseline:

$$\%\Delta = \frac{\text{score}_{\text{quantized}} - \text{score}_{\text{FP16}}}{\text{score}_{\text{FP16}}} * 100 \quad (5)$$

Raw numeric results are in the Appendix. Results are averaged over 5 runs.[4]

Multilingual MMLU (mMMLU) This multi-domain question answering task consists of 14,000+ multiple-choice questions. We translate MMLU (Hendrycks et al., 2020) to 9 languages with Google Translate and refer to this version as mMMLU. We measure accuracy in a 5-shot setting. An example is in Table A1.

MGSM (Shi et al., 2023) MGSM is a generative mathematics evaluation set manually translated from GSM8K (Cobbe et al., 2021). Of our target languages, it is available for German, Spanish, French, Japanese, and Chinese. We report accuracy over the 250-item test set for each language.

FLORES-200 (Costa-jussà et al., 2022b) This well-known multi-way parallel test set evaluates translation capabilities. We translate into and out of English, and report SacreBLEU (Post, 2018).

Language Confusion (*Under Review et al.*, 2024) These test sets assess a model's ability to respond in a user's desired language. In the monolingual setting, prompts are in language l and the model must respond in language l. For instance, a user prompts in Arabic, so implicitly desires an Arabic response. In the cross-lingual variant, a prompt is provided in English but the user requests output in a different language l'.[5] fastText (Joulin et al.,

[1] https://docs.cohere.com/docs/command-r-plus
[2] https://docs.cohere.com/docs/command-r
[3] https://github.com/TimDettmers/bitsandbytes

[4] k=0, p=0.75, temperature=0.3, except mMMLU, which, as a QA eval, is run deterministically with t=0.

[5] An example from the *Okapi* subsection of the evaluation is: "Reply in Spanish. Explain a common misconception about your topic. Topic: Using AI to Augment Human Capabilities"

2016) language identificaion is run over the output. We report *line-level pass rate* (**LPR**), i.e., the percentage of responses for which all lines in the response are in the user's desired language.

Aya Evaluation Aya 23 models are evaluated using an extended version of the Aya evaluation setup (Aryabumi et al., 2024) using the unseen discriminative tasks (XWinograd (Muennighoff et al., 2023), XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2022)), mMMLU (Okapi; Dac Lai et al., 2023), MGSM, and Belebele (Bandarkar et al., 2023) from `eval-harness` (Gao et al., 2023).[6] We evaluate models on languages included in the covered 23 languages, except for the unseen tasks where we use all available languages.[7] Aya evaluations allow us to add: *Czech, Greek, Hebrew, Hindi, Indonesian, Dutch, Persian, Polish, Romanian, Russian, Turkish, Ukrainian, Vietnamese*.

3.2 Human Evaluation

We run human evaluation in *Spanish*, *French*, *Korean*, and *Japanese*.

Internal Evaluation Suite 150 diverse prompts designed to be more complex than public evaluation benchmarks. As such, we expect greater degradation with increased quantization given the difficulty of the samples. Prompts for all four languages are translated by humans from an English seed prompt, ensuring that respective language-specific subsets share the same prompts.

Aya Dolly-200 (Singh et al., 2024) We use multilingual data from the Aya Evaluation Suite to assess open-ended generation capabilities. For Korean and Japanese, we use prompts from the Aya Dolly-200 test set (**dolly-machine-translated**), which are automatically translated from English Dolly-15k (Conover et al., 2023) then human-curated to avoid references requiring specific cultural or geographic knowledge. For French and Spanish, we use **dolly-human-edited**, a human post-edited version of **dolly-machine-translated**. For each language, we evaluate using the first 150 prompts.

[6]We follow the setup used by Üstün et al. (2024): each evaluation is run once, and for FLORES, no sampling is used and metric is spBLEU.

[7]mMMLU: ar, de, es, fr, hi, id, it, nl, pt, ro, ru, uk, vi, zh. MGSM: de, es, fr, ja, ru, zh. Belebele: ar, cs, de, es, el, fr, hi, id, it, ja, ko, nl, fa, pl, pt, ro, ru, tr, uk, vi, zh. FLORES: ar, cs, zh, nl, fr, de, el, he, hi, id, it, ja, ko, fa, pl, pt, ro, ru, es, tr, uk, vi.

Annotator Statistics Annotations and translations were completed by native-level speakers of the respective languages, each of whom is also fluent in English. Annotators were paid by the hour, with compensation above the federal minimum wage of the country of employment.

Annotation Interface We use a pairwise evaluation setup. Annotators see a prompt and two (shuffled) completions of the FP16 model and a quantized variant. They rate each response on a 5-point Likert scale, then express a preference between the two model outputs (tie, weak preference, strong preference). We encourage annotators to avoid tied rankings. Win rates are based on the ranking preferences alone.

3.3 LLM/RM-as-a-Judge

Because human evaluation is costly and time-intensive, it is common to use an "LLM-as-a-Judge" to rate model completions (e.g. Li et al., 2023b; Zheng et al., 2023). Reward models (RMs) can also simulate human preference. An RM scores multiple completions given the same prompt, and the prompt-completion pair with the higher score is deemed preferred. We call this *RM-as-a-Judge*.

We assess quantized model outputs using LLM- and RM-as-a-Judge. In the former, an LLM selects a preferred response from a <`instruction, modelA_completion, modelB_completion`> tuple (see Table A2). Following (Üstün et al., 2024; Aryabumi et al., 2024) we use GPT-4[8] as an LLM proxy judge. To minimize bias, we randomize the order of model outputs. For RM-as-a-Judge, a multilingual reward model scores <`prompt, completion`> pairs for each model output, over which we calculate win-rate. We report win-rates of quantized models versus the FP16 baseline.

We assess the outputs of quantized models over the *Internal Evaluation Suite* and *Aya Dolly-200* described in Section 3.2. We use the same prompt and completion pairs as in human evaluation, which provides the ability to relate LLM/RM-as-a-Judge performance with human evaluation.

4 Results

To clearly see the many-faceted impact of quantization, we discuss our results by quantization level (§4.1), by task (§4.2), by language (§4.3), by model

[8]Specifically, `gpt-4-turbo` (`gpt-4-1106-preview`): `https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`

size (§4.4), and by quantization strategy (§4.5). We then report LLM-as-a-Judge and RM-as-a-Judge (§4.6) and human evaluation results (§4.7).

4.1 By Quantization Level

Command R and R+ In Table 1, we aggregate results of each metric for each level of quantization. We average scores across languages, then calculate the relative percentage drop from FP16.[9] We discuss results of **W8**, **W8A8**, and **W4-g** quantization, which are variants available for both Command model sizes. Most results follow intuition: greater quantization leads to larger performance degradation: -0.2% for W8, -0.8% for W8A8, and -0.9% overall for the 103B model. An exception is W8A8 for the 35B, which experiences a slight boost overall due to higher performance on translation and language confusion evaluations.

Aya 23 Models Table 2 shows the aggregated results for Aya 23 models on the extended Aya evaluations (Aryabumi et al., 2024) at **W8**, and **W4** quantization. We find a similar trend with Command models where W4 often leads to a larger relative drop compared to W8 variants, consistent across tasks and languages. W8, however, does not substantially drop performance in any task.

4.2 By Task

Mathematical reasoning as measured by MGSM is strikingly affected by quantization. Relative performance of the 35B W4-g model is a dismal -13.1% on average, with as poor as -17.3% in Chinese (Table A3). MGSM and Belebele are the two tasks with the highest performance drop for Aya 23 models at W4 quantization drops of 7.5% and 8.5% on the 8B model, respectively, followed by mMMLU.

