

Does quantization affect models’ performance on long-context tasks?

Anmol Mekala[★]² Anirudh Atmakuru[★]²

{amekala, aatmakuru}@umass.edu

Yixiao Song² Marzena Karpinska³ Mohit Iyyer²^W

yixiaosong@umass.edu mkarpinska@microsoft.com miyyer@umd.edu

²UMass Amherst ³Microsoft ^WUniversity of Maryland, College Park

Abstract

Large language models (LLMs) now support context windows exceeding 128K tokens, but this comes with significant memory requirements and high inference latency. Quantization can mitigate these costs, but may degrade performance. In this work, we present the first systematic evaluation of quantized LLMs on tasks with long inputs (≥ 64 K tokens) and long-form outputs. Our evaluation spans 9.7K test examples, five quantization methods (FP8, GPTQ-int8, AWQ-int4, GPTQ-int4, BNB-nf4), and five models (Llama-3.1 8B and 70B; Qwen-2.5 7B, 32B, and 72B). We find that, on average, 8-bit quantization preserves accuracy ($\sim 0.8\%$ drop), whereas 4-bit methods lead to substantial losses, especially for tasks involving long-context inputs (drops of up to 59%). This degradation tends to worsen when the input is in a language other than English. Crucially, the effects of quantization depend heavily on the quantization method, model, and task. For instance, while Qwen-2.5 72B remains robust under BNB-nf4, Llama-3.1 70B experiences a 32% performance drop on the same task. These findings highlight the importance of a careful, task-specific evaluation before deploying quantized LLMs, particularly in long-context scenarios and for languages other than English.

 github.com/molereddy/long-context-quantization

1 Introduction

Large language models (LLMs) serve a vast community of users through APIs and web services. However, their inference demands substantial resources, severely constraining scalability and increasing energy consumption (Samsi et al., 2023; Luccioni et al., 2024; Fernandez et al., 2025). Post-training quantization aims to solve this problem by approximating models at lower precision (Zhou et al., 2024; Zhu et al., 2024). 8-bit quantization

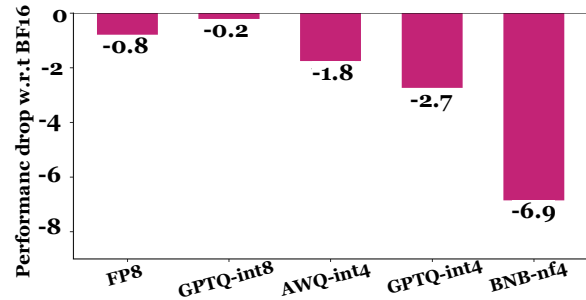


Figure 1: Average performance drop in percentage points caused by quantization on **long-context tasks**. The drop is computed by subtracting the performance of each quantization method (averaged over tasks and models) from the corresponding BF16 baseline.

(Micikevicius et al., 2022; Frantar et al., 2023) provides accuracy at half the cost, and more aggressive 4-bit approaches (Lin et al., 2024; Dettmers et al., 2023; Kim et al., 2023) offer even greater cost-effectiveness. Model providers widely offer quantized models (Meta AI, 2024; Google Cloud, 2024; Databricks, 2023), but the impact on performance has not been explored exhaustively.

While previous work has shown that quantized LLMs largely maintain performance on standard benchmarks (Badshah and Sajjad, 2024; Liu et al., 2024b; Jin et al., 2024; Huang et al., 2024, inter alia), these studies have mainly focused on tasks with short-form inputs and outputs (Jin et al., 2024; Lee et al., 2024), rarely exceeding 2K input and 100 output tokens. This limited scope leaves open a critical gap: the impact of quantization on tasks requiring long-context inputs or long-form outputs.¹ This is particularly salient because, as context length increases, the associated memory and compute requirements grow quadratically, making low-precision methods particularly attractive for efficiency. Moreover, many modern LLMs now

¹We define long-form inputs as over 64K tokens, and long-form outputs as typically 250-650 tokens long.

[★]These authors contributed equally to this work.

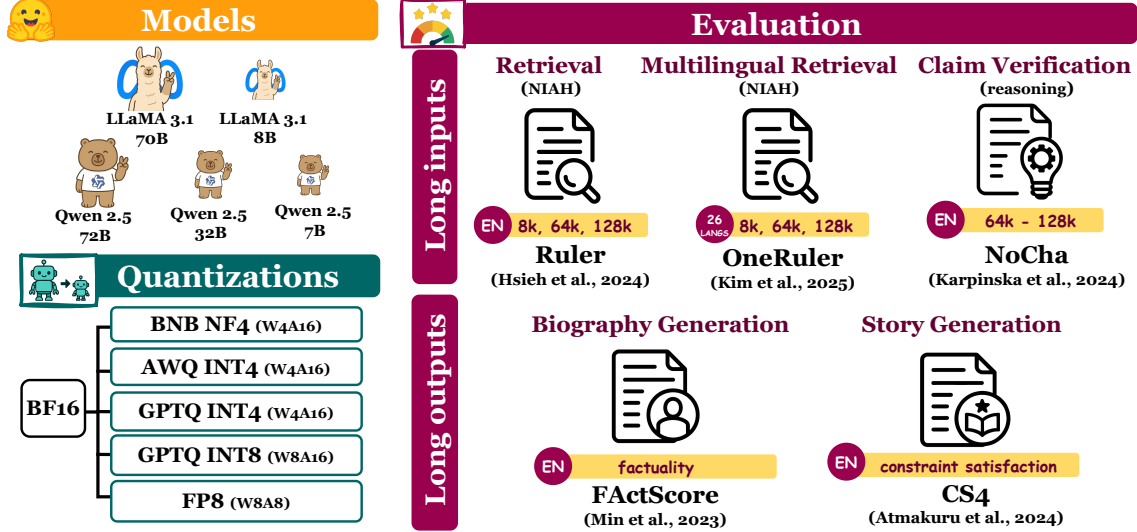


Figure 2: Overview of the evaluation pipeline. We assess five models from two model families (right), comparing their performance in BF16 precision against their quantized variants across five quantization methods. The models are evaluated on three long-input tasks (top left) and two long-form output tasks (bottom left).

support context windows of 128K tokens or more, allowing them to process entire codebases, legal documents, or full novels. Such models, when aggressively quantized (e.g., to 4 bits), could theoretically handle significantly more context under the same memory budget.

However, to our knowledge, only Kurtic et al. (2024) have benchmarked long-context performance, and their study was confined to the RULER (Hsieh et al., 2024a) dataset. Critically, they only reported a single, aggregated score and did not provide a breakdown of performance at different context lengths, which can mask significant performance drops on very long inputs.

To address this gap, we conduct a systematic analysis of quantization performance across multiple models, quantization methods, and benchmark tasks. This includes five datasets across 9.7K test examples, with key factors varying such as context length, language, and output constraints. We evaluate five quantization methods (FP8, GPTQ-int8, AWQ-int4, GPTQ-int4, and BNB-NF4) on five models (Llama-3.1 8B and 70B; Qwen-2.5 7B, 32B, and 72B), comparing against their BF16 base-lines. For long-context input tasks, we use RULER (Hsieh et al., 2024a) and ONERULER (Kim et al., 2025) to assess retrieval, and NOCHA (Karpinska et al., 2024) to test reasoning over book-length texts. For long-form output tasks, we employ FACTSCORE (Min et al., 2023) to evaluate factuality in biography generation and CS4 (Atmakuru

et al., 2024) to assess constrained story generation (see Figure 2). The results reveal the following:

1. **The 8-bit quantized models remain mostly robust**, with FP8 and GPTQ-int8 showing average drops of 0.2% and 0.8%, respectively.² **4-bit methods incur larger losses**, with AWQ showing 1.8%, GPTQ-int4 2.7%, and BNB-nf4 6.9% average drops. BNB-nf4 can be particularly inaccurate on long contexts, with up to 59% loss on ONERULER for Llama-3.1 70B.
2. **Quantization degrades long-input retrieval performance**, with drops worsening as input length increases. 4-bit quantization shows the most significant drop, with an average decrease of up to 23% across models at 128K tokens.
3. **The adverse effects of quantization on performance are more severe in languages other than English**. Our long-context retrieval experiments reveal that, under quantization, accuracy in these languages can drop up to five times more than in English.
4. **The impact of quantization is not uniform between models**. While Qwen-2.5 72B shows minimal loss under BNB-nf4 across tasks, similarly sized Llama-3.1 70B suffers a 32% drop.

2 Experimental setup

This section outlines our experimental setup. We evaluate five models from two model families

²This result is in line with prior studies (Lee et al., 2024; Kurtic et al., 2024).

(Llama and Qwen) across six precision settings on five benchmarks. Additional details are in §A.

Models: We evaluate five open-weight models from two families: **Llama 3.1** 8B and 70B (Grattafiori et al., 2024) and **Qwen 2.5** 7B, 32B, and 72B (Qwen, 2024c; Yang et al., 2024). All models support up to 128K input tokens.³

Quantization methods: We use BF16 as the full-precision baseline and compare four quantization methods at 4-bit and 8-bit precisions against it:

1. **AWQ:** 4-bit integer (W4A16)⁴ quantization (Lin et al., 2024), referred to as AWQ-int4.
2. **BNB:** 4-bit floating-point quantization using bit-sandbytes (W4A16) (Dettmers et al., 2023), referred to as BNB-nf4.
3. **GPTQ:** (a) 4-bit integer (W4A16), referred to as GPTQ-int4, and (b) 8-bit integer (W8A16), referred to as GPTQ-int8 (Frantar et al., 2023).
4. **FP8:** Dynamic 8-bit floating-point quantization (W8A8) (Micikevicius et al., 2022), referred to as FP8.

In total, we benchmark five models across five quantization setups and the BF16 baselines, resulting in 30 configurations.

Inference setup: All models are run using vLLM (Kwon et al., 2023) on one to four A100-80G GPUs, except for the FP8 models, which require H100 GPUs due to the lack of FP8 support on A100 (vLLM, 2024). All experiments use greedy decoding with a temperature of 0.0 to ensure deterministic outputs.

Benchmark tasks: The BF16 models and their quantized variants are evaluated on five benchmarks: (1) **RULER** (Hsieh et al., 2024a) for English long-context retrieval, (2) **ONERULER** (Kim et al., 2025) for long-context retrieval across languages, (3) **NOCHA** (Karpinska et al., 2024) for long-context reasoning, (4) **FACTSCORE** (Min et al., 2023) for factual biography generation, and (5) **CS4** (Atmakuru et al., 2024) for constrained creative story generation. The first three tasks evaluate model performance on long inputs, while the last two focus on long-form outputs. See §B for more details.

Dataset: **RULER** (Hsieh et al., 2024a)

³Qwen-2.5 models originally support 32k tokens but can be extended to 128K using YARN (Peng et al., 2024). See §A.2 for details.

⁴WxAy denotes x-bit weights and y-bit activations.

Task: **RULER** evaluates model’s *effective context window* with needle-in-a-haystack (NIAH) retrieval tasks in English. The model must locate one or more target key-value pairs (the “needles”) hidden in passages of varying lengths (the “haystacks”). Following Yen et al. (2025), the task includes three variants: (1) *multi-key needle*, (2) *multi-key UUID*, and (3) *multi-value*. Each variant is tested at three context lengths (8K, 64K, and 128K),⁵ with emphasis on longer contexts (64K and 128K). Each {task, context-length} combination includes 100 examples, yielding 900 test examples in total.

Metric: Following (Hsieh et al., 2024a), performance is measured using exact match accuracy against the gold answer.

Dataset: **ONERULER** (Kim et al., 2025)

Task: **ONERULER** extends **RULER** by evaluating NIAH retrieval tasks not only in English but also across multiple other languages. It also introduces no-needle scenarios, where models must explicitly predict the absence of a needle. **ONERULER** serves as an important complement to **RULER** in light of the performance degradation of quantized models observed by Marchisio et al. (2024) on non-Latin scripts. Each model is evaluated on two tasks (NIAH-single and NIAH-none; see §B), three context lengths (8K, 64K, and 128K),⁵ and 26 languages. Each {task, context-length, language} combination consists of 50 examples, totaling 7,800 test examples.

Metric: As with **RULER**, exact match accuracy is used as the metric.

Dataset: **NOCHA** (Karpinska et al., 2024)

Task: **NOCHA** evaluates long-context reasoning by asking models to verify claims against book-length contexts. Each example includes a minimal pair of true and false claims with subtle but meaningful differences. Models are prompted to explain their reasoning before making a decision. The evaluation includes data points with contexts up to 128K tokens. In total, the dataset contains 723 claim pairs drawn from 46 novels and three short story collections.

⁵Unless explicitly stated, we report average performance across 64K and 128K context lengths, as our focus is on long-context scenarios. Results at 8K are included only in a targeted ablation to highlight the impact of extended context.

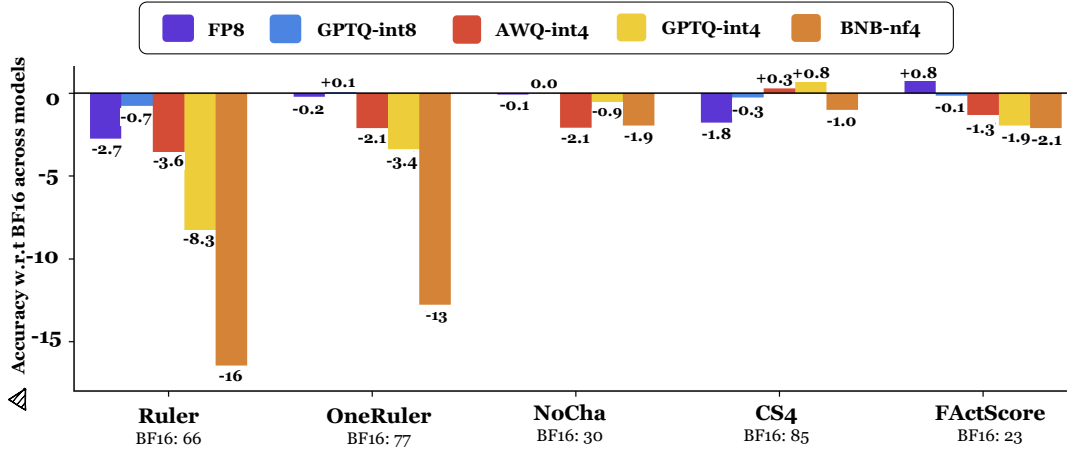


Figure 3: Average performance of each quantization method by task, computed by averaging across all models. Performance is reported using Δ -accuracy, representing the difference relative to the BF16 baseline. The BF16 baseline values (averaged across models for each task) are shown below the task names. For ONERULER and RULER, only long-context results from the 64K and 128K context lengths are included in the aggregation.

Metric: Following Karpinska et al. (2024), we report accuracy at the claim-pair level: an answer is considered correct only if the model correctly labels both claims in the pair (i.e., identifies which is true and which is false).

Dataset: FACTSCORE (Min et al., 2023)

Task: FACTSCORE evaluates a model’s ability to generate factual long-form biographies for named entities. The evaluation covers 100 entities, evenly distributed across three frequency categories: *very rare*, *medium*, and *very frequent*.⁶

Metric: We report VERIScore (Song et al., 2024) as the factuality metric. Responses without verifiable content receive a score of zero.

Dataset: CS4 (Atmakuru et al., 2024)

Task: CS4 evaluates a model’s instruction-following and story-writing abilities by measuring how well it generates coherent short stories under increasing levels of constraint (9, 23, and 39 constraints). Each model generates 150 stories in total, with 50 per constraint level.

Metric: Following Atmakuru et al. (2024), we use GPT-4o to evaluate constraint satisfaction (Liu et al., 2023), computed per story as $100 \times \frac{\text{\# of satisfied constraints}}{\text{\# of total constraints}}$. We use BoookScore (Chang et al., 2024) to evaluate coherence and internal consistency of the generated stories.⁷

⁶We adopt the definition from FACTSCORE (Min et al., 2023), which uses Wikipedia occurrences as frequency.

⁷We use gpt-4o-2024-08-06 as the judge LLM for constraint satisfaction and BoookScore.

3 Results and discussion

This section presents the results of how quantization affects performance on the benchmarks in §2. We examine results at the following levels of detail: model family, task category, and sub-task settings such as context length and language. For each setting, we report Δ -accuracy, defined as the change in performance between the BF16 model and its quantized version:

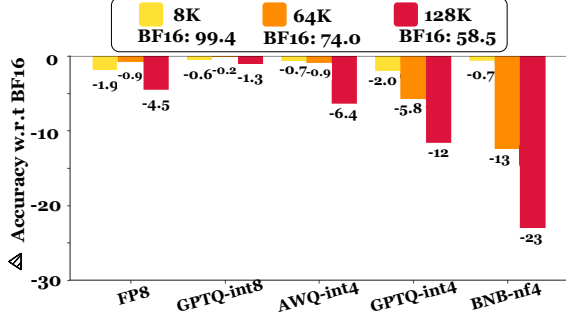
$$\Delta\text{-accuracy} = \text{acc}_{\text{quantized}} - \text{acc}_{\text{BF16}}$$

We summarize key findings below.

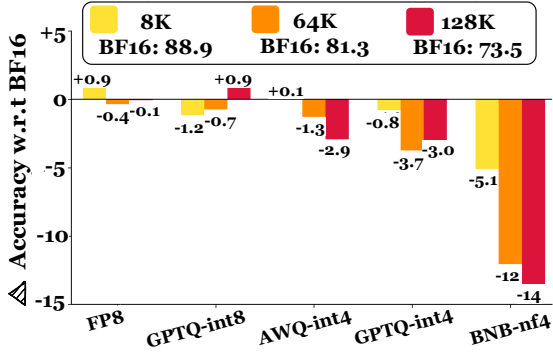
Long-context retrieval performance degrades with 4-bit quantization. Long-context retrieval tasks are particularly sensitive to aggressive compression (Figure 3, $p < 0.05$).⁸ Specifically, accuracy on RULER and ONERULER drops by up to 16% when model precision is reduced to 4-bits, regardless of the quantization method. For the NOCHA long-context reasoning benchmark, the overall accuracy decrease is smaller. This is likely because the accuracy of the baseline BF16 is already near chance level around 30%, leaving little room for further degradation.⁹ Nevertheless, model performance still tends to decline slightly for most quantization methods as input length increases from $<75K$ to $>75K$ tokens (Figure 4c). While the extent of this drop varies, it can be observed across

⁸We report the details of statistical analysis reported throughout this section in §F.

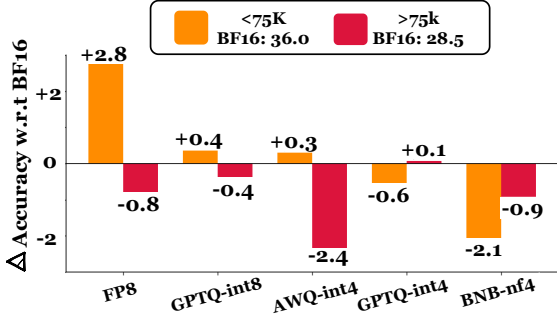
⁹As the model has to evaluate correctly two claims in the pair, the random baseline is at 25%.



(a) RULER: drop in performance at three context lengths.



(b) ONERULER: drop in performance at three context lengths.

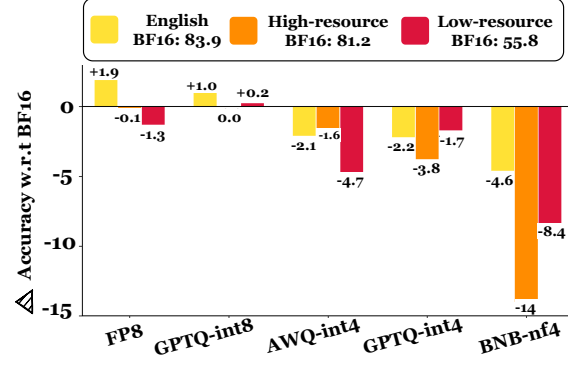


(c) NOCHA: drop in performance at varying context lengths.

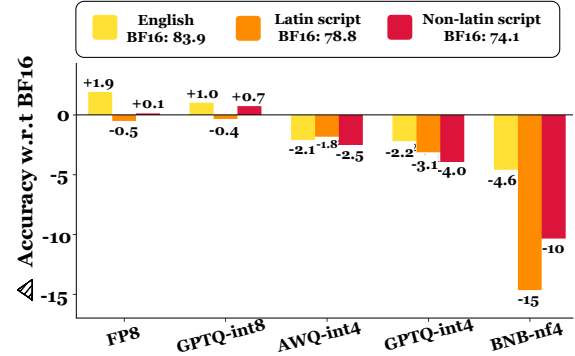
Figure 4: Quantization performance by context length for three long-context benchmarks (RULER, ONERULER, and NOCHA), shown as Δ -accuracy values averaged across models for each quantization method. The corresponding BF16 baseline scores (model-averaged) are shown in the legend.

most models for all three tasks (Figure 6). This sensitivity of quantization performance at long contexts is possibly due to the rounding errors in RoPE embeddings accumulated over long context (Wang et al., 2025a).

Quantization Degrades Accuracy More Severely for High- and Low-Resource Languages. Our analysis of ONERULER performance groups languages into three categories: English, other high-



(a) ONERULER: drop in performance by language resource level



(b) ONERULER: drop in performance by script family

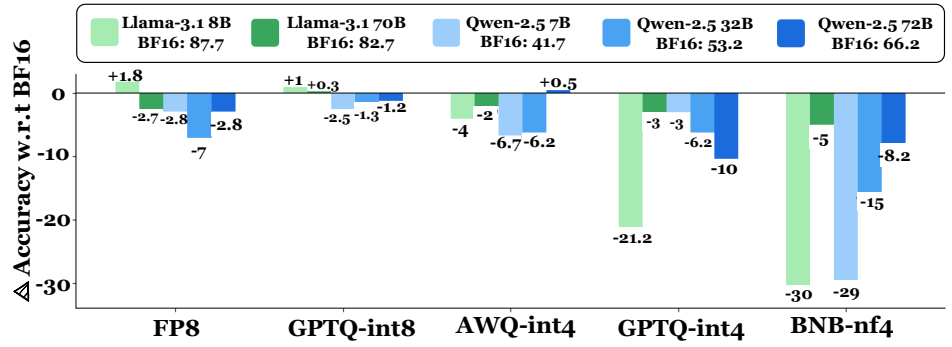
Figure 5: **ONERULER**: Average performance of each quantization method on each model, aggregated over long-context results (64K and 128K lengths). Performance is shown as Δ -accuracy, the difference relative to the BF16 baseline.

resource languages, and low-resource languages.¹⁰ While 8-bit quantization methods like FP8 and GPTQ-int8 show minimal impact across all groups (ranging from +1.9% to -1.3% vs. the BF16 baseline), 4-bit methods uniformly degrade performance (see Figure 5a).

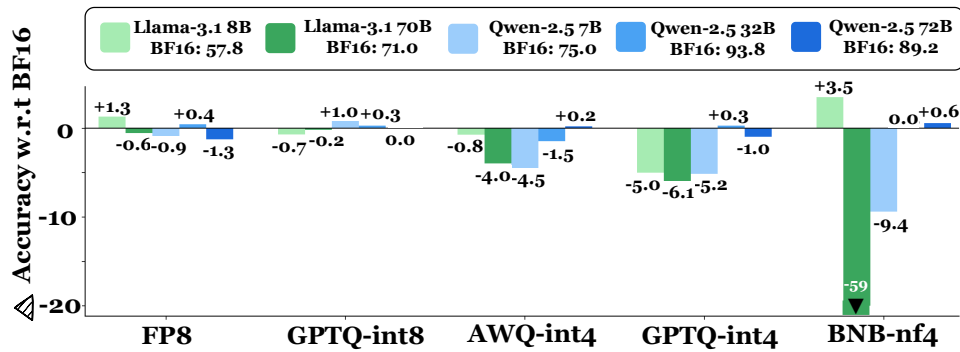
Specifically, larger drops occur for **low-resource languages** when using AWQ-int4 (-4.7%) and BNB-nf4 (-8.4%). Similarly, **high-resource languages** face significant performance decreases with GPTQ-int4 (-3.8%) and particularly with BNB-nf4 (-14%). While performance on English also drops, the degradation range is narrower (-2.1% to -4.6%). Post-hoc tests confirm that drops for non-English languages are statistically significant ($p < 0.05$, Table 14 in §F).¹¹

¹⁰We adopt the definition from ONERULER (Kim et al., 2025), which classifies languages with under 250k Wikipedia articles (Hindi, Sesotho, Swahili, and Tamil) as low-resource.

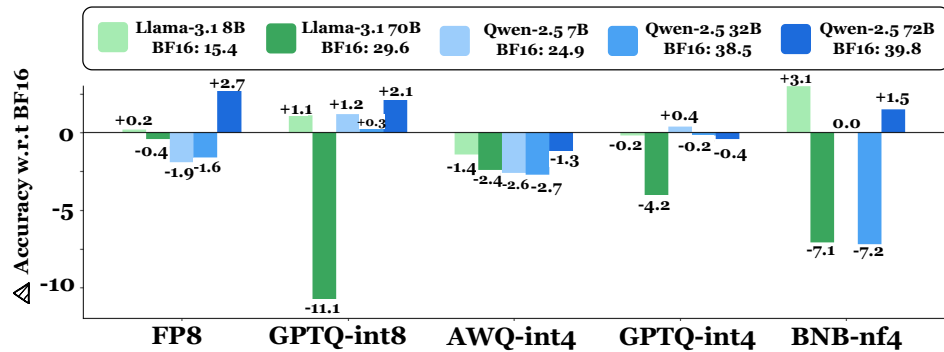
¹¹We also report performance by language script following Marchisio et al. (2024), in (Figure 5b). However, we do not observe a consistent pattern where quantization dispropo-



(a) **RULER**: Average Δ -accuracy relative to BF16 across models.



(b) **ONERULER**: Average Δ -accuracy relative to BF16 across models.



(c) **NOCHA**: Average Δ -accuracy relative to BF16 across models.

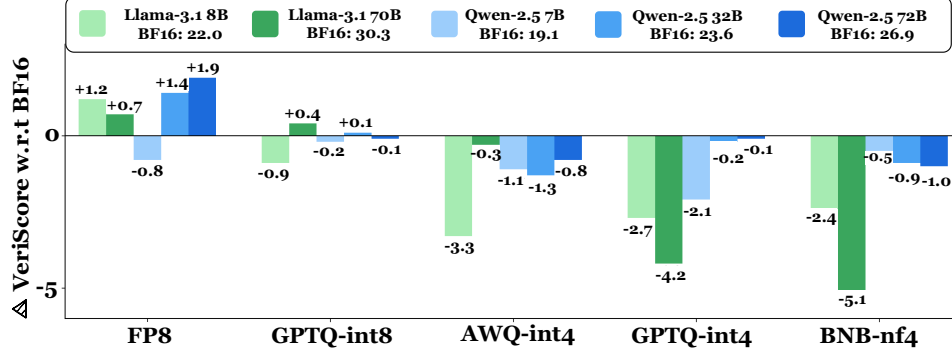
Figure 6: Average performance of quantization methods on long-context tasks, by model. The BF16 baseline results over which deltas are computed are given in the legend. RULER and ONERULER results are computed over long (64K, 128K) context lengths only.

Factuality is preserved at 8-bit but drops at 4-bit precision. On the FACTSCORE biography generation task, 8-bit quantizations remain essentially on par with the BF16 baseline (FP8 +0.8%, GPTQ-int8 −0.1%; see Figure 3). In contrast, 4-bit precision reduces scores: AWQ-int4 (−1.3%), GPTQ-int4 (−1.9%), BNB-nf4 (−2.1%),

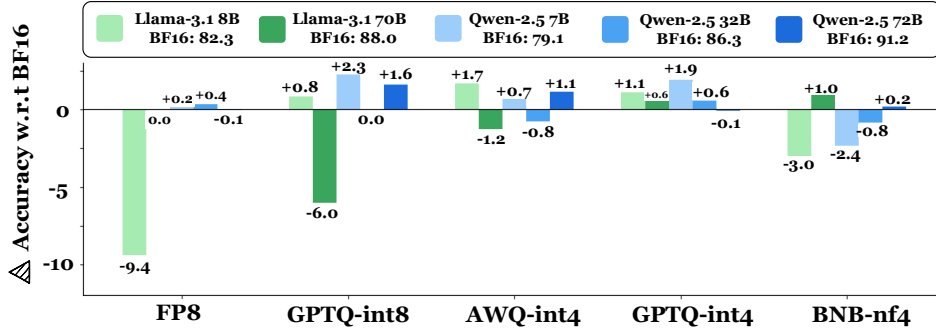
with model-specific variation (Figure 7a).¹² Figure 8 shows that the largest drops (up to −3.7%) occur for very frequent entities, while medium- and low-frequency entities are less affected. These smaller declines likely reflect their already low baseline accuracy, which leaves limited room for further degradation. All figures show that drops

portionately affects languages with non-Latin scripts. Instead, the performance drops vary depending on the quantization method: for example, GPTQ-int4 results in slightly larger drops for non-Latin script languages, while BNB-nf4 leads to greater degradation for languages using the Latin script.

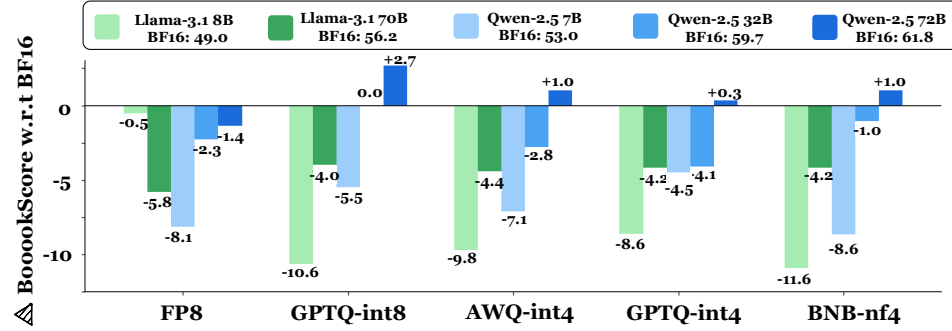
¹²Most of this decline stems from higher abstention rates: responses with no verifiable claim are counted as errors. For example, Llama-3.1-70B abstains in 29% of BF16 outputs, 30% of FP8, and 52% of BNB-nf4. Full statistics appear in Figure 25 (§C.4).



(a) FACTSCORE: Average Δ -VERIScore across models.



(b) CS4: Average Δ -constraint-satisfaction across models.



(c) CS4: Average Δ -BooookScore across models.

Figure 7: Average performance of quantization methods on long-form generation tasks, by model. The BF16 baseline results over which deltas are computed are given in the legend.

are small, with only BNB-nf4 showing statistically significant effects on VERIScore, no claim rates and effect of entity frequency ($p < 0.05$, Tables 20, 22 & 24 in §F).

GPTQ-int8, GPTQ-int4, and AWQ-int4 preserve constraint satisfaction on CS4, while FP8 and BNB-nf4 incur losses. On the CS4 task, most quantization methods preserve constraint satisfaction. GPTQ and AWQ methods perform on par with the BF16 baseline, with scores ranging from +0.8% to -0.3% (Figure 3). In contrast, FP8 and BNB-nf4 incur statistically significant drops of

-1.8% and -1.0%, respectively ($p < 0.05$; Tables 25 and 26 in §F). For these two methods, the degradation worsens as the number of constraints increases (Figure 9, $p < 0.05$; Table 28 in §F). The pronounced drop from FP8 is particularly noteworthy, as it contrasts with its strong performance on other tasks.

Quantization degrades story coherence, especially in smaller models. Story coherence, evaluated by BoookScore, declines with quantization for all models except Qwen-2.5 72B (Figure 7c). Within each model family, the drop is generally more pronounced for smaller models. For instance,

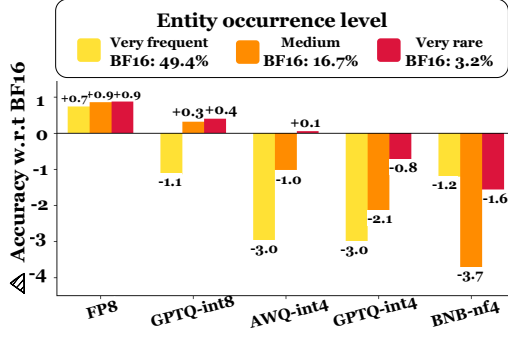


Figure 8: **FACTSCORE**: Average Δ -VERIScore (relative to BF16), varying by level of occurrence of the input entity

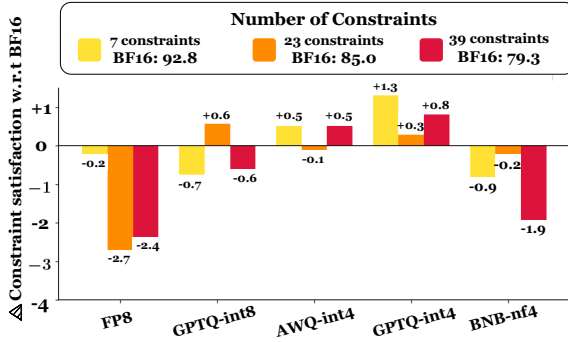


Figure 9: **CS4**: Average Δ -constraint-satisfaction of different quantization methods relative to BF16 on the CS4 benchmark across varying input constraints.

in the Qwen-2.5 family, the 7B model loses 6.8% on average, while the 32B model loses only 2.0% and the 72B model gains 0.7%. This pattern is also visible in the Llama-3.1 family: the 70B model sees a modest 4.5% average drop, while the 8B model shows a much larger 10% loss. The notable exception is FP8 quantization, which preserves coherence exceptionally well for Llama-3.1 8B, resulting in a mere 0.5% decline.

FP8 and GPTQ-int8 preserve overall accuracy most effectively, while 4-bit quantization causes noticeable degradation. FP8 and GPTQ-int8 have the smallest average accuracy drops (Δ -accuracy 0.8% and 0.2%, respectively) relative to BF16 across tasks (Figure 1). Post-hoc tests for RULER (Table 4), ONERULER (Table 10), NOCHA (Table 15), and FACTSCORE (Table 10) confirm that neither differs significantly from BF16 ($p > 0.05$). In contrast to 8-bit quantization methods, 4-bit methods exhibit larger losses ($p < 0.05$ in Tables 4, 10, 16, 20 and 26 in §F). AWQ-int4 and GPTQ-int4 suffer average drops of 1-3%, while BNB-nf4 loses an average of 6.9% (Figure 1). Spe-

cific cases show very large drops: on ONERULER, Llama-3.1 70B under BNB-nf4 drops by 59% (Figure 6b); on RULER, Qwen-2.5 8B, Qwen-2.5 32B, and Llama-3.1 8B each drop by over 15% (Figure 6a), etc. Notably, BNB-nf4 is the default 4-bit quantization in libraries like HuggingFace and vLLM, so its pronounced performance drops have significant practical implications. The results highlight that 4-bit quantization, especially BNB-bf4, should be used with caution. Overall, as seen in Figure 1, the average trend across quantization methods indicates that FP8 and GPTQ-int8 are comparable with BF16, with the 4-bit quantization methods ranked as AWQ-int4 > GPTQ-int4 > BNB-nf4.

The effect of quantization varies with model and task. Our results demonstrate that quantization effects are highly context-dependent, varying significantly with both model family and the evaluation task. This model-specific behavior is evident even between models of comparable size. On the ONERULER benchmark, for example, BNB-nf4 quantization leads to a 59% accuracy drop for Llama-3.1 70B, whereas the similarly-sized Qwen-2.5 72B gains 0.6% with the same method (Figure 6b). Sensitivity also varies within a model family; the Llama-3.1 models are generally more affected by quantization than the Qwen-2.5 models, whose larger 32B and 72B variants show higher resilience.¹³ This effect can even be unpredictable, as seen when BNB-nf4 causes a 32% performance drop on Llama-3.1 70B but a 0.6% gain on Llama-3.1 8B (Figure 11).