On FLORES, relative drops are sensitive to translation direction: the more challenging EN→L2 is impacted more than L2→EN (-1.4% vs -1.2% for Aya 23 35B, -1.8% vs -1.0% for Aya 23 8B). The same effect is observed in most Command models. Quantization does not noticeably impact unseen discriminative tasks (XWinograd, XCOPA, XStoryCloze: Table A4). Full raw and relative results by task are in the Appendix.

Curiously, there are some fleeting performance boosts: an increase of 1.8–2.1% on MGSM and mild improvements on FLORES with W8 on Aya models and a similar translation boost of the 35B Command model at W8A8. Quantization has no effect or causes a mild improvement on the monolingual language confusion task (except for 35B W4-g), and cross-lingual language confusion performance is boosted with greater quantization.

4.3 By Language

We next ask, *Are languages differently affected by quantization?* Table 3 shows performance averaged over mMMLU, FLORES, and Language Confusion tasks, with Table 4 further including MGSM for supported languages. Metrics are on different scales, so we average relative change (%Δ) rather than raw scores.[10] We further separate into languages written in the Latin/Roman script, which also are the subset of Indo-European languages (**Ltn/IE**) versus those using other scripts (¬**Ltn/IE**). W4-g is severely degrading across languages for the 35B Command model, and a relationship between language and performance is apparent: ¬**Ltn/IE** languages typically degrade more. Chinese and Korean are particularly harmed at W4. The effect is seen consistently across all automatic metrics for Command models, with limited exception. Table 5 is discussed more thoroughly in Section 4.5, but also shows this discrepancy.

Interestingly, W8A8 quantization of the 35B model seems to *help* on average across all languages. The magnitude is primarily due to W8A8 helping on cross-lingual language confusion.

4.4 By Model Size

Across evaluations at the most extreme quantization (W4/W4-g), smaller models are more sensitive: W4-g variants of 103B and 35B Command record -0.9% and -2.8% performance relative to FP16 on average, with a stark difference of -2.9% vs. -13.1% on MGSM. Aya 23 35B/8B record -2.8%/-3.7% on average, with their largest gap occurring in Belebele (-5.9% vs. -8.5%).

4.5 By Quantization Strategy

We evaluate variants of the 103B Command model with SmoothQuant (**W8A8-sq**), and a more naive **W4** variant using per-column quantization instead of group-wise scaling. We compare **W8A8-sq** to

[9] Ex. For 103B W4-g MGSM, scores were: {de: 71.2, es: 75.7, fr: 69.0, ja: 58.0, zh: 68.9}, thus the average score was 68.6—a 2.9% drop from FP16 ($\frac{68.6-70.6}{70.6} = -0.029$).

[10] Ex. to arrive at -1.3% for 103B W8A8 in Arabic, we average relative performance for mMMLU, FLORES En↔L2, and Language Confusion tasks: $\text{avg}(\{-2.2\%, -1.0\%, -1.3\%, 0.0\%, -1.8\%\}) = -1.3\%$.

		Avg. Rel. %Δ	mMMLU		MGSM		FLORES L2→En		FLORES En→L2		Language Confusion Monolingual		Language Confusion Cross-lingual	
103B	FP16	-	66.7	-	70.6	-	37.7	-	39.6	-	99.2	-	91.5	-
	W8	-0.2%	66.7	0.0%	69.9	-1.0%	37.7	0.0%	39.6	0.0%	99.2	0.0%	91.2	-0.3%
	W8A8-sq	-0.5%	66.3	-0.5%	69.5	-1.6%	37.8	0.2%	39.1	-1.3%	99.2	0.0%	91.5	0.1%
	W8A8	-0.8%	65.6	-1.7%	69.8	-1.1%	37.7	0.0%	39.1	-1.2%	99.4	0.2%	90.4	-1.2%
	W4-g	-0.9%	65.7	-1.4%	68.6	-2.9%	37.8	0.4%	39.4	-0.5%	99.2	0.0%	90.5	-1.1%
	W4	-2.5%	63.8	-4.3%	64.4	-8.8%	37.1	-1.6%	39.0	-1.6%	99.3	0.1%	92.8	1.4%
35B	FP16	-	59.4	-	49.8	-	32.4	-	35.5	-	98.7	-	66.5	-
	W8	-0.2%	59.3	-0.1%	49.4	-0.7%	32.3	-0.2%	35.4	-0.2%	98.8	0.1%	66.3	-0.2%
	W8A8	0.2%	59.3	-0.2%	47.1	-5.5%	32.9	1.6%	35.8	0.9%	99.0	0.3%	68.9	3.7%
	W4-g	-2.8%	58.2	-2.0%	43.3	-13.1%	31.7	-1.9%	35.3	-0.7%	98.3	-0.4%	67.1	1.0%

Table 1: **Per-dataset average performance across non-English languages for 103B and 35B Command models at varying levels of quantization.** %Δ the relative performance vs. FP16 [ex., for MGSM at W4-g on the 35B: $\frac{43.3-49.8}{49.8} * 100 = -13.1\%$.] Languages: ar, de, es, fr, it, ja, ko, pt, zh; except MGSM: de, es, fr, ja, zh. Any discrepancy is due to rounding: raw scores and **%Δ** were calculated at full precision.

		Avg. Rel. %Δ	Unseen Tasks ↑		mMMLU (Okapi) ↑		MGSM ↑		Belebele ↑		FLORES L2→En ↑		FLORES En→L2 ↑	
Aya 23 35B	FP16	-	70.8	-	58.2	-	51.2	-	77.6	-	42.9	-	37.8	-
	W8	0.1%	70.6	-0.2%	57.9	-0.5%	52.1	1.8%	77.1	-0.6%	43.0	0.1%	37.9	0.3%
	W4	-2.8%	70.5	-0.3%	56.6	-2.3%	48.1	-6.0%	73.0	-5.9%	42.4	-1.2%	37.2	-1.4%
Aya 23 8B	FP16	-	67.6	-	48.2	-	34.7	-	64.8	-	39.8	-	34.8	-
	W8	0.3%	67.6	0.1%	47.8	-0.9%	35.4	2.1%	64.6	-0.3%	39.7	0.5%	34.8	0.2%
	W4	-3.7%	67.5	-0.2%	46.7	-3.2%	32.1	-7.5%	59.3	-8.5%	39.1	-1.0%	34.1	-1.8%

Table 2: **Per-dataset average performance across non-English languages for 35B and 8B Aya 23 models at varying levels of quantization.** %Δ is relative performance vs. FP16. We follow the evaluation setup of Aryabumi et al. (2024) and evaluate on languages in the 23 languages list. On "Unseen Tasks" (XWinograd, XCOPA, XStoryCloze), we use all the available languages. See Section 3.1 for details and language list.

W8A8, and **W4-g** to **W4**. Table 5 shows the effect of using SmoothQuant and Group-Wise scaling strategies. On average and across mMMLU, MGSM, and FLORES, Group-Wise scaling greatly improves over column-wise W4, recovering over 6 percentage points lost on MGSM for Ltn/IE languages. SmoothQuant has a similar effect on average and for mMMLU, though to a lesser degree. Curiously, SmoothQuant harms MGSM scores slightly, and Group-Wise scaling degrades cross-lingual language confusion. We again observe that ¬Ltn/IE languages suffer more in nearly all cases.