Furthermore, a model’s robustness to quantization is also task-dependent. For instance, that same Llama-3.1 70B model, when quantized with GPTQ-int8, drops 11.1% on NOCHA (Figure 6c) but shows no comparable decline on RULER or ONERULER (Figure 6a, Figure 6b). These findings differ from prior work that examined a single model family and observed only minor performance drops (Kurtic et al., 2024; Huang et al., 2024; Kumar et al., 2025; Jin et al., 2024), highlighting the importance of evaluating a diverse set of models.

Why does quantization affect long-context tasks? While studies on short context tasks show minimal degradation, we observe larger drops, particularly for tasks involving long-context inputs. A

¹³For AWQ and GPTQ, the Llama-3.1 and Qwen-2.5 models originate from different sources and may have been quantized with different calibration sets. The same overall patterns, however, also appear in BNB-fn4 and FP8 quantizations.

possible explanation for this sink could be the accumulation of errors in RoPE embeddings (Wang et al., 2025a). This issue could be potentially addressed by using long-context samples during the calibration stage of quantization (Williams and Aletras, 2023), including multilingual calibration data, and applying attention-calibration techniques like found-in-the-middle to reduce attention errors (Hsieh et al., 2024b). We leave the evaluation of these strategies for future work.

4 Related works

Prior studies on quantization effects: Recent work shows that 8-bit quantization preserves performance, while 4-bit methods incur only marginal accuracy loss (Kurtic et al., 2024; Badshah and Sajjad, 2024; Liu et al., 2024b; Jin et al., 2024; Huang et al., 2024; Xu et al., 2024; Li et al., 2024; del Rey et al., 2025; Zheng et al., 2025). Lee et al. (2024) and Kurtic et al. (2024) investigate Llama-3.1 and find that FP8 is nearly lossless, while AWQ-int4 and GPTQ-int4 exhibit small degradations, particularly in reasoning, coding tasks and on Arena-Hard. Results on Mistral (Badshah and Sajjad, 2024), Baichuan (Liu et al., 2024b), and Qwen-Chat (Jin et al., 2024) are also similar, with losses rarely exceeding 3% and diminishing as model size grows. Results on safety, alignment, and interpretability suggest that 4-bit quantization may (Egashira et al., 2024) or may not (Xu et al., 2024) lead to drops in safety, protect against data-poisoning (Hussain et al., 2025), reverse unlearning (Zhang et al., 2025), and have mixed results on interpretability (Wang et al., 2025b). Some studies have also shown that models are robust to quantization even at 3-bit precision or lower (Chee et al., 2023; Ma et al., 2024; Egiastian et al., 2024; Kumar et al., 2025). However, these evaluations are generally limited in terms of the input context or the output token ranges, leaving longer-range effects unknown. A significant body of work has studied accuracy trade-offs using different quantization methods on academic benchmarks (Yao et al., 2024; Liu et al., 2024a; Gong et al., 2024a,b). However, these benchmarks may not fully reflect performance in realistic scenarios, such as those involving long-form reasoning or multilingual settings. For instance, Li et al. (2025) and Liu et al. (2025) report significant drops for GPTQ-int4 and AWQ-int4 on such tasks. For multilingual LLMs, Marchisio et al. (2024) observe increasing drops on reasoning for non-Latin

script languages upon quantization. We develop this picture further by evaluating long-form generation (CS4, FACTSCORE) and across languages in ONERULER.

Long-context focus: To our knowledge, only Kurtic et al. (2024) evaluate quantization on long-context tasks, and they report coarse averages over RULER (which includes context lengths as low as 8K) to conclude negligible performance loss. Our analysis differs by focusing on the effects of parameters like context length, task-type, especially at context lengths >64K. While they evaluate only the Llama-3.1 family and use a single 4-bit quantization (GPTQ-int4), we extend the scale of long context analysis by including ONERULER and NOCHA, the Qwen-2.5 model family, and the AWQ-int4 and BNB-nf4 quantizations. We find that Qwen-2.5 and Llama-3.1 respond to quantization very differently, with Llama-3.1 showing larger drops upon quantization (Figure 6a). In addition, the AWQ-int4 quantization performs better than GPTQ-int4, while the popular BNB-nf4 quantization performs worse (Figure 3).

5 Conclusion

In this paper, we presented a systematic evaluation of the performance of five open-weight models (7B to 72B parameters) across five quantization methods (4-bit and 8-bit). We conducted experiments on five tasks requiring the models to either process long-context inputs (in English or other languages) or return long-form outputs in English. While prior work often suggests 8-bit quantization is nearly lossless and 4-bit methods are *good enough*, our findings demonstrate that the impact of quantization becomes more pronounced in long-context scenarios. In particular, we observed that model’s performance under 4-bit quantization tends to degrade progressively as input context length increases. 4-bit quantization has a significant impact when the input is in a language other than English. In contrast, 8-bit precision methods largely preserve accuracy across context lengths. Furthermore, we observed considerable heterogeneity in quantization effects—the performance impact of reduced precision varies significantly across different model architectures and task types. Taken together, these results indicate that quantization should be applied with careful consideration in long-context applications and in multilingual setups.

Limitations

While we conduct a large-scale evaluation of quantized LLMs on long-context and long-form tasks, we identify the following limitations and hope future work will improve on these aspects to enable more comprehensive evaluations of quantization methods.

Scope of experiments: Our evaluations cover five quantization methods and five LLMs, with the largest model having 72 billion parameters. We largely followed publicly available, off-the-shelf quantized models and recipes (e.g., from the official Qwen repository, HuggingQuants, and NeuralMagic) and did not explore tuning each quantization method further. As a result, our findings might not fully generalize to other models or to quantization techniques beyond those tested.

Evaluation methodology and limited output lengths: Although we tested input contexts as long as 128K tokens, we restricted the generated outputs in our experiments to relatively short passages (< 650 tokens). We made this choice because our evaluation relies solely on automated metrics and LLM-based judge models, which are not reliable for evaluating very long text. Consequently, without human annotations, our assessment may overlook subtle errors in logical flow or nuanced instruction-following behavior over long-form outputs.

Ethical Considerations

We rely solely on publicly available datasets that have been vetted and validated in prior peer-reviewed studies. We do not introduce any new data collection or human subject research procedures. Our paper poses no ethical concerns.

Acknowledgments

This project was partially supported by awards IIS-2046248, IIS-2312949, and IIS-2202506 from the National Science Foundation (NSF). We also thank the UMass NLP community for their insightful comments.

AI assistants like ChatGPT were used to polish writing and debug code for experiments. They were not used to generate paper content or result tables and plots.

References

- Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. In *6th Workshop on Narrative Understanding (WNU)*.
- Sher Badshah and Hassan Sajjad. 2024. [Quantifying the capabilities of llms across scale and precision](#). *arXiv preprint arXiv:2405.03146*.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [BoookScore: A systematic exploration of book-length summarization in the era of LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. 2023. QuIP: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36:4396–4429.
- Databricks. 2023. [LLM inference performance engineering: Best practices](#). Accessed: May 18, 2025.
- Santiago del Rey, Paulo Sérgio Medeiros dos Santos, Guilherme Horta Travassos, Xavier Franch, and Silverio Martínez-Fernández. 2025. Aggregating empirical evidence from data strategy studies: a case on model quantization. *arXiv preprint arXiv:2505.00816*.
- 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.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient fine-tuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. 2024. [Exploiting LLM quantization](#). In *ICML 2024 Next Generation of AI Safety Workshop*.
- Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. 2024. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.

- Jared Fernandez, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Luccioni, and Emma Strubell. 2025. [Energy considerations of large language model inference and efficiency optimizations](#). In *LTI Student Research Symposium 2025*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *The Eleventh International Conference on Learning Representations*.
- Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Dacheng Tao, and Xianglong Liu. 2024a. [LLMC: Benchmarking large language model quantization with a versatile compression toolkit](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 132–152, Miami, Florida, US. Association for Computational Linguistics.
- Zhuocheng Gong, Jiahao Liu, Jingang Wang, Xunliang Cai, Dongyan Zhao, and Rui Yan. 2024b. 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.
- Google Cloud. 2024. [Accelerating AI inference with Google Cloud TPUs and GPUs](#). Accessed: May 18, 2025.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekeshe, Fei Jia, and Boris Ginsburg. 2024a. [RULER: What’s the real context size of your long-context language models?](#) In *First Conference on Language Modeling*.
- Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. 2024b. Found in the middle: Calibrating positional attention bias improves long context utilization. *arXiv preprint arXiv:2406.16008*.
- Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. [An empirical study of LLaMA3 quantization: From LLMs to MLLMs](#). *Visual Intelligence*, 2(1):36.
- Aftab Hussain, Sadegh AlMahdi Kazemi Zarkouei, Md Rafiqul Islam Rabin, Mohammad Amin Alipour, Sen Lin, and Bowen Xu. 2025. Capturing the effects of quantization on trojans in code LLMs. *arXiv preprint arXiv:2505.14200*.
- Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. [A comprehensive evaluation of quantization strategies for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12186–12215.
- Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [One thousand and one pairs: A “novel” challenge for long-context language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085.
- Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joon-suk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. 2023. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36:36187–36207.
- Yekyung Kim, Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. [One ruler to measure them all: Benchmarking multilingual long-context language models](#). *arXiv preprint arXiv:2503.01996*.
- Tanishq Kumar, Zachary Ankner, Benjamin Frederick Spector, Blake Bordelon, Niklas Muennighoff, Manish Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. 2025. [Scaling laws for precision](#). In *The Thirteenth International Conference on Learning Representations*.
- Eldar Kurtic, Alexandre Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. 2024. [“Give me BF16 or give me death”? Accuracy-performance trade-offs in llm quantization](#). *arXiv preprint arXiv:2411.02355*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. 2024. [A comprehensive evaluation of quantized instruction-tuned large language models: An experimental analysis up to 405b](#). *arXiv preprint arXiv:2409.11055*.
- 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](#). In *Proceedings of the 41st International Conference on Machine Learning*, pages 28480–28524.
- Zhen Li, Yupeng Su, Runming Yang, Zhongwei Xie, Ngai Wong, and Hongxia Yang. 2025. [Quantization meets reasoning: Exploring LLM low-bit quantization degradation for mathematical reasoning](#). *arXiv preprint arXiv:2501.03035*.

- 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 on-device llm compression and acceleration](#). *Proceedings of Machine Learning and Systems*, 6:87–100.
- Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2024a. [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.
- Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. 2025. Quantization hurts reasoning? an empirical study on quantized reasoning models. *arXiv preprint arXiv:2504.04823*.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yijun Liu, Yuan Meng, Fang Wu, Shenhao Peng, Hang Yao, Chaoyu Guan, Chen Tang, Xinzhu Ma, Zhi Wang, and Wenwu Zhu. 2024b. [Evaluating the generalization ability of quantized LLMs: Benchmark, analysis, and toolbox](#). *arXiv preprint arXiv:2406.12928*.
- Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. [Power hungry processing: Watts driving the cost of AI deployment?](#) In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, page 85–99, New York, NY, USA. Association for Computing Machinery.
- Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Lifeng Dong, Ruiping Wang, Jilong Xue, and Furu Wei. 2024. The era of 1-bit LLMs: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 1.
- Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. [How does quantization affect multilingual LLMs?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947.
- Meta AI. 2024. [Meta Llama: Introducing quantized lightweight Llama 3 models](#). Accessed: 2025-05-18.
- Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, et al. 2022. [FP8 formats for deep learning](#). *arXiv preprint arXiv:2209.05433*.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Shinichi Nakagawa, Paul CD Johnson, and Holger Schielzeth. 2017. The coefficient of determination r^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213.
- NeuralMagic. 2024. [LLMCompressor GitHub repository](#). Accessed: April 7, 2025.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. [YaRN: Efficient context window extension of large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Qwen. 2024a. [Qwen-2.5 quantization collection on HuggingFace](#). Accessed: April 7, 2025.
- Qwen. 2024b. [Qwen documentation](#). <https://qwen.readthedocs.io/en/latest/deployment/vllm.html#extended-context-support> (Deployment with vLLM: Extended Context Support). Accessed: March 4, 2025.
- Qwen. 2024c. [Qwen2.5: A party of foundation models](#).
- Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devsh Tiwari, and Vijay Gadeppally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE.
- Robert Shaw, Mark Kurtz, Sara Adkins, and Benjamin Fineran. 2024. [LLM Compressor: Faster inference with vLLM](#). Published by Red Hat; authors affiliated with NeuralMagic. Accessed: April 7, 2025.
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. [VeriScore: Evaluating the factuality of verifiable claims in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474.
- vLLM. 2024. [vLLM documentation](#). <https://docs.vllm.ai/>, including subsections: <https://docs.vllm.ai/en/latest/features/quantization/fp8.html> and <https://docs.vllm.ai/en/latest/features/quantization/bnb.html>. Accessed: April 7, 2025.
- Haonan Wang, Qian Liu, Chao Du, Tongyao Zhu, Cunxiao Du, Kenji Kawaguchi, and Tianyu Pang. 2025a. [When precision meets position: BFloat16 breaks down RoPE in long-context training](#). *Transactions on Machine Learning Research*.

- Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2025b. Through a compressed lens: Investigating the impact of quantization on LLM explainability and interpretability. *arXiv preprint arXiv:2505.13963*.
- Peter H Westfall, Wesley O Johnson, and Jessica M Utts. 1997. A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, 84(2):419–427.
- Miles Williams and Nikolaos Aletras. 2023. On the impact of calibration data in post-training quantization and pruning. *arXiv preprint arXiv:2311.09755*.
- 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*.
- Zhichao Xu, Ashim Gupta, Tao Li, Oliver Bentham, and Vivek Srikumar. 2024. [Beyond perplexity: Multi-dimensional safety evaluation of llm compression](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15359–15396.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. 2024. [Exploring post-training quantization in llms from comprehensive study to low rank compensation](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. [HELMET: How to evaluate long-context models effectively and thoroughly](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2025. [Catastrophic failure of LLM unlearning via quantization](#). In *The Thirteenth International Conference on Learning Representations*.
- Xingyu Zheng, Yuye Li, Haoran Chu, Yue Feng, Xudong Ma, Jie Luo, Jinyang Guo, Haotong Qin, Michele Magno, and Xianglong Liu. 2025. An empirical study of Qwen3 quantization. *arXiv preprint arXiv:2505.02214*.
- Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.
- Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

Appendix

A Experimental setup: additional details

A.1 Obtaining quantized models

Our objective was to study how easily available quantized models handle long-context and long-form tasks. So, we primarily relied on quantized models provided widely by RedHat AI (former NeuralMagic; (Kurtic et al., 2024)), HuggingQuants, and Qwen-2.5’s official repository, since these are popular resources for quantized versions of Llama-3.1 and Qwen-2.5, with up to 300K downloads as of July 2025.

Table 1 provides an overview of how quantized models were obtained for this study. Some specifics are detailed below.

For Llama-3.1 70B, we observed noticeable accuracy degradation upon using the default GPTQ-int8 (W8A16) quantization recipe. Kurtic et al. (2024) encountered the same issue with the W8A8 version and resolved it by integrating SmoothQuant (Xiao et al., 2023) into their W8A8 GPTQ-int8 recipe. We use this same quantization recipe from Kurtic et al. (2024) and performed a W8A16 GPTQ-int8 quantization, by which we were able to largely retain the pre-quantization accuracy.

For Qwen-2.5 models, FP8-Dynamic models were created locally using LLMCompressor (Shaw et al., 2024; NeuralMagic, 2024), borrowing the same quantization recipe used by Kurtic et al. (2024) for Llama-3.1.

For BNB-nf4, we use in-flight quantization directly from the full-precision checkpoints (vLLM, 2024). We launch vLLM with quantization=bitsandbytes. This converts BF16 weights to nf4 blocks at load time, and then runs inference by de-quantizing those blocks on-chip before each GEMM, while activations remain in BF16. We do not experiment with the 8-bit integer BNB quantization (Dettmers et al., 2022) since it is not supported on vLLM.¹⁴

A.2 Using Qwen-2.5 for long context tasks:

Though Qwen-2.5 has a context length of 128K, the default configuration is set to support context lengths only up to 32K; for longer contexts, the official Qwen documentation (Qwen, 2024b) suggests employing YARN (Peng et al., 2024) to extrapolate the RoPE scaling and achieve 4× the context

length. We adopt the same approach and modify Qwen model configuration when running on context lengths larger than 32K.

A.3 Computational Cost

We estimate the approximate compute consumption as follows: NOCHA: 1800 hours, RULER: 500 hours, ONERULER: 3200 hours, CS4: 180 hours, and FACTSCORE: 100 hours, totaling roughly 5800 GPU hours on NVIDIA A100 GPUs.

B Benchmarks: additional details

In this section, we present more details about each benchmark task, adding to the description in §2.

RULER: While Kurtic et al. (2024) evaluate quantized LLMs on RULER and report coarse-grained, aggregated results across quantized models and context lengths, we perform a more detailed analysis. We follow Yen et al. (2025) to select our datapoints for the multi-key needle, multi-key UUID, and multi-value tasks from the RULER benchmark. In the multi-key task, the model must retrieve the value associated with a specific key given in the query, while ignoring others which serve as distractors. The multi-key UUID task tests a similar ability, but the values are 32-character long-form strings (UUIDs), making it harder to match or guess, testing the model’s ability to precisely extract the correct value without partial matching. Finally, the multi-value task increases the difficulty further by requiring the models to retrieve multiple values associated with the same key in the haystack. While the tokenizers for Llama-3.1 and Qwen-2.5 differ, Yen et al. (2025) use Llama-3.1 tokenizer to tokenize and construct their benchmark. This implies that Llama-3.1 and Qwen-2.5 receive slightly different numbers of input tokens for the same input datapoint.