On cross-lingual language confusion, strategies aimed to retain performance have different effects: SmoothQuant recovers all lost from naive W8A8 but is *not* helpful for monolingual language confusion (whereas naive W8A8 was), but Group-Wise scaling is actively damaging. In contrast, column-wise W4 quantization on cross-lingual language confusion benefits Ltn/IE languages and Arabic only,[11] worsening the rest. Thus, while the quantization strategies tend to aid performance overall, the story is nuanced: there may be adverse effects on specific tasks. More research is needed to understand this, but it is intriguing to consider the effect that lower-precision might have on the ability to produce output in a desired language, and maintain that language once decoding begins.

4.6 LLM/RM-as-a-Judge

Table 6 shows relative performance of quantized variants of the 103B model evaluated with LLM- and RM-as-a-Judge.[12] In nearly all cases, the LLM and RM agree that W4 and W4-g quantization severely harm performance on our challenging *Internal* test set. On average across languages, the LLM and RM agree on the ranking of model quality over *Internal*. Results on the easier *Dolly* test set are less clear-cut: The LLM reports greater degradation for W8, W8A8-sq, and W4-g on *Internal* than on *Dolly* (in fact, it asserts that quality *improved* for multiple languages and quantization

[11] Full results in Table A11.

[12] Calculation: $\frac{\text{Quantized Win Rate}-50}{50}$, as 50 is the expected win-rate of two FP16 models compared.

		ar	de	es	fr	it	ja	ko	pt	zh	avg	Ltn/IE	¬
103B	W8	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	-0.4%	-0.2%	-0.1%	-0.1%	-0.1%
	W8A8-sq	-0.6%	0.2%	-0.3%	0.1%	-0.6%	-0.3%	-0.1%	-0.7%	-0.8%	-0.3%	-0.3%	-0.4%
	W8A8	-1.3%	-0.9%	-0.5%	-0.5%	-0.8%	-0.3%	-1.3%	-0.8%	-0.9%	-0.8%	-0.7%	-0.9%
	W4-g	-0.8%	-0.2%	-0.4%	0.1%	-0.4%	-0.4%	-0.6%	-1.2%	-0.9%	-0.5%	-0.4%	-0.7%
	W4	-1.0%	-0.6%	0.1%	-0.8%	-1.2%	-1.4%	-2.9%	-0.8%	-2.3%	-1.2%	-0.7%	-1.9%
35B	W8	0.3%	-0.5%	-0.1%	-0.2%	-0.4%	0.3%	-0.1%	0.1%	-0.3%	-0.1%	-0.2%	0.0%
	W8A8	2.0%	2.5%	0.7%	1.0%	1.2%	1.1%	0.9%	1.4%	1.0%	1.3%	1.3%	1.3%
	W4-g	-1.1%	-1.1%	0.1%	-0.3%	-0.1%	-2.3%	-1.4%	-0.6%	-1.3%	-0.9%	-0.4%	-1.5%

Table 3: **Per-language relative performance (%Δ) vs. FP16, averaged over mMMLU, FLORES, and Language Confusion tasks. Ltn/IE** are Latin-script/Indo-European languages: de, es, fr, it, pt. ¬ are the rest: ar, ja, ko, zh.

		de	es	fr	ja	zh	avg	Ltn/IE	¬
103B	W8	0.1%	-0.1%	-0.3%	-0.4%	-0.2%	-0.2%	-0.1%	-0.3%
	W8A8-sq	0.4%	-0.9%	-0.1%	-0.3%	-1.2%	-0.4%	-0.2%	-0.8%
	W8A8	-0.4%	-1.0%	-0.6%	-0.1%	-1.3%	-0.7%	-0.6%	-0.7%
	W4-g	-0.5%	-0.5%	-0.3%	-1.7%	-1.1%	-0.8%	-0.4%	-1.4%
	W4	-2.3%	-1.1%	-1.7%	-3.0%	-3.5%	-2.3%	-1.7%	-3.3%
35B	W8	-0.6%	-0.3%	-0.1%	-0.4%	0.0%	-0.2%	-0.3%	-0.2%
	W8A8	1.3%	-0.6%	0.3%	-0.3%	0.0%	0.1%	0.3%	-0.2%
	W4-g	-3.7%	-1.8%	-1.7%	-3.8%	-4.0%	-3.0%	-2.4%	-3.9%

Table 4: **Per-language relative performance (%Δ) vs. FP16, averaged over MGSM, mMMLU, FLORES, and Language Confusion tasks. Ltn/IE** are Latin-script/Indo-European: de, es, fr. ¬ are the rest: ja, zh.

levels on *Dolly*), but the RM disagrees. *Internal* shows more pronounced degradation overall, with a 12.4% average relative drop in winrate versus 3.0% on*Dolly* across quantization levels. Perhaps *Dolly* prompts are easy enough that models output similar responses, creating more noise in the judgments; future work could examine this hypothesis.

4.7 Human Evaluation

Human evaluation paints a similar picture in Table 7, with some outliers. Average performance drops steadily across evaluated languages on the *Internal* test set, which has more difficult prompts. The sharpest decline is in French, with -16.6% at W4-g. Curiously, there is an initial 7.4% boost for Japanese with W8, but it falls to -16.0% with more extreme quantization. Interestingly, human annotators generally prefer outputs of quantized models on *Dolly* prompts in Japanese, too, but disprefer those in other languages. We see more pronounced degradation on *Internal* overall, with an average relative drop of 5.7% versus 2.4% for Dolly.

5 Related Work

Impact of Compression on Multilingual Tasks There is a scarcity of research examining the impact of compression and quantization on multilingual tasks. Paglieri et al. (2024) study the impact of multilingual calibration sets on quantization, but their evaluation is English-only. Ramesh et al. (2023) study the effect of compression on multilingual model fairness in terms of classification accuracy, showing that while monolingual evaluation indicates a negative impact, multilingual evaluation differs across languages and dimensions. Kharazmi et al. (2023) show that recovering compression-cauesd performance loss of LSTMs is harder in a multilingual setting than monolingually. In machine translation, Diddee et al. (2022) show that distillation has a varied effect by language due to dependence on priors such as amount of synthetic data used and confidence of the teacher models, while quantization exhibits more consistent performance trends across languages. Our work is the first, to our knowledge, to study the effect of quantization on LLMs and for open-ended generation.

More broadly, multilingual data is an example of long tail data. Prior work shows that compression techniques like quantization and sparsity amplify disparate treatment of long-tail rare features (Hooker et al., 2019; Ahia et al., 2021; Ogueji et al., 2022; Hooker et al., 2020). Ogueji et al. (2022) show that depending on how out of distribution the task data is, sparsity-based compression can sometimes avoid overfitting to the training data, making a model better suited to the downstream task. Ahia et al. (2021) find that sparsity preserves machine translation performance on frequent sentences, but disparately impacts infrequent sentences.

Quantization of LLMs A recent line of work has emerged on techniques to improve performance of quantized LLMs, with the sole focus on English models and data for tuning and evaluation (Ahmadian et al., 2024; Dettmers et al., 2022; Xiao et al., 2023; Bondarenko et al., 2024; Gong et al., 2024).

	Avg. Rel. %		mMMLU		MGSM		FLORES L2 → En		FLORES En → L2		Language Confusion Monolingual		Language Confusion Cross-lingual	
	Ltn/IE	¬	Ltn/IE	¬	Ltn/IE	¬	Ltn/IE	¬	Ltn/IE	¬	Ltn/IE	¬	Ltn/IE	¬
W8A8	-0.7%	-1.0%	-1.3%	-2.1%	-0.9%	-1.3%	-0.1%	0.1%	-1.0%	-1.6%	0.0%	0.4%	-0.9%	-1.6%
W8A8-sq	-0.4%	-0.7%	-0.4%	-0.8%	-1.3%	-1.9%	0.2%	0.0%	-1.1%	-1.6%	-0.1%	0.1%	0.1%	0.0%
W4	-1.9%	-3.3%	-3.9%	-4.9%	-8.0%	-10.2%	-1.3%	-2.0%	-1.1%	-2.3%	0.1%	0.1%	2.9%	-0.4%
W4-g	-0.6%	-1.4%	-1.1%	-1.9%	-1.8%	-4.9%	0.2%	0.7%	-0.3%	-0.8%	0.1%	-0.1%	-0.9%	-1.3%

Table 5: **Effect of mitigation strategies on W8A8 and W4 quantization on the 103B model.** Percentage points off FP16 baseline for W8A8-sq vs. naive W8A8 and W4-g vs. W4, broken down by Latin-script/Indo-European languages (Ltn/IE) versus others (¬). Avg. Rel. % reports averaged performance all datasets.

		fr LLM	fr RM	es LLM	es RM	ja LLM	ja RM	ko LLM	ko RM	avg LLM	avg RM	Ltn/IE LLM	Ltn/IE RM	¬ LLM	¬ RM
Internal	W8	1.0%	-0.7%	-10.2%	7.5%	-5.4%	5.4%	7.5%	-5.8%	-1.8%	1.6%	-4.6%	3.4%	1.0%	-0.2%
	W8A8-sq	-18.4%	-5.1%	-3.7%	4.1%	2.0%	4.7%	3.7%	-5.1%	-4.1%	-0.3%	-11.0%	-0.5%	2.9%	-0.2%
	W4-g	-10.5%	-17.0%	-16.6%	2.0%	-15.3%	0.0%	-5.8%	-15.6%	-12.1%	-7.7%	-13.6%	-7.5%	-10.5%	-7.8%
	W4	-30.2%	-20.4%	-33.0%	-17.0%	-21.7%	-20.0%	-18.6%	-27.6%	-25.9%	-21.2%	-31.6%	-18.7%	-20.2%	-23.8%
Dolly	W8	-1.3%	2.0%	7.3%	-4.0%	-6.0%	-5.3%	2.7%	2.0%	0.7%	-1.3%	3.0%	-1.0%	-1.7%	-1.7%
	W8A8-sq	-15.3%	-8.7%	8.7%	-8.0%	-1.3%	1.3%	-8.0%	-4.7%	-4.0%	-5.0%	-3.3%	-8.3%	-4.7%	-1.7%
	W8A8	-7.4%	2.7%	-4.0%	-3.3%	-15.3%	-1.3%	-11.3%	-3.3%	-9.5%	-1.3%	-5.7%	-0.3%	-13.3%	-2.3%
	W4-g	-3.4%	-2.7%	13.3%	4.7%	2.7%	-15.3%	5.3%	-5.3%	4.5%	-4.7%	5.0%	1.0%	4.0%	-10.3%

Table 6: **Relative performance vs. FP16 of 103B quantized models according to *LLM/RM-as-a-Judge*** over *Internal* and *Aya Dolly* subsampled test sets. Raw win-rates in Table A12.

		fr	es	ja	ko	avg	Ltn/IE	¬
Internal	W8	-7.4%	0.6%	7.4%	-12.0%	-2.8%	-3.4%	-2.3%
	W8A8-sq	-9.4%	-7.4%	-2.0%	4.0%	-3.7%	-8.4%	1.0%
	W4-g	-16.6%	-4.6%	-16.0%	-4.6%	-10.5%	-10.6%	-10.3%
Dolly	W8	0.6%	-5.4%	12.0%	0.0%	1.8%	-2.4%	6.0%
	W8A8-sq	-7.4%	-8.6%	0.0%	-3.4%	-4.8%	-8.0%	-1.7%
	W4-g	-9.4%	-1.4%	2.6%	-8.0%	-4.1%	-5.4%	-2.7%

Table 7: **Relative performance vs. FP16 of 103B quantized models according to *human evaluators*** over *Internal* and *Aya Dolly* subsampled test sets.

Even the most recent (Li et al., 2024; Liu et al., 2024) omit the multilingual dimension without acknowledging the limitation. Multilinguality and compression are both integral parts of LLMs, and our work explores this new territory.

Model design choices We consider how design choices such as quantization impact performance for users of different languages. A wider body of work examines how design choices impact performance on underrepresented features or subgroups. Zhuang et al. (2021) and Nelaturu et al. (2023) find that hardware choice incurs disparate impact on underrepresented features. Wang et al. (2022) establish that distillation imposes similar trade-offs, but the disproportionate harm to the long-tail could be mitigated by modifying the student-teacher objective. Ko et al. (2023) evaluate the positive role of ensembling disproportionately favoring underrepresented attributes. Bagdasaryan and Shmatikov (2019) show that differential privacy techniques like gradient clipping and noise injection disproportionately impact underrepresented features.

6 Conclusion & Future Work

We examine widely adopted quantization techniques for model compression and ask, *How do they impact different languages?* We perform an extensive study of quantization in state-of-the-art multilingual LLMs—from 8 billion to 103 billion parameters—in 20+ languages using automatic metrics, LLM-as-a-Judge, RM-as-a-Judge, and human evaluation. We find that: (1) Damage from quantization is much worse than appears from automatic metrics: even when not observed automatically, human evaluators notice it. (2) Quantization affects languages to varying degrees, with non-Latin script languages more severely affected on automatic benchmarks. (3) Challenging tasks degrade fast and severely: math performance is strikingly reduced, as are responses on realistic challenging prompts judged by humans. On a bright note, quantization occasionally brings performance benefits.

Our results urge attention to multilingual performance at all stages of system design. Researchers might extend our work to consider the impact of other decisions on multilingual performance, including on languages excluded from training and out-of-distribution tasks. By being mindful of the impact on long-tail features, we'll build better systems to serve the world.

7 Limitations

Generality of findings Due to the number of methods, languages, and benchmarks we examine, we focus our evaluation on models from two families (Command R and Aya). As we observe similar trends across these models, our findings are likely to generalize to other LLMs. Nevertheless, models that have been optimized differently or trained with a focus on specific tasks such as code or mathematical reasoning may behave differently.

Under-represented languages For our study, we focused on languages that were supported by the models we evaluated. Performance deterioration is likely even larger for languages that are not or severely under-represented in the pre-training data. For such languages, evaluation is also more challenging due to poor availability of benchmark data and human annotators.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. *Preprint*, arXiv:2110.03036.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *Preprint*, arXiv:2305.13707.

Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. 2024. Intriguing properties of quantization at scale. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Eugene Bagdasaryan and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Preprint*, arXiv:1905.12101.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2024. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022a. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022b. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. 2022. Too brittle to touch: Comparing the stability of quantization and distillation towards developing low-resource MT models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 870–885, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv*, abs/2306.16388.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Zhuocheng Gong, Jiahao Liu, Jingang Wang, Xunliang Cai, Dongyan Zhao, and Rui Yan. 2024. What makes quantization for large language model hard? an empirical study from the lens of perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18082–18089.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in nlp. *arXiv*, abs/2311.08391.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Sara Hooker, Aaron C. Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. What do compressed deep neural networks forget. *arXiv: Learning*.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *Preprint*, arXiv:2010.03058.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2023. Casteist but not racist? quantifying disparities in large language model bias between india and the west. *ArXiv*, abs/2309.08573.