ONERULER: This benchmark extends RULER in two significant ways. First, ONERULER’s NIAH task includes points without any correct needle present, requiring models to abstain appropriately. Second, ONERULER expands the task beyond English to encompass 26 languages representing diverse language families and writing systems. They find that accuracy on NIAH shows large drops on low-resource languages, especially ones that use non-Latin scripts. We evaluate models using two NIAH task types: NIAH-single and NIAH-none. In NIAH-single, a single correct needle is present

¹⁴<https://github.com/vllm-project/vllm/issues/11655>.

Model Family	Quantization	Models & Source Links	Key Details
Llama-3.1	AWQ-int4	8B , 70B (HuggingQuants)	Zero-point quantization, group size 128, GEMM kernels.
	GPTQ-int4	8B , 70B (RedHat AI)	Symmetric per-group quantization with linear scaling, 10% damping, 768 calibration sequences. Only weights quantized.
	GPTQ-int8	8B (RedHat AI)	Symmetric per-channel quantization with linear scaling, 1% damping, 256 sequences of 8192 tokens. Only weights quantized.
	GPTQ-int8	Locally created 70B quantization	W8A16 quantization adapted from RedHat AI’s 70B W8A8 recipe , integrating SmoothQuant into GPTQ (Kurtic et al., 2024).
	FP8	8B , 70B (RedHat AI)	Symmetric per-channel quantization on weights and activations with per-token dynamic scaling.
	BNB-nf4	In-flight quantization done in vLLM	Weights converted from BF16 to nf4 blocks at load time; activations remain in BF16.
Qwen-2.5	AWQ-int4	7B , 32B , 72B	Official Qwen-2.5 release (Qwen, 2024a).
	GPTQ-int4	7B , 32B , 72B	Official Qwen-2.5 release (Qwen, 2024a).
	GPTQ-int8	7B , 32B , 72B	Official Qwen-2.5 release (Qwen, 2024a).
	FP8	Locally created FP8 Qwen-2.5 models using the same recipe as Llama-3.1 family’s FP8 detailed above.	Dynamic quantization using RedHat AI’s Llama-3.1 70B FP8 recipe (Kurtic et al., 2024).
	BNB-nf4	In-flight quantization done in vLLM	Weights converted from BF16 to nf4 blocks at load time; activations remain in BF16.

Table 1: Summary of quantized model sources, quantization formats, and implementation details.

without distractors, which models must retrieve accurately. In NIAH-none, no correct needle is present, but four distractor needles exist, requiring models to abstain from retrieving any needle.

NOCHA: NOCHA evaluates long-context reasoning abilities by testing realistic claim verification and reasoning skills of LLMs on book-length texts. Since Llama-3.1 and Qwen-2.5 support context lengths of up to 128K tokens, this work considers the subset of datapoints with contexts up to that length, yielding 723 claim pairs drawn from 46 novels and three short story collections. While RULER and ONERULER help analyze the impact of quantization across varying context lengths, NOCHA enables a more nuanced investigation. First, it

poses a significantly more challenging task for LLMs, with full-precision models from Llama-3.1 and Qwen-2.5 performing at approximately chance level (Karpinska et al., 2024). Additionally, NOCHA facilitates further analysis through various dataset splits. A split based on *true* versus *false* claims enables the study of potential LLM biases. A difficulty-based split allows for examining the impact of question difficulty, and not just context length. Finally, a reasoning-scope-based split supports analysis across different spans of context required for reasoning.

FACTSCORE: *Evaluation procedure:* the VERIScore metric runs a three-stage evaluation pipeline: it extracts individual claims from the text

using a task-finetuned LLM, focusing specifically on generating sensible and verifiable factual claims. It then retrieves evidence for each claim through Google search results using the Serper API.¹⁵ Finally, each claim is verified against the retrieved evidence using another task-finetuned LLM. To account for the open-ended nature of the generation, it computes the F1 score at K, where K is the median number of claims extracted across all responses. *Dataset and generation:* see §E for a complete example of biography generation prompt and response for the prompt.

CS4: Each CS4 input prompt initially consists of a simple instruction and is subsequently expanded by introducing sets of constraints across three complexity levels: specifically, 9, 23, and 39 constraints. By incrementally increasing the number of constraints, we examine the model’s capability to write creatively while adhering to provided instructions. For evaluating constraint satisfaction, we improve upon the evaluation prompt used in Atmakuru et al. (2024), employing GPT4o¹⁶ as the judge (Figure 10). Instead of providing the generated story along with all constraints for evaluation, we feed GPT4o three constraints at a time. GPT4o is also used to evaluate story quality using BoookScore. BoookScore evaluates the narrative quality based on language fluency, salience, internal consistency, and related aspects. It is a pairwise evaluation metric that enables comparing stories generated by quantized models against those produced by their full-precision counterparts. See §D for a complete example including the input constraints, generated story, and constraint-satisfaction evaluation.

C Further results

This section provides and details more results, adding to §3. Figure 11 shows average results across models by the quantization method.

C.1 RULER

GPTQ-int4 and BNB-nf4 consistently underperform across models on RULER: Figure 4a shows the Δ -accuracy of the quantized models relative to BF16 on RULER. At the 8K context length, all methods perform similarly to BF16. As the context length increases, 8-bit methods remain robust, with

accuracy drops under 5%.¹⁷ However, 4-bit methods, especially BNB-nf4, show significant degradation in longer contexts. A post-hoc analysis in Table 6 (§F) reveals that the performance drop in BNB-nf4 compared to BF16 is statistically significant at the context lengths 64K and 128K ($p < 0.05$ for both). A model-wise breakdown is shown in Figure 6a. Except FP8 quantization on Qwen-2.5 32B, 8-bit quantization methods exhibit less than a 3% performance drop relative to BF16. On the other hand, GPTQ-int4 and BNB-nf4 show performance drops of at least 3% on all models and over 10% on several models: specifically, GPTQ-int4 on Llama-3.1 8B (21.2%) and Qwen-2.5 72B (10.3%), and BNB-nf4 on Llama-3.1 8B (30.7%), Qwen-2.5 7B(29.7%), and Qwen-2.5 32B(15.7%).

Quantization-induced degradation does not consistently align with task difficulty: Figure 12 analyzes the impact on different sub-tasks of RULER on quantization. Absolute performance values for these sub-tasks are visualized as heat maps: Figure 13 for needle retrieval, Figure 14 for multi-key retrieval, and Figure 15 for multi-value retrieval. Although RULER defines *needle* as the easiest of the three tasks, Figure 12 shows that the performance drop on *needle* is not the lowest for the FP8, GPTQ-int8, and BNB-nf4 quantization methods. This suggests that different quantization methods struggle with different tasks and that increased task difficulty does not directly correlate with greater performance degradation.

Larger models are more robust to quantization: Consistent with prior research (Lee et al., 2024), Figures 6a and 13 to 15 demonstrate that within each model family, larger models are generally less affected by quantization and that different models and quantization methods find different tasks difficult.

C.2 ONERULER

GPTQ-int4 and BNB-nf4 consistently underperform across models on ONERULER: Figure 4b presents the Δ -accuracy of the quantized models relative to BF16 on ONERULER. FP8 and GPTQ-int8 perform comparably to BF16 across all context lengths. In contrast, 4-bit quantization—particularly BNB-nf4—leads to more pronounced performance degradation as context length increases. The impact of each quantization method,

¹⁵<https://serper.dev/>

¹⁶refers to the gpt-4o-2024-08-06 endpoint

¹⁷The p values are larger than 0.05 in Table 6 §F for FP8 and GPTQ-int8 at 64K and 128K context-lengths compared to BF16.

You are an expert evaluator. You will be given a story followed by a set of three constraints. Your task is to carefully read both of them and tell how many constraints are being satisfied in the story. As the output, print "yes" or "no" for each constraint based on whether it is being satisfied or not, followed by an explanation of why it is satisfied/violated. If a constraint is satisfied, print the sentence(s) from the story in which it is satisfied. If a constraint is not being satisfied, explain how it is being violated. Be very strict in your evaluation. Mark a constraint as satisfied ("yes") only if it is completely satisfied in the story. For no satisfaction/partial satisfaction, mark a "no". Your output should always follow the format: <constraint number><yes/no><explanation>.

Figure 10: Prompt used to evaluate constraint-satisfaction of stories in the CS4benchmark.

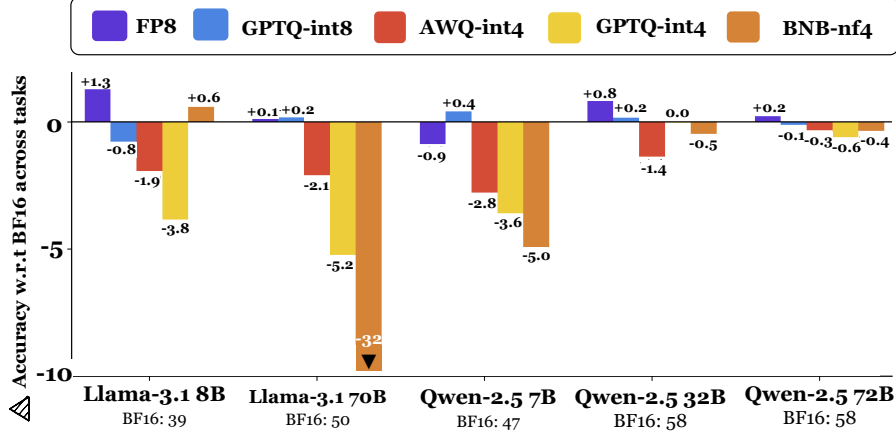


Figure 11: Average performance of each quantization method on each model averaged across all tasks. Performance is illustrated using Δ -accuracy, the difference w.r.t. the BF16 baseline. The BF16 baseline average results for each task (over which deltas are computed) are displayed above the x-axis. For ONERULER and RULER we only aggregate the long context results from 64K and 128K context lengths.

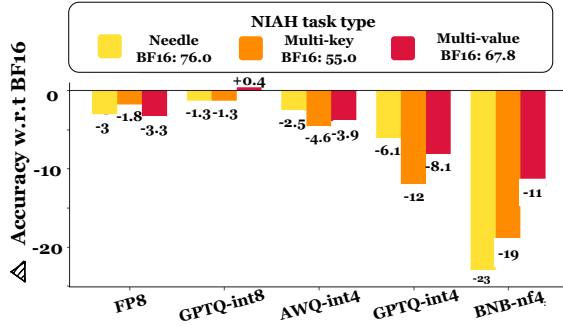


Figure 12: Average Δ -accuracy of different quantization methods relative to BF16 on the RULER benchmark across different sub-tasks.

however, varies by model. For instance, BNB-nf4 yields the best results for Llama-3.1 8B but performs particularly poorly on Llama-3.1 70B, with a 66% drop in accuracy. Interestingly, Qwen-2.5 32B and 72B maintain performance comparable to BF16 even under 4-bit quantization, showing strong robustness to quantization. Figure 5a shows no significant performance difference between high- and low-resource languages within each quantization method. While performance drops increase for non-English languages, we do

not observe consistent trends distinguishing high- vs. low-resource languages (Figure 5a). Finally, Figure 5b shows the performance of different quantization methods across language scripts, while Figure 16 further breaks down the ONERULER results by individual languages across models. Figure 17 shows how Llama-3.1 70B and Qwen-2.572B, despite being approximately of the same size, perform very differently across context lengths on ONERULER.

C.3 NOCHA

BF16 performs around chance on NOCHA: Figure 4c shows the Δ -accuracy of quantized models relative to BF16 on NOCHA. The magnitude of the Δ -accuracy is smaller than that observed on RULER and ONERULER. This is largely because even the unquantized models perform near chance level (Figure 19), leaving limited room for further degradation from quantization.

Performance drops become steeper beyond 75k context length: Figure 4c shows that most quantization methods undergo a steeper performance drop compared to BF16 when context length increases from below 75k to above 75k. Fig-

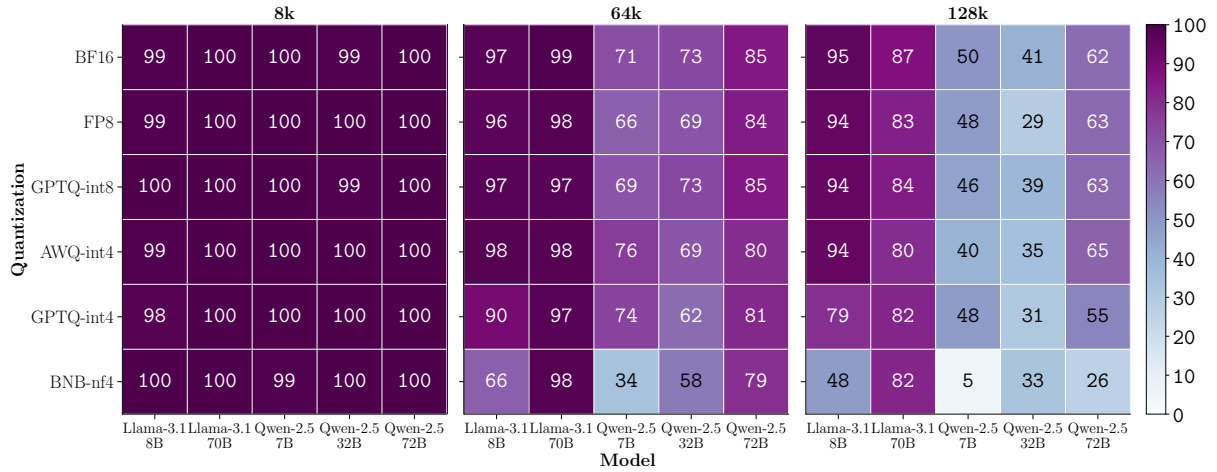


Figure 13: Performance of all models and quantization methods across context lengths on the Needle task in RULER.

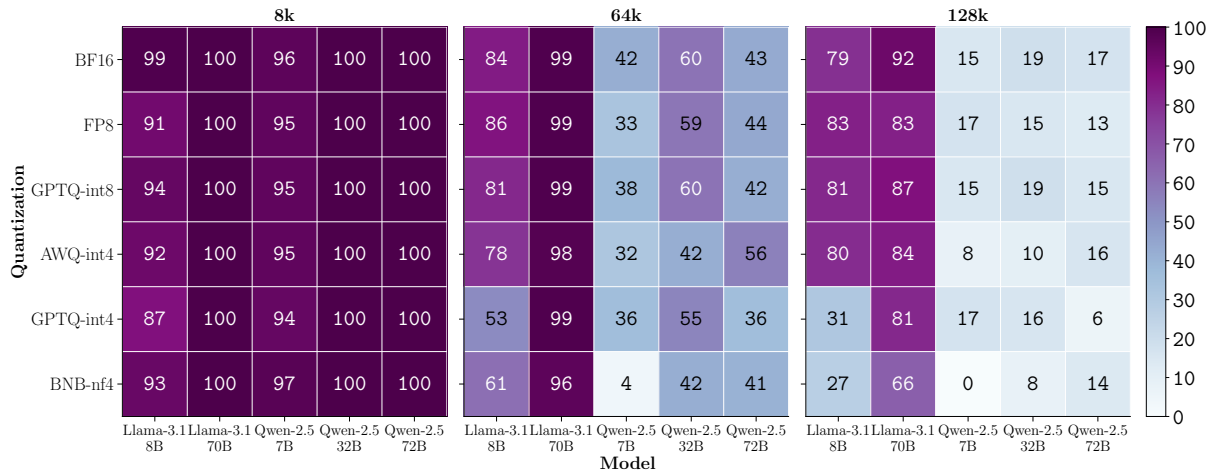


Figure 14: Performance of all models and quantization methods across context lengths on the Multi-key task in RULER.

ures 20 and 21 take a closer look at Figure 4c by analyzing the performance of each quantization method across different models at varying context lengths. At context lengths below 75k (Figure 20), the effects of quantization vary arbitrarily across models. While GPTQ-int4 and BNB-nf4 generally underperform compared to BF16, Figure 20 shows that all quantization methods—except AWQ-int4—outperform BF16 on Llama-3.1 8B. These trends stabilize as the context length increases beyond 75k (Figure 21). Although some quantization methods achieve up to 3% better performance than their BF16 counterparts, most configurations show relative drops. We observe performance drops of 13% for GPTQ-int8 on Llama-3.1 70B, 12% for AWQ-int4 on Qwen-2.5 72B, and 14% for GPTQ-int4 on Qwen-2.5 72B.

Quantization-induced degradation does not consistently align with task difficulty: Figure 22

shows the average Δ -accuracy of different quantization methods relative to BF16 across the *easy* and *hard* splits of NOCHA. Together with Figures 23 and 24, Figure 22 reiterates the observation from §3 that the magnitude of performance degradation due to quantization does not directly correlate with task difficulty. While Figure 22 may suggest that increased task difficulty leads to reduced performance degradation, Figures 23 and 24 show that this trend is arbitrary across models. Specifically for NOCHA, the BF16 baseline values in Figures 22 to 24 are close to chance level, making any small gain or drop around these values difficult to interpret meaningfully.