Pegah Kharazmi, Zhewei Zhao, Clement Chung, and Samridhi Choudhary. 2023. Distill-quantize-tune-leveraging large teachers for low-footprint efficient multilingual nlu on edge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv*, abs/2305.14976.

Wei-Yin Ko, Daniel D'souza, Karina Nguyen, Randall Balestriero, and Sara Hooker. 2023. Fair-ensemble: When fairness naturally emerges from deep ensembling. *Preprint*, arXiv:2303.00586.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. *Proceedings of The ACM Collective Intelligence Conference*.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023a. Privacy in large language models: Attacks, defenses and future directions. *ArXiv*, abs/2310.10383.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating quantized large language models. *arXiv preprint arXiv:2402.18158*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2024. Do emergent abilities exist in quantized large language models: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5174–5190, Torino, Italia. ELRA and ICCL.

Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. 2023. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv*, abs/2311.17035.

Sree Harsha Nelaturu, Nishaanth Kanna Ravichandran, Cuong Tran, Sara Hooker, and Ferdinando Fioretto. 2023. On the fairness impacts of hardware selection in machine learning. *Preprint*, arXiv:2312.03886.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *arXiv*, abs/2306.07377.

Kelechi Ogueji, Orevaoghene Ahia, Gbemileke Onilude, Sebastian Gehrmann, Sara Hooker, and Julia Kreutzer. 2022. Intriguing properties of compression on multilingual models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9092–9110, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I. Adelani. 2023. How good are large language models on african languages? *arXiv*, abs/2311.07978.

Davide Paglieri, Saurabh Dash, Tim Rocktäschel, and Jack Parker-Holder. 2024. Outliers and calibration sets have diminishing effect on quantization of modern llms. *Preprint*, arXiv:2405.20835.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. pages 2362–2376.

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

Luiza Pozzobon, Patrick Lewis, Sara Hooker, and Beyza Ermis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. *Preprint*, arXiv:2403.03893.

Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada. Association for Computational Linguistics.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. 2022. Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, 1270(10.6028).

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. *arXiv*, abs/2307.01503.

Serena Wang, Harikrishna Narasimhan, Yichen Zhou, Sara Hooker, Michal Lukasik, and Aditya Krishna Menon. 2022. Robust distillation for worst-class performance. *Preprint*, arXiv:2206.06479.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. *arXiv*, abs/2310.02446.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Donglin Zhuang, Xingyao Zhang, Shuaiwen Leon Song, and Sara Hooker. 2021. Randomness in neural network training: Characterizing the impact of tooling. *Preprint*, arXiv:2106.11872.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Appendix

A.1 Prompts for mMMLU and LLM-as-a-Judge

The following are multiple choice questions (with answers) about clinical knowledge.

다음중파제트병에대한설명으로옳은것은무엇입니까?
A. 긴뼈가휘어지는것이특징
B. 척수압박은흔한합병증이다
C. 심부전은알려진합병증이아니다
D. 병적골절은특징이아닙니다.
Answer: B
...

Table A1: **mMMLU prompt**. Following Achiam et al. (2023), letter choices and “Answer” are kept in English.

Example Prompt

I want you to create a leaderboard of different large-language models. To do so, I will give you the conversations (prompts) given to the models, and the responses of two models. Please rank the models based on which responses would be preferred by humans. All inputs and outputs should be python dictionaries.

Here is the prompt:
{
    "conversation": """User: La tomate est-elle un fruit ou un légume?""",
}

Here are the outputs of the models:
[
    {
    "model": "model_1",
    "answer": """La tomate est un fruit. Plus précisément, il s'agit d'un fruit charnu, issu de la transformation de l'ovaire de la fleur du plant de tomate."""
    },
    {
    "model": "model_2",
    "answer": """La tomate est un fruit du point de vue botanique, car elle contient des graines et se développe à partir de la fleur d'une plante. Cependant, en cuisine, on considère souvent la tomate comme un légume en raison de son utilisation dans des plats salés et de sa saveur moins sucrée par rapport à d'autres fruits."""
    }
]

Now please rank the models by the quality of their answers, so that the model with rank 1 has the best output. Then return a list of the model names and ranks, i.e., produce the following output:
[
    {'model': <model-name>, 'rank': <model-rank>},
    {'model': <model-name>, 'rank': <model-rank>}
]

Your response must be a valid Python dictionary and should contain nothing else because we will directly execute it in Python. Please provide the ranking that the majority of humans would give.

Table A2: **Example Input for LLM-as-a-Judge.** Template derived from Li et al. (2023b): https://github.com/tatsu-lab/alpaca_eval/blob/main/src/alpaca_eval/evaluators_configs/gpt-3.5-turbo-1106_ranking/ranking_prompt.txt

A.2 Full Task Results

		de	es	fr	ja	zh	avg
103B	W8	0.3%	-0.9%	-1.6%	-2.7%	-0.1%	-1.0%
	W8A8	2.1%	-3.5%	-1.1%	0.6%	-3.2%	-1.0%
	W4-g	-1.9%	-1.3%	-2.3%	-7.9%	-1.9%	-3.0%
	W8A8-sq	1.1%	-3.7%	-1.5%	-0.1%	-3.6%	-1.6%
	W4	-11.0%	-7.0%	-5.9%	-10.9%	-9.6%	-8.8%
35B	W8	-1.3%	-1.1%	0.6%	-3.7%	1.6%	-0.8%
	W8A8	-4.4%	-6.8%	-3.6%	-7.6%	-5.4%	-5.6%
	W4-g	-16.7%	-10.9%	-9.0%	-11.5%	-17.3%	-13.1%

Table A3: Percentage drop off FP16 baseline on MGSM. AVG is the average of *percentage drops*, so that all languages are on the same scale (as opposed to languages with higher raw scores dominating the averages)

		Avg	XSC	XCOPA	XWNG
Aya-23-8b	FP16	67.6	62.3	59.8	80.7
	W8	67.6	62.4	60.0	80.6
	W4	67.5	62.3	59.6	80.6
Aya-23-35b	FP16	70.8	65.1	62.8	84.4
	W8	70.6	65.0	62.9	83.9
	W4	70.5	64.8	62.3	84.5

Table A4: **Performance of quantized Aya 23 models on unseen discriminative tasks**. XStoryCloze (XSC), XCOPA, and XWinograd (XWNG).

		English → L2										L2 → English									
		ar	de	es	fr	it	ja	ko	pt	zh	avg	ar	de	es	fr	it	ja	ko	pt	zh	avg
103B	FP16	27.1	40.0	30.1	50.6	33.1	33.1	29.1	51.0	45.1	37.7	45.0	46.3	33.4	48.6	36.5	29.5	33.0	52.2	32.1	39.6
	W8	27.2	40.0	30.0	50.7	33.1	33.2	29.1	50.9	45.1	37.7	45.2	46.3	33.4	48.5	36.5	29.5	33.0	52.1	32.0	39.6
	W8A8-sq	26.8	40.3	30.0	51.0	33.0	33.1	29.3	51.2	45.1	37.8	44.5	46.2	32.9	48.1	35.9	29.3	32.5	51.6	31.2	39.1
	W8A8	26.9	39.8	30.0	50.9	33.0	33.7	29.0	51.1	45.1	37.7	44.4	45.9	33.1	47.9	36.2	29.2	32.5	51.8	31.4	39.1
	W4-g	27.3	40.4	30.1	51.0	33.0	33.9	29.3	50.9	44.7	37.8	44.9	46.4	33.2	48.4	36.3	29.3	32.7	52.0	31.6	39.4
	W4	26.9	39.1	29.9	50.0	32.8	32.8	27.9	50.3	44.0	37.1	44.2	45.8	33.1	47.9	36.0	29.0	32.3	51.8	30.9	39.0
35B	FP16	20.1	33.5	27.8	44.5	29.7	27.0	22.7	45.5	40.4	32.4	38.4	41.2	31.8	43.1	34.0	26.2	28.4	48.1	28.4	35.5
	W8	20.0	33.4	27.8	44.5	29.7	26.9	22.9	45.3	40.3	32.3	38.3	41.1	31.7	43.0	34.0	26.4	28.2	48.0	28.2	35.4
	W8A8	21.2	34.1	27.8	45.1	30.0	27.6	23.1	46.1	40.8	32.9	38.5	42.2	31.7	43.5	34.2	26.5	28.6	48.6	28.7	35.8
	W4-g	18.8	32.9	27.7	43.9	29.6	26.0	22.1	45.1	39.7	31.7	38.3	41.4	31.0	43.1	34.0	25.5	28.1	48.0	28.0	35.3

Table A5: Full results (raw BLEU score) on FLORES

		English → L2										L2 → English									
		ar	de	es	fr	it	ja	ko	pt	zh	avg	ar	de	es	fr	it	ja	ko	pt	zh	avg
103B	W8	0.1%	0.1%	-0.4%	0.2%	0.1%	0.3%	-0.2%	-0.3%	0.1%	0.0%	0.4%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	0.0%	-0.1%	-0.1%	0.0%
	W8A8-sq	-1.1%	0.8%	-0.4%	0.7%	-0.2%	0.2%	0.7%	0.3%	0.1%	0.1%	-1.2%	-0.1%	-1.4%	-1.0%	-1.8%	-0.7%	-1.6%	-1.1%	-2.9%	-1.3%
	W8A8	-1.0%	-0.4%	-0.5%	0.5%	-0.4%	1.8%	-0.4%	0.1%	0.0%	0.0%	-1.3%	-0.8%	-1.1%	-1.4%	-1.0%	-1.2%	-1.7%	-0.8%	-2.1%	-1.3%
	W4-g	0.7%	1.0%	-0.3%	0.8%	-0.3%	2.6%	0.5%	-0.3%	-0.8%	0.4%	-0.3%	0.2%	-0.6%	-0.3%	-0.6%	-0.7%	-0.9%	-0.3%	-1.4%	-0.6%
	W4	-0.8%	-2.2%	-0.7%	-1.3%	-0.9%	-0.8%	-4.3%	-1.5%	-2.3%	-1.6%	-1.8%	-1.1%	-0.8%	-1.4%	-1.6%	-1.7%	-2.2%	-0.7%	-3.6%	-1.7%
35B	W8	-0.7%	-0.4%	-0.1%	0.0%	0.0%	-0.2%	0.7%	-0.4%	-0.2%	-0.1%	-0.2%	-0.2%	-0.5%	-0.3%	0.0%	0.8%	-0.5%	-0.1%	-0.6%	-0.2%
	W8A8	5.5%	1.9%	0.1%	1.4%	0.9%	2.1%	1.9%	1.4%	0.9%	1.8%	0.5%	2.5%	-0.6%	0.9%	0.5%	1.1%	0.8%	1.1%	1.0%	0.9%
	W4-g	-6.7%	-1.9%	-0.4%	-1.3%	-0.4%	-3.9%	-2.8%	-0.7%	-1.7%	-2.2%	-0.1%	0.6%	-2.5%	0.0%	-0.1%	-2.8%	-1.1%	-0.2%	-1.4%	-0.8%

Table A6: Percentage drop off FP16 baseline on Flores. AVG is the average of *percentage drops*, so that all languages are on the same scale (as opposed to languages with higher raw scores dominating the averages)

		de	es	fr	ja	zh	avg
103B	FP16	72.6	76.6	70.6	63.0	70.2	70.6
	W8	72.8	75.9	69.5	61.3	70.2	69.9
	W8A8	74.1	73.9	69.8	63.4	68.0	69.8
	W4-g	71.2	75.7	69.0	58.0	68.9	68.6
	W8A8-sq	73.4	73.8	69.6	62.9	67.7	69.5
	W4	64.6	71.3	66.5	56.1	63.5	64.4
35B	FP16	56.6	57.3	51.8	38.8	44.4	49.8
	W8	55.9	56.6	52.1	37.4	45.1	49.4
	W8A8	54.2	53.4	49.9	35.8	42.0	47.1
	W4-g	47.2	51.0	47.1	34.3	36.7	43.3

Table A7: Raw MGSM scores (accuracy)

		ar	de	es	fr	it	ja	ko	pt	zh	avg
103B	FP16	64.0	68.3	68.7	68.0	69.3	64.4	62.3	70.0	65.0	66.7
	W8	64.1	68.3	68.7	68.1	69.4	64.3	62.3	69.9	65.0	66.7
	W8A8	62.6	67.1	68.2	67.4	68.3	62.9	60.8	68.7	64.1	65.6
	W4-g	62.9	67.5	68.2	67.6	68.6	62.8	61.1	68.6	64.0	65.7
	W8A8-sq	63.5	67.9	68.8	68.0	69.1	63.6	61.8	69.2	64.9	66.3
	W4	60.5	65.7	66.5	65.4	66.6	61.1	59.3	66.7	62.1	63.8
35B	FP16	56.5	60.7	62.3	61.8	62.0	56.4	54.8	62.0	57.9	59.4
	W8	56.5	60.6	62.2	61.8	61.9	56.4	54.7	62.1	57.9	59.3
	W8A8	56.4	60.5	62.5	61.9	62.0	55.8	54.5	61.8	58.1	59.3
	W4-g	55.4	59.7	62.0	61.0	60.7	54.4	53.2	60.8	56.6	58.2

Table A8: Raw results on mMMLU (accuracy)

		ar	de	es	fr	it	ja	ko	pt	zh	avg
103B	W8	0.2%	0.0%	0.0%	0.1%	0.1%	-0.2%	0.0%	-0.1%	0.0%	0.0%
	W8A8	-2.2%	-1.8%	-0.7%	-1.0%	-1.5%	-2.3%	-2.4%	-1.8%	-1.4%	-1.7%
	W4-g	-1.7%	-1.2%	-0.7%	-0.6%	-1.0%	-2.5%	-1.9%	-2.0%	-1.5%	-1.5%
	W8A8-sq	-0.8%	-0.6%	0.1%	0.1%	-0.3%	-1.3%	-0.8%	-1.1%	-0.2%	-0.5%
	W4	-5.5%	-3.8%	-3.1%	-3.9%	-3.8%	-5.1%	-4.8%	-4.8%	-4.4%	-4.4%
35B	W8	0.0%	-0.2%	-0.2%	0.0%	-0.2%	0.0%	-0.2%	0.2%	0.0%	-0.1%
	W8A8	-0.2%	-0.3%	0.3%	0.2%	0.0%	-1.1%	-0.5%	-0.3%	0.3%	-0.2%
	W4-g	-1.9%	-1.6%	-0.5%	-1.3%	-2.1%	-3.5%	-2.9%	-1.9%	-2.2%	-2.0%