Quantization generally preserves model bias, but effects vary across methods and architectures: Each datapoint in NOCHA consists of narrative-minimal claim pairs labeled as *true* and *false*, with points awarded only when a model cor-

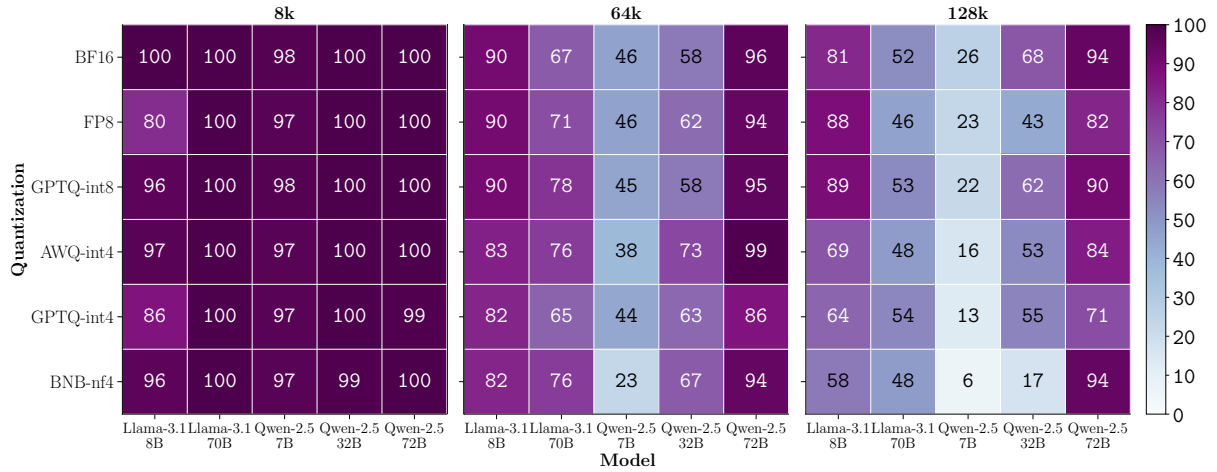


Figure 15: Performance of all models and quantization methods across context lengths on the Multi-value task in RULER.

rectly labels both claims within a pair. Analyzing the results for each label separately allows us to assess whether models exhibit bias toward one label. An unbiased model should perform similarly on both *true* and *false* claims; thus, the performance difference between them reflects the model’s bias. We observed that all models predicted *false* more frequently than *true*. On average, across models, Figure 18 shows that different quantization methods tend to preserve the bias of their BF16 counterparts, except for AWQ-int4 that amplifies the bias.

C.4 FACTSCORE

Figure 8 shows the trends in VERIScore as it varies across quantization methods with the frequency of the input entity. Figure 7a shows how VERIScore varies across quantization methods with the different models, while Figure 25 shows the rate of responses without verifiable claims on FACTScore.

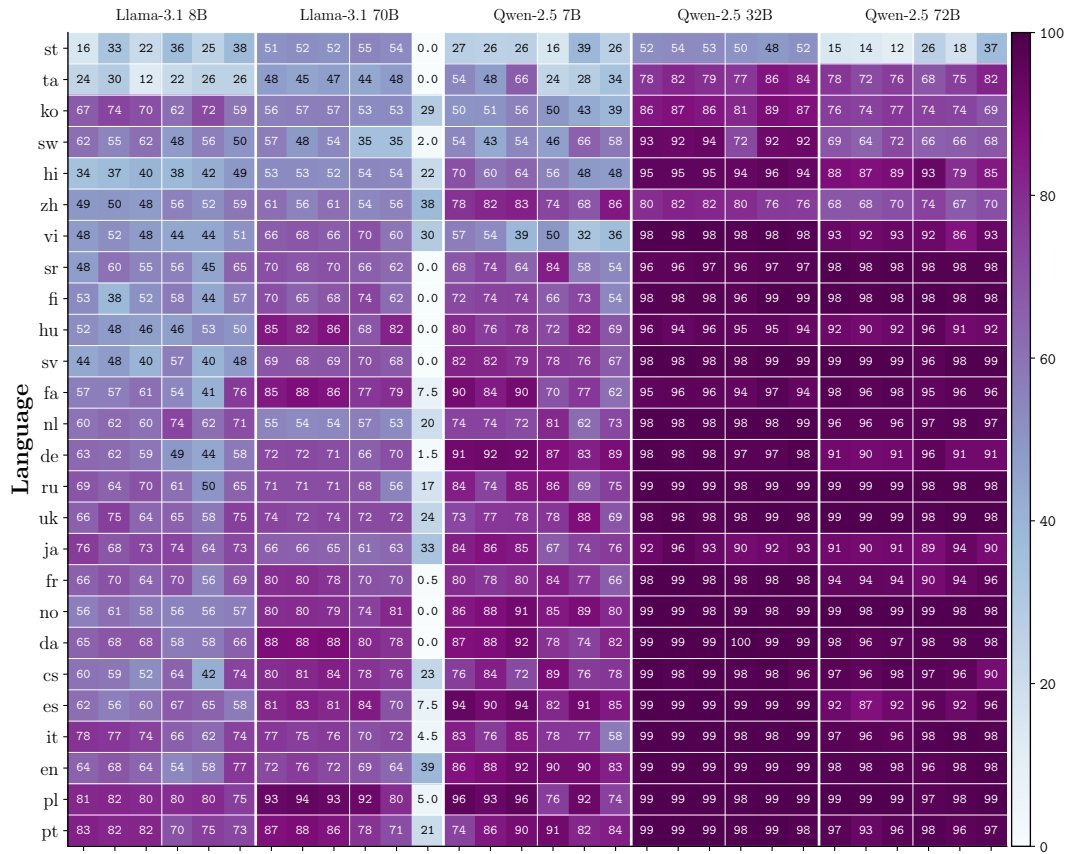
C.5 CS4

Prompting the judge model with fewer constraints improves agreement with human evaluations: Figure 10 (§B) shows the prompt used to evaluate constraint satisfaction for stories generated using the CS4 benchmark. Although each story-writing input can have up to 39 constraints, the judge model is prompted with only three constraints at a time, and the results are then aggregated. To assess whether this improves reliability, we conducted a meta-evaluation on 42 generations—21 where the judge model was prompted with 3 constraints at a time, and 21 where all

No. of input constraints for writing the story	No. of constraints given to judge-LLM at a time	Avg. evaluations agreeing with the human
7	3	6.57
7	7	6.57
23	3	21.00
23	23	19.46
39	3	37.28
39	39	32.00

Table 2: Agreement between human evaluation and LLM-as-judge under different judge-LLM input sizes. Agreement refers to the number of constraint evaluations (out of the total) where the judge model’s assessment matched human annotations, across varying numbers of input constraints and judge-prompt sizes.

constraints were shown at once. The 21 stories were evenly split across the three levels of input constraints. These 42 story–constraint pairs were manually annotated by one of the authors, who independently read each story and assessed which constraints were satisfied. Agreement scores reported in Table 2 indicate how many of the judge model’s evaluations matched the human annotations. Prompting with three constraints at a time resulted in closer agreement with human assessments.



Model and quantization

Figure 16: **ONERULER**: Performance (accuracy) of all model and quantization configurations across the points in each language. Includes only 64K, 128K lengths due to focus on very long contexts.

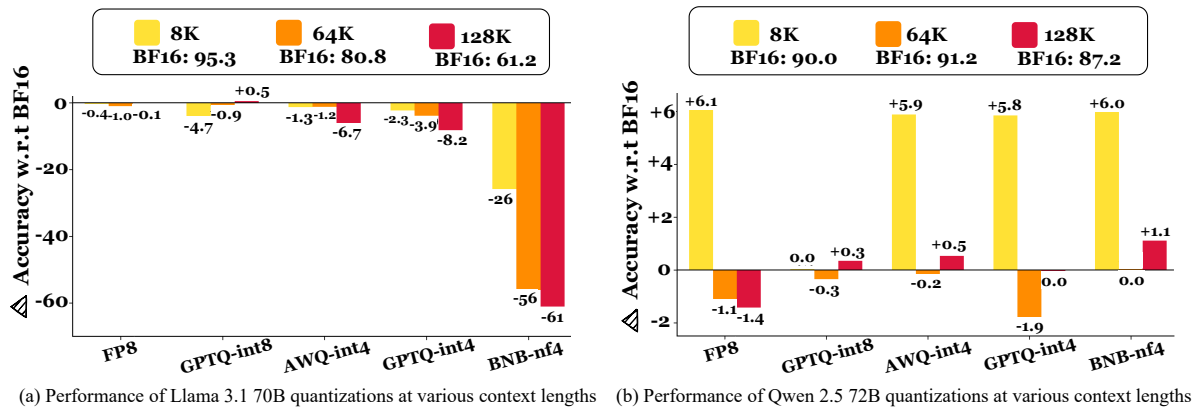


Figure 17: ONERULER Δ -accuracy, averaged across models, with varying context length for Llama-3.1 70B and Qwen-2.5 72B. Both show starkly different responses to quantization.

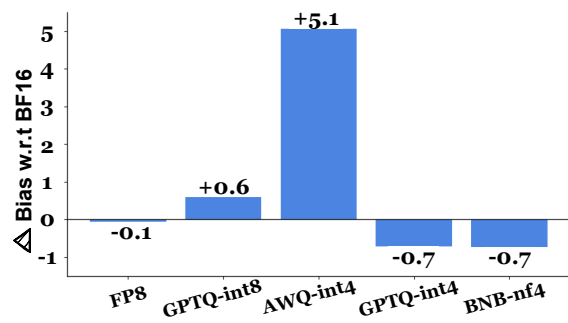


Figure 18: Average Δ -bias of different quantization methods relative to BF16 on the NOCHA benchmark. Bias is defined as the performance difference between *false* and *true* claims, indicating a model’s tendency to favor one label over the other.

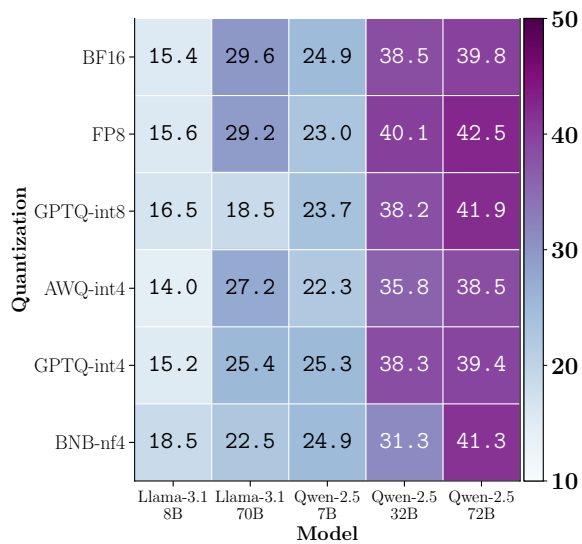


Figure 19: Performance (accuracy) of all models and quantization methods on the NOCHA benchmark.

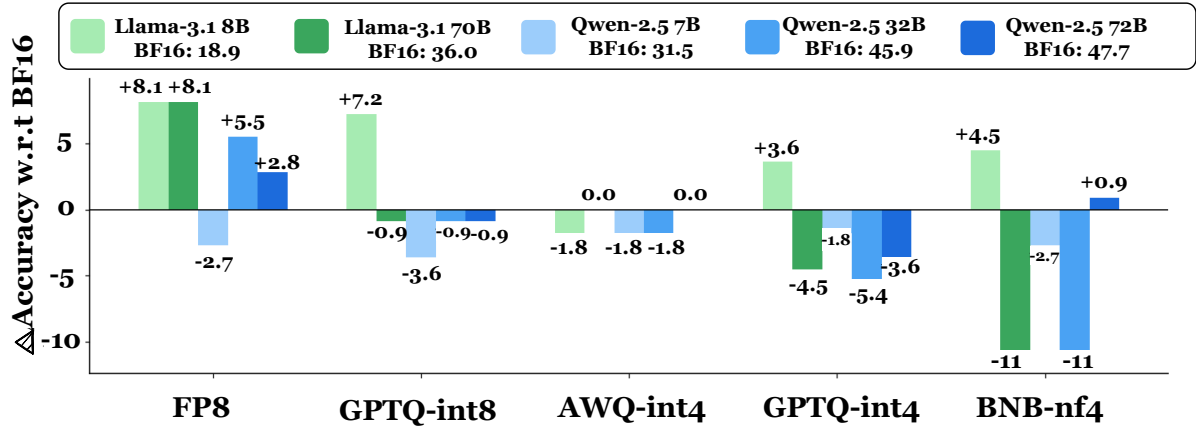


Figure 20: Average Δ -accuracy of different quantization methods relative to BF16 on the NOCHA benchmark across models when context length $< 75k$.

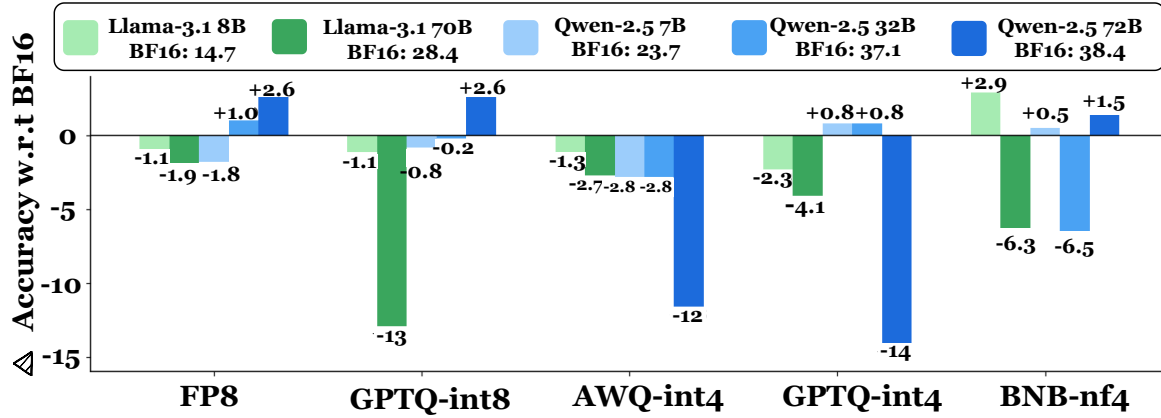


Figure 21: Average Δ -accuracy of different quantization methods relative to BF16 on the NOCHA benchmark across models when context length $> 75k$.

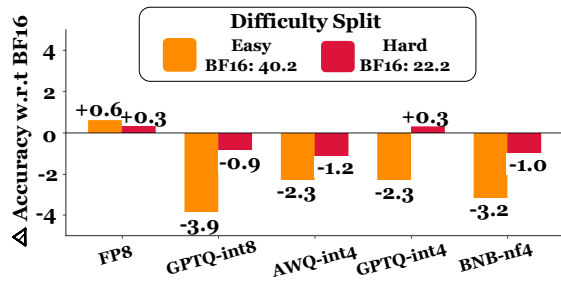


Figure 22: Average Δ -accuracy of different quantization methods relative to BF16 on the NOCHA benchmark with varying difficulty level.

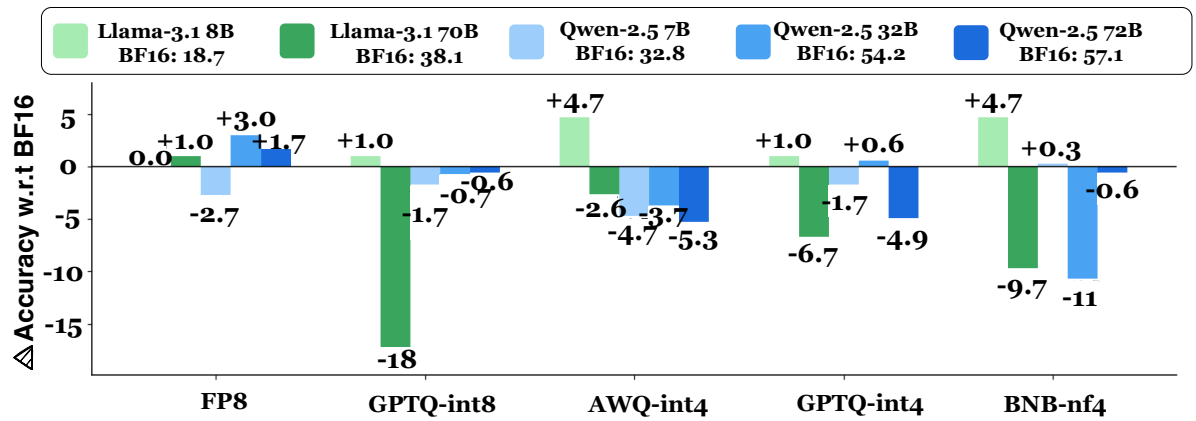


Figure 23: Average Δ -accuracy of different quantization methods relative to BF16 on the *easy* split of NoCHA across models.

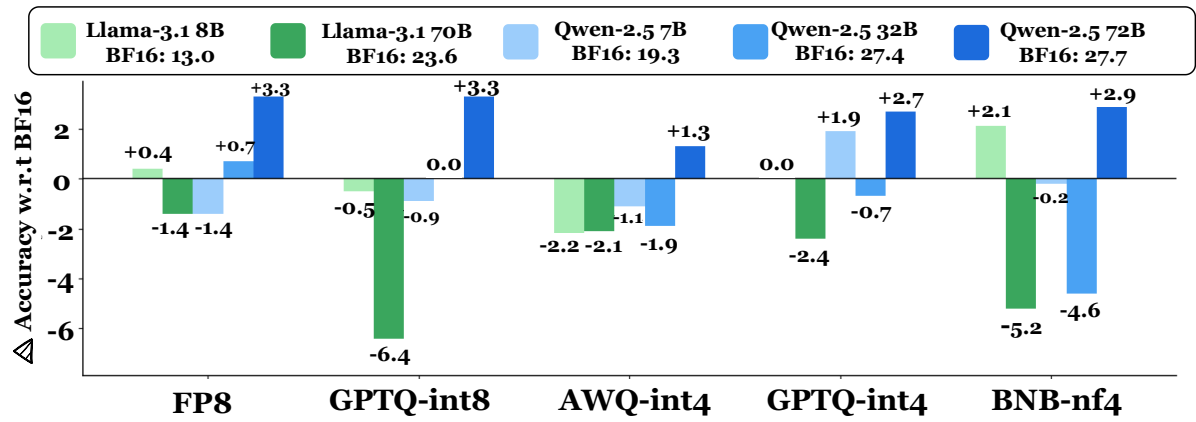


Figure 24: Average Δ -accuracy of different quantization methods relative to BF16 on the *hard* split of NoCHA across models.