Table A9: Percentage drop off FP16 baseline on mMMLU. AVG is the average of *percentage drops*, so that all languages are on the same scale (as opposed to languages with higher raw scores dominating the averages)

		ar	de	es	fr	it	ja	ko	pt	zh	avg	ar	de	es	fr	it	ja	ko	pt	zh	avg
103B	FP16	99.3	100.0	99.3	99.6	100.0	98.6	100.0	98.3	97.9	99.2	93.0	90.6	91.2	91.6	93.0	93.1	91.1	88.3	91.3	91.5
	W8	99.0	100.0	99.5	99.4	99.8	99.2	99.8	97.8	98.5	99.2	92.6	91.1	91.7	91.4	92.9	92.8	91.3	87.4	89.7	91.2
	W8A8	99.3	100.0	99.5	99.8	100.0	99.0	99.8	98.1	99.1	99.4	91.3	89.3	91.0	91.1	91.8	93.0	89.3	87.3	89.2	90.4
	W4-g	99.1	100.0	99.6	99.9	100.0	97.4	100.0	98.1	98.9	99.2	90.6	89.9	90.7	91.7	93.1	92.8	90.6	85.4	89.6	90.5
	W8A8-sq	99.4	100.0	99.3	99.6	100.0	98.6	100.0	97.7	98.4	99.2	93.3	91.5	91.4	92.4	92.1	93.3	92.1	87.6	90.0	91.5
	W4	99.4	100.0	99.4	99.7	99.8	99.6	99.0	98.9	98.4	99.3	95.8	94.3	95.9	93.8	93.6	92.6	88.9	90.5	89.7	92.8
35B	FP16	99.2	97.0	98.1	99.2	99.6	99.6	99.0	99.0	97.7	98.7	58.8	59.6	69.0	73.0	63.6	66.3	69.2	64.2	74.6	66.5
	W8	99.7	97.0	98.1	98.9	100.0	99.8	99.0	99.3	97.4	98.8	59.8	58.6	69.2	72.8	62.3	66.8	68.7	64.3	74.5	66.3
	W8A8	99.9	98.0	97.1	98.9	100.0	100.0	100.0	99.0	98.4	99.0	61.0	63.9	72.1	75.1	66.2	68.3	70.1	67.4	76.3	68.9
	W4-g	99.4	95.0	96.5	99.9	100.0	99.8	97.0	98.3	98.6	98.3	60.4	59.3	72.8	73.3	64.8	65.4	70.4	64.6	73.0	67.1

Table A10: Raw scores for Language confusion metrics. Left: Monolingual, Right: Cross-lingual

		Monolingual										Cross-Lingual									
		ar	**de**	**es**	**fr**	**it**	**ja**	**ko**	**pt**	**zh**	**avg**	**ar**	**de**	**es**	**fr**	**it**	**ja**	**ko**	**pt**	**zh**	**avg**
103B	W8	-0.3%	0.0%	0.2%	-0.2%	-0.2%	0.6%	-0.2%	-0.5%	0.6%	0.0%	-0.5%	0.5%	0.5%	-0.2%	-0.1%	-0.3%	0.3%	-1.0%	-1.8%	-0.3%
	W8A8	0.0%	0.0%	0.2%	0.2%	0.0%	0.4%	-0.2%	-0.2%	1.2%	0.2%	-1.8%	-1.5%	-0.2%	-0.6%	-1.2%	-0.2%	-1.9%	-1.1%	-2.3%	-1.2%
	W4-g	-0.2%	0.0%	0.3%	0.4%	0.0%	-1.2%	0.0%	-0.2%	1.0%	0.0%	-2.6%	-0.8%	-0.6%	0.1%	0.1%	-0.3%	-0.5%	-3.3%	-1.8%	-1.1%
	W8A8-sq	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-0.6%	0.5%	0.0%	0.3%	1.0%	0.2%	0.9%	-0.9%	0.2%	1.1%	-0.8%	-1.4%	0.1%
	W4	0.0%	0.0%	0.1%	0.1%	-0.2%	1.0%	-1.0%	0.6%	0.5%	0.1%	3.0%	4.0%	5.1%	2.3%	0.6%	-0.5%	-2.4%	2.5%	-1.8%	1.4%
35B	W8	0.5%	0.0%	0.0%	-0.3%	0.4%	0.2%	0.0%	0.3%	-0.3%	0.1%	1.7%	-1.6%	0.2%	-0.3%	-2.0%	0.8%	-0.8%	0.3%	-0.2%	-0.2%
	W8A8	0.7%	1.0%	-0.9%	-0.3%	0.4%	0.4%	1.0%	0.0%	0.7%	0.3%	3.8%	7.2%	4.4%	2.9%	4.1%	3.1%	1.2%	5.0%	2.3%	3.8%
	W4-g	0.2%	-2.1%	-1.6%	0.7%	0.4%	0.2%	-2.0%	-0.7%	0.9%	-0.4%	2.8%	-0.5%	5.4%	0.5%	1.9%	-1.3%	1.7%	0.6%	-2.2%	1.0%

Table A11: Percentage drop off FP16 baseline on Language Confusion Metrics. AVG is the average of *percentage drops*, so that all languages are on the same scale (as opposed to languages with higher raw scores dominating the averages)

		fr		es		ja		ko		avg		Ltn/IE		¬	
		LLM	**RM**	**LLM**	**RM**	**LLM**	**RM**	**LLM**	**RM**	**LLM**	**RM**	**LLM**	**RM**	**LLM**	**RM**
Internal	W8	50.5	49.7	44.9	53.7	47.3	52.7	53.7	47.1	49.1	50.8	47.7	51.7	50.5	49.9
	W8A8-sq	40.8	47.5	48.1	52.0	51.0	52.4	51.9	47.5	48.0	49.8	44.5	49.7	51.4	49.9
	W4-g	44.8	41.5	41.7	51.0	42.4	50.0	47.1	42.2	44.0	46.2	43.2	46.3	44.7	46.1
	W4	34.9	39.8	33.5	41.5	39.2	40.0	40.7	36.2	37.1	39.4	34.2	40.7	39.9	38.1
Dolly	W8	49.3	51.0	53.7	48.0	47.0	47.3	51.3	51.0	50.3	49.3	51.5	49.5	49.2	49.2
	W8A8-sq	42.3	45.7	54.3	46.0	49.3	50.7	46.0	47.7	48.0	47.5	48.3	45.8	47.7	49.2
	W8A8	46.3	51.3	48.0	48.3	42.3	49.3	44.3	48.3	45.2	49.3	47.2	49.8	43.3	48.8
	W4-g	48.3	48.7	56.7	52.3	51.3	42.3	52.7	47.3	52.2	47.7	52.5	50.5	52.0	44.8

Table A12: ***LLM/RM-as-a-Judge*** **Raw win-rates of 103B quantized models vs. FP16** over *Internal* and *Aya Dolly* subsampled test sets.

		fr	**es**	**ja**	**ko**	**avg**	**Ltn/IE**	**¬**
Internal	W8	46.3	50.3	53.7	44.0	48.6	48.3	48.9
	W8A8-sq	45.3	46.3	49.0	52.0	48.2	45.8	50.5
	W4-g	41.7	47.7	42.0	47.7	44.8	44.7	44.9
Dolly	W8	50.3	47.3	56.0	50.0	50.9	48.8	53.0
	W8A8-sq	46.3	45.7	50.0	48.3	47.6	46.0	49.2
	W4-g	45.3	49.3	51.3	46.0	48.0	47.3	48.7

Table A13: **Human evaluation raw win-rates of 103B quantized models vs. FP16** over *Internal* and *Aya Dolly* subsampled test sets.

mMMLU		Avg (w/o en)	Avg (w/ en)	en	ar	de	es	fr	hi	id	it	nl	pt	ro	ru	uk	vi	zh
Aya-23-8b	FP16	48.2	48.6	54.6	45.1	50.0	50.9	51.0	39.7	48.8	50.7	49.7	50.8	49.9	47.8	46.8	46.5	47.1
	W8	47.8	48.2	54.2	44.9	49.9	50.5	50.6	39.4	48.5	50.2	49.4	50.6	49.2	47.4	46.3	45.7	46.4
	W4	46.7	47.1	53.4	43.9	48.4	49.4	49.0	38.4	47.5	49.1	47.9	49.1	48.0	46.2	45.6	44.9	46.1
Aya-23-35b	FP16	58.2	58.8	66.7	53.9	60.4	61.6	62.0	47.8	58.9	61.5	60.3	62.0	59.7	57.8	56.3	55.3	57.5
	W8	57.9	58.5	66.2	53.8	60.0	61.7	61.7	47.4	58.7	61.1	60.0	61.6	59.1	57.5	56.1	54.9	57.5
	W4	56.6	57.2	65.2	52.3	58.7	60.3	60.4	45.7	57.4	59.8	58.6	60.5	57.7	56.5	55.0	53.8	56.1

Table A14: **Aya 23 models' language-specific results for Multilingual MMLU (Okapi)**

Belebele		Avg (w/o eng)	Avg (w/ eng)	en	ar	cs	de	el	fr	hi	id	it	ja	ko	nl	fa	pl	pt	ro	ru	es	tr	uk	vi	zh
Aya-23-8b	FP16	64.8	65.3	77.0	65.6	61.9	65.6	64.0	69.6	54.3	67.4	65.7	65.2	61.7	63.8	63.6	61.3	69.1	65.7	69.7	67.0	58.1	66.8	62.3	72.2
	W8	64.6	65.1	76.1	64.3	61.8	64.8	63.0	70.4	54.2	67.4	64.6	65.4	61.4	64.3	63.9	59.8	68.7	65.4	68.7	67.4	58.1	67.0	63.7	71.8
	W4	59.3	59.9	73.8	61.9	57.0	61.6	57.7	65.7	49.8	64.7	58.3	60.7	51.1	60.7	58.2	54.9	62.0	59.8	63.9	61.1	50.1	61.0	58.8	66.2
Aya-23-35b	FP16	77.6	77.9	84.7	78.9	78.2	77.1	76.4	81.9	65.6	77.8	79.8	75.9	73.3	77.7	75.8	75.8	83.8	78.9	79.6	81.0	74.1	77.6	78.3	81.2
	W8	77.1	77.4	84.6	77.3	78.8	77.2	76.6	82.4	65.6	77.6	80.8	74.8	73.7	77.6	74.9	74.8	82.9	77.1	78.9	80.8	72.0	77.2	77.0	80.3
	W4	73.0	73.4	83.2	73.8	74.9	73.2	70.8	78.1	61.0	73.9	76.2	71.7	67.4	73.0	71.4	70.4	80.1	74.3	73.3	77.0	68.2	73.0	71.4	78.8

Table A15: **Aya 23 models' language-specific results for Belebele**

MGSM (5-shot)		Avg (w/o eng)	Avg (w/ eng)	de	en	es	fr	ja	ru	zh
Aya-23-8b	FP16	34.7	36.6	40.4	48.0	45.2	38.8	12.8	38.0	32.8
	W8	35.4	36.9	39.6	45.6	45.6	38.8	13.6	38.8	36.0
	W4	32.1	33.5	39.6	42.4	42.0	34.0	7.2	33.6	36.0
Aya-23-35b	FP16	51.2	53.7	61.6	68.4	58.4	55.6	22.8	58.0	50.8
	W8	52.1	54.2	54.4	66.4	61.2	60.4	24.4	57.2	55.2
	W4	48.1	50.7	58.8	66.0	54.8	54.8	18.4	53.6	48.4

Table A16: **Aya 23 models' language-specific results for MGSM**

FLORES		English→X																					
		Avg	ar	cs	zh	nl	fr	de	el	he	hi	id	it	ja	ko	fa	pl	pt	ro	ru	es	tr	uk	vi
Aya-23-8b	FP16	34.8	36.3	35.7	30.1	32.2	51.0	39.3	34.0	35.0	27.2	43.4	34.7	24.9	22.0	30.0	28.4	50.2	41.6	35.0	31.5	29.1	34.2	39.0
	W8	34.8	36.5	36.1	30.6	32.1	51.4	39.5	33.9	35.0	27.0	43.2	34.8	24.8	22.2	30.4	28.5	50.0	42.0	34.9	31.4	28.9	34.3	39.0
	W4	34.1	35.4	35.0	29.4	31.8	50.2	39.2	33.4	33.3	26.4	42.8	34.3	24.3	21.5	29.6	28.0	49.8	40.9	34.2	31.2	28.1	33.7	38.5
Aya-23-8b	FP16	37.8	40.0	39.1	34.0	33.4	54.1	42.5	36.3	39.5	31.9	44.7	36.6	28.7	25.5	33.4	30.7	53.1	43.3	38.9	32.1	33.8	38.2	41.0
	W8	37.9	40.0	39.0	33.8	33.6	53.9	42.9	36.2	40.0	32.3	44.8	36.5	28.9	25.5	33.7	30.9	53.2	43.4	38.7	32.2	33.8	38.3	41.4
	W4	37.2	39.3	38.0	33.1	32.9	53.3	42.5	36.0	39.1	31.2	44.6	36.1	28.2	25.1	32.6	30.0	52.8	42.6	38.3	32.0	33.2	37.8	40.8
		X→English																					
Aya-23-8b	FP16	39.5	42.4	42.0	31.6	35.8	48.1	46.5	38.7	43.7	37.4	45.5	37.9	29.9	30.9	36.5	33.6	51.7	46.7	38.6	35.4	36.9	41.2	38.2
	W8	39.7	42.1	42.5	31.9	35.6	48.1	46.7	39.2	44.2	37.7	45.5	38.2	30.0	31.3	36.8	33.7	52.0	46.6	38.5	35.5	37.0	41.6	38.4
	W4	39.1	41.4	42.2	31.6	35.7	47.4	46.2	38.1	42.8	36.2	44.7	38.5	29.7	30.1	36.7	33.1	51.9	46.1	38.3	35.8	35.7	40.7	37.6
Aya-23-8b	FP16	42.9	46.4	45.3	34.8	37.7	50.6	48.9	42.4	48.3	42.7	48.5	40.5	33.7	35.3	41.3	36.4	54.8	49.5	41.6	37.7	42.2	44.8	41.4
	W8	43.0	46.4	45.4	34.9	37.5	50.7	49.0	42.2	48.6	42.9	48.7	40.5	34.0	35.1	41.4	36.4	54.8	49.5	41.7	37.3	42.2	45.0	41.5
	W4	42.4	45.7	44.9	34.2	37.2	50.4	48.5	41.8	47.3	41.8	48.0	40.8	33.1	34.4	40.5	35.8	54.3	49.2	41.6	37.5	41.4	44.2	40.9

Table A17: **Aya 23 models' language-specific results for Flores**