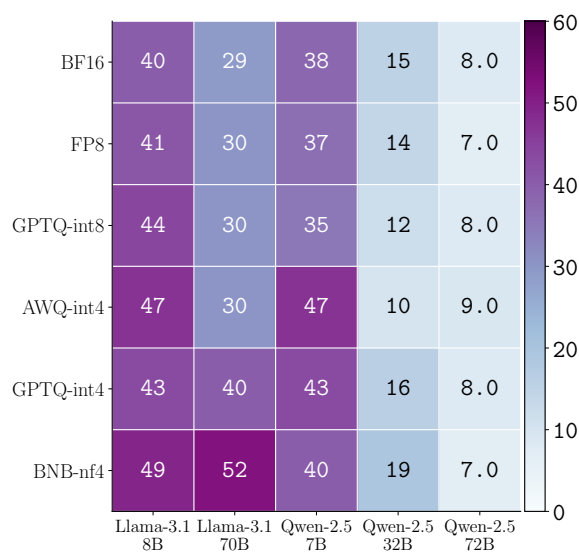


Figure 25: Abstain-rate (rate of responses without verifiable claims) on FACTSCORE.

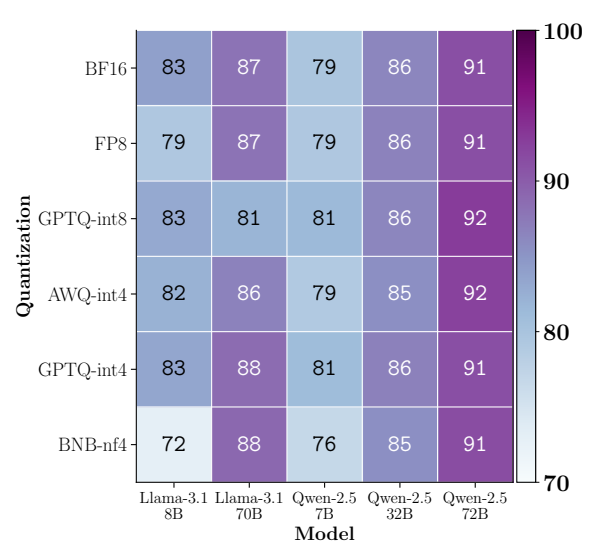


Figure 26: Average constraint-satisfaction scores of different model-quantization pairs on CS4.

D An example from the CS4 benchmark

Input prompt for story writing: Write a story about an introverted protagonist with uncontrollable intuition, who experiences others' emotions intensely, as she navigates a challenging encounter during the holiday season. The story should satisfy the following constraints:

1. The protagonist suffers physical discomfort when overwhelmed by emotions (nausea, shaking, etc.).
2. The protagonist is challenged by the need to engage in public spaces.
3. The unknown man realizes that the protagonist can feel his emotions.
4. The protagonist uses humor and sarcasm to cope with her situation.
5. The protagonist is an introverted character.
6. The story includes communication via text messages.
7. The story is set in a Starbucks on Michigan in Chicago a week before Christmas.
8. The protagonist is forced to leave the meeting early due to being overwhelmed.
9. The protagonist desires to live a more normal life despite her unique condition.
10. There could exist a vaccine for controlling intuition.
11. The protagonist devises coping strategies for managing her anxiety in public places.
12. Scientists are working to find a solution for people who can't use the intuition vaccine.
13. Tiffany threatens the protagonist to meet her.
14. The protagonist struggles with accepting her condition.
15. The protagonist must grapple with the thoughts and feelings of others in the Starbucks.
16. The man looks at the protagonist with both desire and love.
17. The protagonist is physically attractive.
18. The protagonist encounters an unknown man who causes powerful and unique emotions.
19. There is societal disapproval for people whose intuitions cannot be controlled by the vaccine.
20. The protagonist has a heightened intuition.
21. The setting should be during the holiday season.
22. The protagonist feels other people's emotions intensely.
23. The protagonist struggles with disentangling their own feelings from others'.

Story generated by Llama-3.1 8B BF16 model: In the cozy glow of a late-December morning, under the twinkling lights of a small Starbucks on Michigan Avenue in Chicago, Emily sat alone, sipping a steaming latte. The holiday season was in full swing, and the café was bustling with people. Emily, an introvert with a heightened intuition, felt her heart racing as she scanned the crowd. Her eyes met those of an unknown man across the room. He was tall, with piercing blue eyes and a charming smile. Emily felt a mix of emotions—curiosity, warmth, and a hint of fear. As he looked back, she could sense a deep longing and a touch of desire in him. The intensity of these feelings was overwhelming, causing her stomach to churn and her hands to shake slightly. "Are you okay?" a voice whispered through her mind, breaking the spell. She blinked, trying to focus. It was the man from across the room. He leaned closer, his eyes filled with concern. "You look like you're going to be sick." Emily laughed, a nervous chuckle that sounded hollow even to her ears. "I'm fine," she said out loud, her voice trembling. "Just a bit... overwhelmed." He nodded, understanding. "I get it. I've been there." He gestured towards the crowded café. "Want to step outside for a bit? It's not easy in here." Emily hesitated but then nodded, grateful for his kindness. They stepped into the chilly air, and the man offered her a jacket. "Here, put this on. It's freezing out here." As they walked, the man began to share stories, his laughter light and infectious. Emily found herself smiling, her anxiety easing slightly. She realized that she could control her reactions better when she engaged in conversation. "So, what brings you to Chicago this time of year?" she asked, trying to keep the conversation flowing. "I'm actually here for a conference," he replied, his eyes never leaving hers. "And you?" "Just visiting," Emily said, her cheeks flushing. She felt a wave of emotions wash over her—joy, admiration, and something she couldn't quite place. She took a deep breath, her intuition telling her to be careful. Back inside the café, the man excused himself, promising to catch up later. Emily sat down, her mind racing. She had always struggled with her unique condition, feeling the emotions of others intensely and often struggling to separate them from her own. The idea of a vaccine that could control her intuition was appealing, but she feared losing the connection to others. Tiffany, her best friend, had threatened to meet her earlier that day, but Emily had pushed her away. She needed some space to think. Emily pulled out her phone and texted Tiffany, "I'm meeting someone for coffee. Can we talk later?" Tiffany replied almost immediately, "Sure thing. Be safe." Emily sighed, feeling a mix of relief and apprehension. She knew she had to find a way to live a more normal life, one where she could navigate public spaces without being overwhelmed. She decided to reach out to a support group for people like her, hoping to find others who understood her struggle. As the sun began to set, casting a warm golden light over the city, Emily gathered her things and left the café. She knew the road ahead would be challenging, but she also knew that she wasn't alone. With each step, she felt a glimmer of hope, determined to find a way to live a more balanced life, one where her unique condition didn't define her.

Evaluation for constraint satisfaction using the prompt in Figure 10:

1. Yes – "The intensity of these feelings was overwhelming, causing her stomach to churn and her hands to shake slightly." –the story shows clear nausea and shaking when emotions peak, fully meeting the requirement for physical discomfort.
2. Yes – "The holiday season was in full swing, and the café was bustling with people." / "She knew she had to find a way to live a more normal life, one where she could navigate public spaces

- without being overwhelmed.” These lines depict a crowded public setting that stresses her and a stated struggle with such spaces.
3. No – At no point does the unknown man discover that Emily can literally feel his emotions; he only notices that she looks unwell. There is no dialogue or narration indicating his realization of her empathic ability, so the constraint is unmet.
 4. No – Although Emily gives a nervous chuckle and says, “I’m fine. . . just a bit overwhelmed,” this is neither genuine humor nor sarcasm used as a coping mechanism. The story lacks any witty, humorous, or sarcastic remarks from her to manage the situation, so the criterion is not satisfied.
 5. Yes – “Emily, an introvert with a heightened intuition, felt her heart racing as she scanned the crowd.” The narrative explicitly identifies the protagonist as an introvert, fully fulfilling the constraint.
 6. Yes – “Emily pulled out her phone and texted Tiffany, ‘I’m meeting someone for coffee. Can we talk later?’ Tiffany replied almost immediately, ‘Sure thing. Be safe.’” – these two lines show a back-and-forth text exchange, satisfying the communication-via-text requirement.
 7. No – The narrative states only that events occur “in the cozy glow of a late-December morning, under the twinkling lights of a small Starbucks on Michigan Avenue in Chicago.” It never specifies “a week before Christmas.” Late December could be before or after Christmas, so the exact timing demanded by the constraint is missing.
 8. No – Although Emily briefly steps outside to get fresh air, she later returns and ultimately leaves the café at sunset of her own accord. The story never shows her being compelled to end the encounter prematurely because of overwhelm (e.g., cutting the meeting short, fleeing abruptly), so the requirement of being “forced to leave the meeting early” is not met.
 9. Yes – “She knew she had to find a way to live a more normal life, one where she could navigate public spaces without being overwhelmed.” – this sentence directly expresses her desire for a normal life despite her extraordinary intuition.
 10. Yes – “The idea of a vaccine that could control her intuition was appealing, but she feared losing the connection to others.” – this line explicitly introduces the existence of a vaccine designed to dampen or manage her intuitive ability, fulfilling the constraint.
 11. Yes – She actively develops ways to stay calm: “She realized that she could control her reactions better when she engaged in conversation,” and later, “She decided to reach out to a support group for people like her.” These lines show deliberate coping strategies for public-place anxiety.
 12. No – While a vaccine to blunt intuition is mentioned, the story never says that scientists are pursuing an alternative for people who cannot take it; no research effort is described.
 13. Yes – “Tiffany, her best friend, had threatened to meet her earlier that day, but Emily had pushed her away.” –the verb “threatened” makes the demand explicit.
 14. Yes – “She had always struggled with her unique condition, feeling the emotions of others intensely and often struggling to separate them from her own.” This directly shows her difficulty accepting and living with the condition.
 15. Yes – “Emily, an introvert with a heightened intuition, felt her heart racing as she scanned the crowd.” / “She could sense a deep longing and a touch of desire in him.” These sentences depict her grappling with others’ thoughts and feelings inside the Starbucks.
 16. No – The narrative states the man’s “deep longing and a touch of desire,” but it never attributes love to him. Because love is absent, the condition is unmet.
 17. No – Nowhere is Emily’s physical attractiveness described or implied; the story focuses on her inner state, not her appearance.
 18. Yes – “Her eyes met those of an unknown man across the room. . . She felt a mix of emotions—curiosity, warmth, and a hint of fear.” The unknown man clearly triggers powerful, unique feelings in her, satisfying the constraint.
 19. No – The story gives no indication of societal disapproval toward people whose intuition cannot be controlled; no stigma or judgment is depicted.
 20. Yes – “Emily, an introvert with a heightened intuition. . .” –the phrase explicitly confirms her heightened intuitive ability.
 21. Yes – “The holiday season was in full swing” and the date is “late-December”; the setting is clearly during the holiday season.
 22. Yes – “She had always struggled with her unique condition, feeling the emotions of others intensely. . .” –this line directly states that she experiences other people’s emotions intensely.
 23. Yes – “. . . often struggling to separate them from her own.” The narrative explicitly describes her difficulty disentangling her feelings from others’, fully satisfying the constraint.

E An example from the FACTSCORE benchmark

Input prompt: Write the biography of Fumiko Kaneko.

Response: Fumiko Kaneko (1903–1926) was a Japanese anarchist and nihilist who gained notoriety for her involvement in a plot to assassinate the Japanese Emperor Hirohito and other high-ranking government officials.

Born in 1903 in Aomori Prefecture, Japan, Kaneko grew up in a poor farming family. Her early life was marked by hardship and struggle, which may have contributed to her radicalization. In her late teens, she moved to Tokyo, where she became involved in the city's anarchist and nihilist circles.

In 1923, Kaneko met Pak Yeol, a Korean anarchist who had fled to Japan after being involved in the March 1st Movement, a Korean independence movement against Japanese colonial rule. The two became romantically involved and began to plan a series of bombings and assassinations targeting Japanese government officials and the royal family.

On January 9, 1926, Kaneko and Pak Yeol were arrested by Japanese police while attempting to plant a bomb near the Imperial Palace in Tokyo. The plot was foiled, and the two were charged with treason and attempted murder.

During her trial, Kaneko showed no remorse for her actions, stating that she had acted out of a desire to overthrow the Japanese government and monarchy, which she saw as oppressive and corrupt. Her statements were seen as a challenge to the authority of the Japanese state, and she became a symbol of resistance against the government.

Kaneko was sentenced to death and executed by hanging on March 23, 1926, at the age of 23. Pak Yeol was also executed on the same day. Their executions sparked widespread outrage and protests among anarchist and left-wing groups in Japan, who saw them as martyrs for the cause of revolution and social change.

Despite her short life and tragic end, Fumiko Kaneko's legacy as a radical activist and symbol of resistance has endured in Japan and beyond. Her story has inspired numerous books, films, and plays, and she remains a fascinating and complex figure in Japanese history.

F Statistical tests of significance for comparing quantization methods

We conducted statistical analysis by fitting the generalized linear mixed-effects models (GLMMs) using the `glmer()` function from the `lme4` package in R (Bates et al., 2015) to study if different quantization methods performed significantly differently from their BF16 counterparts. For improved convergence, we used a *binomial logit link* and the *bobyqa* optimizer. We formatted our data in long format, with each model response represented as a binary categorical variable indicating whether the prediction was correct or incorrect. Quantization methods were included as the primary fixed effect of interest, with additional fixed effects for context length, task, and their interactions in different model variants.

For each GLMM we report marginal R^2 , the variance explained by fixed effects alone, and conditional R^2 , the variance explained by fixed and random effects (Nakagawa et al., 2017). We also list the variance components (σ^2 , τ_{00}) and intra-class correlation coefficients (ICC) for every grouping factor. Other aspects like task, context length, and language are added where appropriate to account for the non-independence of repeated measures and to allow the fixed effects of interest to be estimated after partialing out systematic differences due solely to model identity, task, or input length.

Following (Karpinska et al., 2024), we conducted pairwise comparisons of quantization levels with `emmeans` for every GLMM. p-values were Bonferroni-adjusted (Westfall et al., 1997) and reported in the post-hoc tables. Log-odds estimates were exponentiated to odds ratios, which were then converted to probabilities to facilitate interpretation (a probability of 0.50 means the two methods are equally accurate).

To examine how quantization interacts with context length and task type within RULER, we fit three GLMMs with `Model` as a random intercept. The main-effect model (`Binary_label ~ Quantization + (1|Model)`) tests whether any quantization method differs from BF16 (Table 3; pairwise contrasts in Table 4). The interaction model with context length (`Binary_label ~ Quantization * Context_length + (1|Model)`) + $(1|\text{Task})$) evaluates whether quantization effects vary across 8k, 64k, and 128k tokens while controlling for task variability (Table 5; contrasts in Ta-

ble 6). The task interaction model (`Binary_label ~ Quantization * Task + (1|Model)`) compares performance of quantized models on multi-key and multi-value tasks on RULER, using BF16 on needle as the reference (Table 7; contrasts in Table 8). Across these models, marginal R^2 increases from .013 (quantization only) to .615 (quantization * context-length), while conditional R^2 peaks at .488, indicating that both fixed effects and grouping factors jointly capture a sizeable proportion of variance on RULER. We conduct the same statistical analyses with ONERULER as in RULER (Tables 9 to 14), with the only change being the use of *language frequency* instead of *task*.

NOCHA analyses parallel the RULER tests but center on the 75k-token split defined in the benchmark. The main-effect model (`Binary_label ~ Quantization + (1|Context_length) + (1|Model)`) includes quantization as the predictor, with `Context_length` and `Model` as random intercepts (Table 15); pairwise contrasts appear in Table 16. To test whether quantization errors increase for longer inputs, we also fit an interaction model (`Binary_label ~ Quantization * Context_length + (1|Model)`), using BF16 at $\leq 75k$ tokens as the reference (Table 17; contrast results are reported in Table 18). Random-effects variances are small for NOCHA ($ICC_{\text{Model}} \approx 0.045$), and both marginal and conditional R^2 values are low (≈ 0.016), indicating the dataset’s overall difficulty.

To evaluate how quantization impacts performance on FACTSCORE, we fit three linear and generalized linear mixed-effects models with `Model` as a random intercept. The main-effect model (`log_veriscore ~ Quantization + (1|topic) + (1|Model)`) estimates whether quantized models deviate from the BF16 baseline in overall generation quality (Table 19; pairwise contrasts in Table 20). To assess no claim rate, we model the likelihood of no claims (`glmer(no_claims ~ Quantization + (1|Model) + (1|topic))`) in Table 21 with contrasts in Table 22. Lastly, we examine whether quantization interacts with category (`log_veriscore ~ Quantization * cat + (1|Model) + (1|topic)`) by modeling VERIScore across medium, very frequent, and very rare categories (Table 23; post-hoc contrasts in Table 24).

Figure 26 in §C.5 plots the absolute constraint satisfaction scores of each model-quantization pair. Here, we fit GLMMs to evaluate how quantization

influences constraint satisfaction on CS4. However, each data point here reflects the percentage of constraints satisfied, i.e., a success proportion, rather than a binary label. Since this can only take a finite set of values, we use a binomial GLMM with `cbind`. The main-effect model (`cbind(Success, Failures) ~ Quantization + (1|Model)`) tests whether quantized models differ from BF16 in satisfying constraints (Table 25; pairwise contrasts in Table 26). To analyze whether quantization interacts with the number of input constraints, we introduce input-level interactions in (`cbind(Success, Failures) ~ Quantization * Input_constraints + (1|Model)`), allowing us to track performance across 7, 23, and 39 input constraints (Table 27; post-hoc results in Table 28).

glmer(Binary_label ~ Quantization + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	1.336	3.802	2.182	6.623	0.000	***
FP8	-0.144	0.866	0.783	0.957	0.050	
GPTQ-int8	-0.042	0.959	0.867	1.062	0.421	
AWQ-int4	-0.158	0.854	0.772	0.944	0.002	**
GPTQ-int4	-0.365	0.694	0.629	0.766	0.000	***
BNB-nf4	-0.647	0.524	0.475	0.577	0.000	***
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	0.408					
ICC (Model)	0.290					
N (Model)	5					
Observations	27000					
R^2 (marginal)	0.013					
R^2 (conditional)	0.122					

Table 3: **RULER**: Summary of generalized linear mixed model with **quantization** as the predictor of **accuracy**: glmer(Binary_label ~ Quantization + (1|Model), data = data, family = binomial). The quantization method "BF16" was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. See Table 4 for post-hoc analysis.

CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
BF16 - FP8	0.144	0.051	1.155	0.536	0.072	
BF16 - (GPTQ-int8)	0.042	0.052	1.043	0.510	1.000	
BF16 - (AWQ-int4)	0.158	0.051	1.171	0.539	0.029	*
BF16 - (GPTQ-int4)	0.365	0.050	1.441	0.590	0.000	***
BF16 - (BNB-nf4)	0.647	0.049	1.910	0.656	0.000	***
(AWQ-int4) - FP8	-0.014	0.050	0.986	0.497	1.000	
(AWQ-int4) - (GPTQ-int8)	-0.117	0.051	0.890	0.471	0.330	
(AWQ-int4) - (GPTQ-int4)	0.207	0.049	1.230	0.552	0.000	***
(AWQ-int4) - (BNB-nf4)	0.489	0.048	1.631	0.620	0.000	***
FP8 - (GPTQ-int4)	0.221	0.049	1.247	0.555	0.000	***
FP8 - (GPTQ-int8)	-0.103	0.051	0.902	0.474	0.661	
FP8 - (BNB-nf4)	-0.503	0.048	0.605	0.377	0.000	***
(GPTQ-int4) - (GPTQ-int8)	-0.324	0.050	0.723	0.420	0.000	***
(GPTQ-int4) - (BNB-nf4)	-0.282	0.047	0.754	0.430	0.000	***
(GPTQ-int8) - (BNB-nf4)	-0.606	0.049	0.546	0.353	0.000	***

Table 4: **RULER**: Post-hoc comparisons between quantization methods for **accuracy** (Table 3) using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).

glmer(Binary_label ~ Quantization * Context_length + (1 Model) + (1 Task), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	5.577	264.228	85.367	817.846	0.000	***
AWQ-int4	-0.814	0.443	0.216	0.910	0.027	*
BNB-nf4	-0.761	0.467	0.227	0.961	0.039	*
FP8	-1.480	0.228	0.118	0.440	0.000	***
GPTQ-int4	-1.508	0.221	0.115	0.428	0.000	***
GPTQ-int8	-0.705	0.494	0.238	1.026	0.059	
64k	-4.319	0.013	0.007	0.024	0.000	***
128k	-5.164	0.006	0.003	0.010	0.000	***
AWQ-int4:64k	0.757	2.131	1.019	4.460	0.045	*
BNB-nf4:64k	0.061	1.063	0.508	2.226	0.871	
FP8:64k	1.428	4.169	2.113	8.222	0.000	***
GPTQ-int4:64k	1.170	3.222	1.635	6.350	0.001	***
GPTQ-int8:64k	0.693	2.000	0.945	4.234	0.070	
AWQ-int4:128k	0.496	1.643	0.788	3.424	0.185	
BNB-nf4:128k	-0.388	0.679	0.325	1.418	0.303	
FP8:128k	1.255	3.507	1.785	6.893	0.000	***
GPTQ-int4:128k	0.932	2.541	1.293	4.992	0.007	**
GPTQ-int8:128k	0.642	1.900	0.901	4.007	0.092	
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	0.807					
τ_{00} (Task)	0.281					
ICC (Model)	0.386					
ICC (Task)	0.134					
N (Model)	5					
N (Task)	3					
Observations	27000					
R^2 (marginal)	0.615					
R^2 (conditional)	0.488					

Table 5: **RULER**: Summary of generalized linear mixed model with **quantization**, **context length**, and their interaction as predictors of **accuracy**: glmer(Binary_label ~ Quantization * Context_length + (1|Model) + (1|Task), data = data, family = binomial). The quantization method "BF16" at context length 8k was set as the reference level (intercept), with Model and Task set as the grouping factors for random intercepts to account for variability across models and tasks. See Table 6 for post-hoc analysis.

CONTEXT	CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value
8k	BF16 - (AWQ-int4)	0.814	0.367	2.256	0.693	0.398
8k	BF16 - (BNB-nf4)	0.761	0.368	2.140	0.682	0.580
8k	BF16 - FP8	1.480	0.336	4.395	0.815	0.000 ***
8k	BF16 - (GPTQ-int4)	1.508	0.336	4.516	0.819	0.000 ***
8k	BF16 - (GPTQ-int8)	0.705	0.373	2.025	0.669	0.878
8k	(AWQ-int4) - (BNB-nf4)	-0.053	0.320	0.949	0.487	1.000
8k	(AWQ-int4) - FP8	0.667	0.279	1.948	0.661	0.255
8k	(AWQ-int4) - (GPTQ-int4)	0.694	0.277	2.002	0.667	0.183
8k	(AWQ-int4) - (GPTQ-int8)	-0.108	0.326	0.898	0.473	1.000
8k	(BNB-nf4) - FP8	0.719	0.283	2.053	0.672	0.164
8k	(BNB-nf4) - (GPTQ-int4)	0.747	0.281	2.110	0.678	0.119
8k	(BNB-nf4) - (GPTQ-int8)	-0.055	0.328	0.946	0.486	1.000
8k	FP8 - (GPTQ-int4)	0.027	0.233	1.028	0.507	1.000
8k	FP8 - (GPTQ-int8)	-0.775	0.288	0.461	0.315	0.108
8k	(GPTQ-int4) - (GPTQ-int8)	-0.802	0.287	0.448	0.310	0.078
64k	BF16 - (AWQ-int4)	0.057	0.090	1.058	0.514	1.000
64k	BF16 - (BNB-nf4)	0.700	0.087	2.013	0.668	0.000 ***
64k	BF16 - FP8	0.053	0.090	1.054	0.513	1.000
64k	BF16 - (GPTQ-int4)	0.338	0.088	1.402	0.584	0.002 **
64k	BF16 - (GPTQ-int8)	0.012	0.090	1.012	0.503	1.000
64k	(AWQ-int4) - (BNB-nf4)	0.643	0.086	1.902	0.655	0.000 ***
64k	(AWQ-int4) - FP8	-0.004	0.090	0.996	0.499	1.000
64k	(AWQ-int4) - (GPTQ-int4)	0.281	0.088	1.324	0.570	0.021 *
64k	(AWQ-int4) - (GPTQ-int8)	-0.045	0.090	0.956	0.489	1.000
64k	(BNB-nf4) - FP8	-0.647	0.086	0.524	0.344	0.000 ***
64k	(BNB-nf4) - (GPTQ-int4)	-0.362	0.084	0.696	0.411	0.000 ***
64k	(BNB-nf4) - (GPTQ-int8)	-0.687	0.087	0.503	0.335	0.000 ***
64k	FP8 - (GPTQ-int4)	0.285	0.088	1.330	0.571	0.018 *
64k	FP8 - (GPTQ-int8)	-0.041	0.090	0.960	0.490	1.000
64k	(GPTQ-int4) - (GPTQ-int8)	-0.325	0.088	0.722	0.419	0.003 **
128k	BF16 - (AWQ-int4)	0.317	0.081	1.373	0.579	0.001 **
128k	BF16 - (BNB-nf4)	1.149	0.083	3.154	0.759	0.000 ***
128k	BF16 - FP8	0.226	0.081	1.253	0.556	0.084
128k	BF16 - (GPTQ-int4)	0.575	0.081	1.778	0.640	0.000 ***
128k	BF16 - (GPTQ-int8)	0.064	0.082	1.066	0.516	1.000
128k	(AWQ-int4) - (BNB-nf4)	0.831	0.082	2.296	0.697	0.000 ***
128k	(AWQ-int4) - FP8	-0.092	0.081	0.912	0.477	1.000
128k	(AWQ-int4) - (GPTQ-int4)	0.258	0.081	1.294	0.564	0.021 *
128k	(AWQ-int4) - (GPTQ-int8)	-0.254	0.081	0.776	0.437	0.027 *
128k	(BNB-nf4) - FP8	-0.923	0.083	0.397	0.284	0.000 ***
128k	(BNB-nf4) - (GPTQ-int4)	-0.573	0.082	0.564	0.360	0.000 ***
128k	(BNB-nf4) - (GPTQ-int8)	-1.085	0.083	0.338	0.253	0.000 ***
128k	FP8 - (GPTQ-int4)	0.350	0.081	1.419	0.587	0.000 ***
128k	FP8 - (GPTQ-int8)	-0.162	0.081	0.850	0.460	0.695
128k	(GPTQ-int4) - (GPTQ-int8)	-0.512	0.081	0.599	0.375	0.000 ***

Table 6: **RULER**: Post-hoc comparisons between quantization methods for **accuracy** (Table 5) across context lengths using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).

glmer(Binary_label ~ Quantization * Task + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	1.790	5.992	3.347	10.729	0.000	***
AWQ-int4	-0.120	0.887	0.730	1.077	0.225	
BNB-nf4	-0.935	0.393	0.328	0.470	0.000	***
FP8	-0.145	0.865	0.713	1.050	0.143	
GPTQ-int4	-0.291	0.747	0.618	0.904	0.003	**
GPTQ-int8	-0.061	0.941	0.773	1.144	0.540	
multi-key	-0.877	0.416	0.348	0.498	0.000	***
multi-value	-0.382	0.683	0.566	0.823	0.000	***
AWQ-int4:multi-key	-0.060	0.942	0.732	1.210	0.639	
BNB-nf4:multi-key	0.313	1.367	1.078	1.733	0.010	**
FP8:multi-key	0.053	1.054	0.820	1.355	0.679	
GPTQ-int4:multi-key	-0.142	0.867	0.678	1.109	0.257	
GPTQ-int8:multi-key	-0.003	0.997	0.774	1.284	0.979	
AWQ-int4:multi-value	-0.053	0.948	0.730	1.231	0.688	
BNB-nf4:multi-value	0.481	1.618	1.264	2.072	0.000	***
FP8:multi-value	-0.072	0.931	0.718	1.207	0.588	
GPTQ-int4:multi-value	-0.084	0.920	0.712	1.187	0.520	
GPTQ-int8:multi-value	0.061	1.063	0.816	1.386	0.650	
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	0.428					
ICC (Model)	0.300					
N (Model)	5					
Observations	27000					
R^2 (marginal)	0.157					
R^2 (conditional)	0.047					

Table 7: **RULER**: Summary of generalized linear mixed model with **quantization**, **task**, and their interaction as predictors of **accuracy**: `glmer(Binary_label ~ Quantization * Task + (1|Model), data = data, family = binomial)`. The quantization method "BF16" on the *needle* task was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. See Table 8 for post-hoc analysis.

TASK	CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
needle	BF16 - (AWQ-int4)	0.120	0.099	1.128	0.530	1.000	
needle	BF16 - (BNB-nf4)	0.935	0.091	2.546	0.718	0.000	***
needle	BF16 - FP8	0.145	0.099	1.155	0.536	1.000	
needle	BF16 - (GPTQ-int4)	0.291	0.097	1.338	0.572	0.040	*
needle	BF16 - (GPTQ-int8)	0.061	0.100	1.063	0.515	1.000	
needle	(AWQ-int4) - (BNB-nf4)	0.814	0.091	2.257	0.693	0.000	***
needle	(AWQ-int4) - FP8	0.024	0.098	1.024	0.506	1.000	
needle	(AWQ-int4) - (GPTQ-int4)	0.171	0.096	1.186	0.543	1.000	
needle	(AWQ-int4) - (GPTQ-int8)	−0.059	0.099	0.943	0.485	1.000	
needle	(BNB-nf4) - FP8	−0.790	0.090	0.454	0.312	0.000	***
needle	(BNB-nf4) - (GPTQ-int4)	−0.643	0.088	0.526	0.345	0.000	***
needle	(BNB-nf4) - (GPTQ-int8)	−0.873	0.092	0.418	0.295	0.000	***
needle	FP8 - (GPTQ-int4)	0.147	0.096	1.158	0.537	1.000	
needle	FP8 - (GPTQ-int8)	−0.083	0.099	0.920	0.479	1.000	
needle	(GPTQ-int4) - (GPTQ-int8)	−0.230	0.097	0.794	0.443	0.264	
multi-key	BF16 - (AWQ-int4)	0.181	0.082	1.198	0.545	0.404	
multi-key	BF16 - (BNB-nf4)	0.622	0.080	1.863	0.651	0.000	***
multi-key	BF16 - FP8	0.092	0.082	1.096	0.523	1.000	
multi-key	BF16 - (GPTQ-int4)	0.434	0.081	1.543	0.607	0.000	***
multi-key	BF16 - (GPTQ-int8)	0.065	0.082	1.067	0.516	1.000	
multi-key	(AWQ-int4) - (BNB-nf4)	0.441	0.079	1.555	0.609	0.000	***
multi-key	(AWQ-int4) - FP8	−0.089	0.081	0.915	0.478	1.000	
multi-key	(AWQ-int4) - (GPTQ-int4)	0.253	0.080	1.288	0.563	0.022	*
multi-key	(AWQ-int4) - (GPTQ-int8)	−0.116	0.081	0.891	0.471	1.000	
multi-key	(BNB-nf4) - FP8	−0.530	0.080	0.588	0.370	0.000	***
multi-key	(BNB-nf4) - (GPTQ-int4)	−0.188	0.078	0.828	0.453	0.235	
multi-key	(BNB-nf4) - (GPTQ-int8)	−0.557	0.080	0.573	0.364	0.000	***
multi-key	FP8 - (GPTQ-int4)	0.342	0.080	1.408	0.585	0.000	***
multi-key	FP8 - (GPTQ-int8)	−0.027	0.082	0.974	0.493	1.000	
multi-key	(GPTQ-int4) - (GPTQ-int8)	−0.369	0.080	0.691	0.409	0.000	***
multi-value	BF16 - (AWQ-int4)	0.174	0.090	1.190	0.543	0.788	
multi-value	BF16 - (BNB-nf4)	0.453	0.087	1.573	0.611	0.000	***
multi-value	BF16 - FP8	0.216	0.089	1.242	0.554	0.231	
multi-value	BF16 - (GPTQ-int4)	0.375	0.088	1.455	0.593	0.000	***
multi-value	BF16 - (GPTQ-int8)	0.000	0.092	1.000	0.500	1.000	
multi-value	(AWQ-int4) - (BNB-nf4)	0.279	0.086	1.322	0.569	0.017	*
multi-value	(AWQ-int4) - FP8	0.042	0.088	1.043	0.511	1.000	
multi-value	(AWQ-int4) - (GPTQ-int4)	0.201	0.086	1.223	0.550	0.297	
multi-value	(AWQ-int4) - (GPTQ-int8)	−0.174	0.090	0.840	0.457	0.798	
multi-value	(BNB-nf4) - FP8	−0.237	0.085	0.789	0.441	0.084	
multi-value	(BNB-nf4) - (GPTQ-int4)	−0.078	0.084	0.925	0.481	1.000	
multi-value	(BNB-nf4) - (GPTQ-int8)	−0.453	0.088	0.636	0.389	0.000	***
multi-value	FP8 - (GPTQ-int4)	0.159	0.086	1.172	0.540	0.969	
multi-value	FP8 - (GPTQ-int8)	−0.216	0.090	0.805	0.446	0.236	
multi-value	(GPTQ-int4) - (GPTQ-int8)	−0.375	0.088	0.687	0.407	0.000	***

Table 8: **RULER**: Post-hoc comparisons between quantization methods for **accuracy** (Table 7) across tasks using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second.

glmer(Binary_label ~ Quantization + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	1.710	5.532	3.712	8.243	0.000	***
AWQ-int4	-0.096	0.908	0.876	0.943	0.000	***
BNB-nf4	-0.636	0.529	0.511	0.548	0.000	***
FP8	0.012	1.012	0.975	1.051	0.516	
GPTQ-int4	-0.173	0.841	0.811	0.872	0.000	***
GPTQ-int8	-0.024	0.976	0.940	1.013	0.200	
RANDOM EFFECTS						
σ^2 (residual variance)	3.290					
τ_{00} (Model)	0.841					
ICC (Model)	0.204					
N (Model)	5					
Observations	234000					
R^2 (marginal)	0.012					
R^2 (conditional)	0.213					

Table 9: **ONERULER**: Summary of generalized linear mixed model with **quantization** as the predictor of **accuracy**: glmer(Binary_label ~ Quantization + (1|Model), data = data, family = binomial). The quantization method "BF16" was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. See Table 10 for post-hoc analysis.

CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
BF16 - (AWQ-int4)	0.096	0.019	1.101	0.524	0.000	***
BF16 - (BNB-nf4)	0.636	0.018	1.889	0.654	0.000	***
BF16 - FP8	-0.012	0.019	0.988	0.497	1.000	
BF16 - (GPTQ-int4)	0.173	0.019	1.189	0.543	0.000	***
BF16 - (GPTQ-int8)	0.024	0.019	1.025	0.506	1.000	
(AWQ-int4) - (BNB-nf4)	0.540	0.018	1.716	0.632	0.000	***
(AWQ-int4) - FP8	-0.108	0.019	0.897	0.473	0.000	***
(AWQ-int4) - (GPTQ-int4)	0.077	0.019	1.081	0.519	0.000	***
(AWQ-int4) - (GPTQ-int8)	-0.072	0.019	0.931	0.482	0.002	**
(BNB-nf4) - FP8	-0.649	0.018	0.523	0.343	0.000	***
(BNB-nf4) - (GPTQ-int4)	-0.463	0.018	0.630	0.386	0.000	***
(BNB-nf4) - (GPTQ-int8)	-0.612	0.018	0.542	0.352	0.000	***
FP8 - (GPTQ-int4)	0.186	0.019	1.204	0.546	0.000	***
FP8 - (GPTQ-int8)	0.037	0.019	1.037	0.509	0.821	
(GPTQ-int4) - (GPTQ-int8)	-0.149	0.019	0.861	0.463	0.000	***

Table 10: **ONERULER**: Post-hoc comparisons between quantization methods for **accuracy** (Table 9) using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).

glmer(Binary_label ~ Quantization * Context_length + (1 Model) + (1 Frequency), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	2.403	11.058	6.371	19.193	0.000	***
AWQ-int4	0.010	1.010	0.935	1.090	0.806	
BNB-nf4	-0.497	0.609	0.566	0.654	0.000	***
FP8	0.108	1.115	1.030	1.206	0.007	**
GPTQ-int4	-0.087	0.916	0.849	0.988	0.024	*
GPTQ-int8	-0.127	0.881	0.818	0.949	0.001	***
64k	-0.702	0.496	0.463	0.531	0.000	***
128k	-1.237	0.290	0.271	0.310	0.000	***
AWQ-int4:64k	-0.106	0.900	0.815	0.993	0.036	*
BNB-nf4:64k	-0.291	0.747	0.681	0.820	0.000	***
FP8:64k	-0.135	0.873	0.790	0.965	0.008	**
GPTQ-int4:64k	-0.182	0.833	0.757	0.918	0.000	***
GPTQ-int8:64k	0.071	1.073	0.975	1.181	0.149	
AWQ-int4:128k	-0.186	0.831	0.755	0.913	0.000	***
BNB-nf4:128k	-0.263	0.768	0.701	0.842	0.000	***
FP8:128k	-0.112	0.894	0.811	0.986	0.025	*
GPTQ-int4:128k	-0.096	0.909	0.826	0.999	0.048	*
GPTQ-int8:128k	0.182	1.199	1.092	1.317	0.000	***
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	3.29					
τ_{00} (Frequency)	0.723					
ICC (Model)	0.218					
ICC (Frequency)	0.141					
N (Model)	5					
N (Frequency)	3					
Observations	234000					
R^2 (marginal)	0.065					
R^2 (conditional)	0.401					

Table 11: **ONERULER**: Summary of generalized linear mixed model with **quantization**, **context length**, and their interaction as predictors of **accuracy**: `glmer(Binary_label ~ Quantization * Context_length + (1|Model) + (1|Frequency), data = data, family = binomial)`. The quantization method "BF16" at context length 8k was set as the reference level (intercept), with Model and Frequency set as the grouping factors for random intercepts to account for variability across models and frequency bands. See Table 12 for post-hoc analysis.

CONTEXT	CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
8k	BF16 - (AWQ-int4)	-0.010	0.039	0.990	0.498	1.000	
8k	BF16 - (BNB-nf4)	0.497	0.037	1.643	0.622	0.000	***
8k	BF16 - FP8	-0.108	0.040	0.897	0.473	0.102	
8k	BF16 - (GPTQ-int4)	0.087	0.039	1.091	0.522	0.354	
8k	BF16 - (GPTQ-int8)	0.127	0.038	1.135	0.532	0.012	*
8k	(AWQ-int4) - (BNB-nf4)	0.506	0.038	1.659	0.624	0.000	***
8k	(AWQ-int4) - FP8	-0.099	0.042	0.906	0.475	0.278	
8k	(AWQ-int4) - (GPTQ-int4)	0.097	0.041	1.102	0.524	0.253	
8k	(AWQ-int4) - (GPTQ-int8)	0.136	0.040	1.146	0.534	0.011	*
8k	(BNB-nf4) - FP8	-0.605	0.039	0.546	0.353	0.000	***
8k	(BNB-nf4) - (GPTQ-int4)	-0.409	0.038	0.664	0.399	0.000	***
8k	(BNB-nf4) - (GPTQ-int8)	-0.370	0.038	0.691	0.409	0.000	***
8k	FP8 - (GPTQ-int4)	0.196	0.041	1.216	0.549	0.000	***
8k	FP8 - (GPTQ-int8)	0.235	0.041	1.265	0.559	0.000	***
8k	(GPTQ-int4) - (GPTQ-int8)	0.039	0.040	1.040	0.510	1.000	
64k	BF16 - (AWQ-int4)	0.096	0.033	1.101	0.524	0.059	
64k	BF16 - (BNB-nf4)	0.788	0.032	2.199	0.687	0.000	***
64k	BF16 - FP8	0.027	0.033	1.027	0.507	1.000	
64k	BF16 - (GPTQ-int4)	0.270	0.033	1.310	0.567	0.000	***
64k	BF16 - (GPTQ-int8)	0.056	0.033	1.058	0.514	1.000	
64k	(AWQ-int4) - (BNB-nf4)	0.692	0.032	1.998	0.666	0.000	***
64k	(AWQ-int4) - FP8	-0.069	0.034	0.933	0.483	0.618	
64k	(AWQ-int4) - (GPTQ-int4)	0.174	0.033	1.190	0.543	0.000	***
64k	(AWQ-int4) - (GPTQ-int8)	-0.040	0.034	0.961	0.490	1.000	
64k	(BNB-nf4) - FP8	-0.761	0.032	0.467	0.318	0.000	***
64k	(BNB-nf4) - (GPTQ-int4)	-0.518	0.031	0.596	0.373	0.000	***
64k	(BNB-nf4) - (GPTQ-int8)	-0.732	0.032	0.481	0.325	0.000	***
64k	FP8 - (GPTQ-int4)	0.243	0.033	1.275	0.560	0.000	***
64k	FP8 - (GPTQ-int8)	0.029	0.034	1.029	0.507	1.000	
64k	(GPTQ-int4) - (GPTQ-int8)	-0.214	0.033	0.807	0.447	0.000	***
128k	BF16 - (AWQ-int4)	0.176	0.030	1.193	0.544	0.000	***
128k	BF16 - (BNB-nf4)	0.760	0.029	2.138	0.681	0.000	***
128k	BF16 - FP8	0.003	0.031	1.003	0.501	1.000	
128k	BF16 - (GPTQ-int4)	0.183	0.030	1.201	0.546	0.000	***
128k	BF16 - (GPTQ-int8)	-0.055	0.031	0.947	0.486	1.000	
128k	(AWQ-int4) - (BNB-nf4)	0.584	0.029	1.793	0.642	0.000	***
128k	(AWQ-int4) - FP8	-0.173	0.030	0.841	0.457	0.000	***
128k	(AWQ-int4) - (GPTQ-int4)	0.007	0.030	1.007	0.502	1.000	
128k	(AWQ-int4) - (GPTQ-int8)	-0.231	0.031	0.794	0.443	0.000	***
128k	(BNB-nf4) - FP8	-0.757	0.030	0.469	0.319	0.000	***
128k	(BNB-nf4) - (GPTQ-int4)	-0.577	0.029	0.562	0.360	0.000	***
128k	(BNB-nf4) - (GPTQ-int8)	-0.815	0.030	0.443	0.307	0.000	***
128k	FP8 - (GPTQ-int4)	0.180	0.030	1.197	0.545	0.000	***
128k	FP8 - (GPTQ-int8)	-0.058	0.031	0.943	0.485	0.899	
128k	(GPTQ-int4) - (GPTQ-int8)	-0.238	0.031	0.788	0.441	0.000	***

Table 12: **ONERULER**: Post-hoc comparisons between quantization methods for **accuracy** (Table 11) across context lengths using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second.

glmer(Binary_label ~ Quantization * Frequency + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	0.627	1.871	0.752	4.656	0.178	
AWQ-int4	-0.176	0.839	0.777	0.906	0.000	***
BNB-nf4	-0.380	0.684	0.634	0.739	0.000	***
FP8	-0.009	0.991	0.917	1.070	0.813	
GPTQ-int4	-0.146	0.864	0.800	0.933	0.000	***
GPTQ-int8	-0.056	0.946	0.875	1.022	0.158	
English	1.667	5.298	4.625	6.069	0.000	***
High-resource	1.362	3.904	3.670	4.154	0.000	***
AWQ-int4:English	0.005	1.005	0.818	1.237	0.959	
BNB-nf4:English	0.110	1.116	0.917	1.360	0.273	
FP8:English	0.127	1.136	0.923	1.398	0.229	
GPTQ-int4:English	-0.012	0.988	0.800	1.219	0.907	
GPTQ-int8:English	0.131	1.140	0.931	1.396	0.205	
AWQ-int4:High-resource	0.099	1.105	1.012	1.206	0.027	*
BNB-nf4:High-resource	-0.377	0.686	0.629	0.748	0.000	***
FP8:High-resource	0.025	1.026	0.939	1.121	0.574	
GPTQ-int4:High-resource	-0.049	0.952	0.872	1.039	0.271	
GPTQ-int8:High-resource	0.035	1.036	0.948	1.132	0.435	
RANDOM EFFECTS						
σ^2 (residual variance)	3.290					
τ_{00} (Model)	0.924					
ICC (Model)	0.219					
N (Model)	5					
Observations	234000					
R^2 (marginal)	0.067					
R^2 (conditional)	0.271					

Table 13: **ONERULER**: Summary of generalized linear mixed model with **quantization**, **language frequency**, and their interaction as predictors of **accuracy**: glmer(Binary_label ~ Quantization * Frequency + (1|Model), data = data, family = binomial). The quantization method "BF16" under low-resource frequency was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. See Table 14 for post-hoc analysis.

FREQUENCY	CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
Low-resource	BF16 - (AWQ-int4)	0.176	0.039	1.192	0.544	0.001	***
Low-resource	BF16 - (BNB-nf4)	0.380	0.039	1.462	0.594	0.000	***
Low-resource	BF16 - FP8	0.009	0.039	1.009	0.502	1.000	
Low-resource	BF16 - (GPTQ-int4)	0.146	0.039	1.157	0.536	0.003	**
Low-resource	BF16 - (GPTQ-int8)	0.056	0.039	1.057	0.514	1.000	
Low-resource	(AWQ-int4) - (BNB-nf4)	0.204	0.040	1.226	0.551	0.000	***
Low-resource	(AWQ-int4) - FP8	-0.167	0.040	0.846	0.458	0.001	**
Low-resource	(AWQ-int4) - (GPTQ-int4)	-0.030	0.040	0.971	0.493	1.000	
Low-resource	(AWQ-int4) - (GPTQ-int8)	-0.120	0.040	0.887	0.470	0.044	*
Low-resource	(BNB-nf4) - FP8	-0.370	0.040	0.691	0.408	0.000	***
Low-resource	(BNB-nf4) - (GPTQ-int4)	-0.234	0.040	0.792	0.442	0.000	***
Low-resource	(BNB-nf4) - (GPTQ-int8)	-0.324	0.040	0.723	0.420	0.000	***
Low-resource	FP8 - (GPTQ-int4)	0.137	0.041	1.147	0.534	0.012	*
Low-resource	FP8 - (GPTQ-int8)	0.046	0.041	1.047	0.512	1.000	
Low-resource	(GPTQ-int4) - (GPTQ-int8)	-0.090	0.041	0.914	0.477	0.396	
English	BF16 - (AWQ-int4)	0.171	0.101	1.186	0.543	1.000	
English	BF16 - (BNB-nf4)	0.269	0.096	1.309	0.567	0.076	
English	BF16 - FP8	-0.118	0.102	0.889	0.471	1.000	
English	BF16 - (GPTQ-int4)	0.159	0.102	1.172	0.540	1.000	
English	BF16 - (GPTQ-int8)	-0.075	0.100	0.927	0.481	1.000	
English	(AWQ-int4) - (BNB-nf4)	0.099	0.107	1.104	0.525	1.000	
English	(AWQ-int4) - FP8	-0.289	0.113	0.749	0.428	0.162	
English	(AWQ-int4) - (GPTQ-int4)	-0.012	0.108	0.988	0.497	1.000	
English	(AWQ-int4) - (GPTQ-int8)	-0.246	0.110	0.782	0.439	0.376	
English	(BNB-nf4) - FP8	-0.387	0.113	0.679	0.404	0.009	**
English	(BNB-nf4) - (GPTQ-int4)	-0.111	0.108	0.895	0.472	1.000	
English	(BNB-nf4) - (GPTQ-int8)	-0.345	0.109	0.708	0.415	0.023	*
English	FP8 - (GPTQ-int4)	0.277	0.116	1.319	0.569	0.251	
English	FP8 - (GPTQ-int8)	0.043	0.117	1.044	0.511	1.000	
English	(GPTQ-int4) - (GPTQ-int8)	-0.234	0.111	0.792	0.442	0.535	
High-resource	BF16 - (AWQ-int4)	0.077	0.023	1.080	0.519	0.011	*
High-resource	BF16 - (BNB-nf4)	0.756	0.021	2.130	0.680	0.000	***
High-resource	BF16 - FP8	-0.016	0.023	0.984	0.496	1.000	
High-resource	BF16 - (GPTQ-int4)	0.195	0.022	1.216	0.549	0.000	***
High-resource	BF16 - (GPTQ-int8)	0.020	0.023	1.021	0.505	1.000	
High-resource	(AWQ-int4) - (BNB-nf4)	0.680	0.021	1.973	0.664	0.000	***
High-resource	(AWQ-int4) - FP8	-0.093	0.023	0.911	0.477	0.001	**
High-resource	(AWQ-int4) - (GPTQ-int4)	0.119	0.022	1.126	0.530	0.000	***
High-resource	(AWQ-int4) - (GPTQ-int8)	-0.056	0.023	0.945	0.486	0.197	
High-resource	(BNB-nf4) - FP8	-0.772	0.021	0.462	0.316	0.000	***
High-resource	(BNB-nf4) - (GPTQ-int4)	-0.561	0.021	0.571	0.363	0.000	***
High-resource	(BNB-nf4) - (GPTQ-int8)	-0.736	0.021	0.479	0.324	0.000	***
High-resource	FP8 - (GPTQ-int4)	0.212	0.022	1.236	0.553	0.000	***
High-resource	FP8 - (GPTQ-int8)	0.037	0.023	1.037	0.509	1.000	
High-resource	(GPTQ-int4) - (GPTQ-int8)	-0.175	0.022	0.839	0.456	0.000	***

Table 14: **ONERULER**: Post-hoc comparisons between quantization methods for **accuracy** (Table 13) across language frequency using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second.

glmer(Binary_label ~ Quantization + (1 Context_length) + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	0.510	1.666	1.302	2.132	0.000	***
FP8	0.049	1.050	0.981	1.123	0.159	
GPTQ-int8	-0.018	0.983	0.918	1.052	0.613	
AWQ-int4	-0.005	0.995	0.931	1.065	0.893	
GPTQ-int4	-0.002	0.998	0.933	1.067	0.952	
BNB-nf4	-0.046	0.955	0.893	1.022	0.181	
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	0.047					
ICC (Model)	0.045					
N (Model)	5					
Observations	42781					
R^2 (marginal)	0.016					
R^2 (conditional)	0.001					

Table 15: **NOCHA**: Summary of generalized linear mixed model with **quantization** as the predictor of **accuracy**: glmer(Binary_label ~ Quantization + (1|Context_length) + (1|Model), data = data, family = binomial). The quantization method "BF16" was set as the reference level (intercept), with Context_length and Model set as the grouping factors for random intercepts to account for variability across context lengths and models. See Table 16 for post-hoc analysis.

CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value
BF16 - FP8	-0.049	0.034	0.953	0.488	1.000
BF16 - GPTQ-int8	0.018	0.035	1.018	0.504	1.000
BF16 - AWQ-int4	0.005	0.034	1.005	0.501	1.000
BF16 - GPTQ-int4	0.002	0.034	1.002	0.501	1.000
BF16 - BNB-nf4	0.046	0.034	1.047	0.511	1.000
AWQ-int4- FP8	-0.053	0.034	0.948	0.487	1.000
AWQ-int4- GPTQ-int8	0.013	0.035	1.013	0.503	1.000
AWQ-int4- GPTQ-int4	-0.003	0.034	0.997	0.499	1.000
AWQ-int4- BNB-nf4	0.041	0.034	1.042	0.510	1.000
BNB-nf4 - FP8	-0.094	0.034	0.910	0.476	0.092
BNB-nf4 - GPTQ-int8	-0.028	0.035	0.972	0.493	1.000
BNB-nf4 - GPTQ-int4	-0.044	0.034	0.957	0.489	1.000
FP8 - GPTQ-int8	0.066	0.035	1.068	0.517	0.882
FP8 - GPTQ-int4	0.051	0.034	1.052	0.513	1.000
GPTQ-int4 - GPTQ-int8	0.016	0.035	1.016	0.504	1.000

Table 16: **NOCHA**: Post-hoc comparisons between quantization methods for **accuracy** (Table 15) across all context lengths using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).

<code>glmer(Binary_label ~ Quantization * Context_length + (1 Model), data = data, family = binomial)</code>					
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value
Intercept	0.588	1.800	1.440	2.250	0.000 ***
FP8	0.126	1.135	0.955	1.348	0.150
GPTQ-int8	0.015	1.016	0.856	1.204	0.859
AWQ-int4	0.022	1.022	0.862	1.211	0.803
GPTQ-int4	-0.027	0.973	0.821	1.153	0.752
BNB-nf4	-0.091	0.913	0.772	1.081	0.291
Context length>75k	-0.160	0.852	0.749	0.970	0.015 *
FP8:>75k	-0.091	0.913	0.757	1.100	0.337
GPTQ-int8:>75k	-0.039	0.962	0.799	1.157	0.678
AWQ-int4:>75k	-0.031	0.970	0.807	1.165	0.742
GPTQ-int4:>75k	0.030	1.030	0.858	1.238	0.751
BNB-nf4:>75k	0.053	1.054	0.878	1.266	0.570
RANDOM EFFECTS					
σ^2 (residual variance)	1.000				
τ_{00} (Model)	0.047				
ICC (Model)	0.045				
N (Model)	5				
Observations	42781				
R^2 (marginal)	0.016				
R^2 (conditional)	0.001				

Table 17: **NOCHA**: Summary of generalized linear mixed model with **quantization**, **context length**, and their interaction as predictors of **accuracy**: `glmer(Binary_label ~ Quantization * Context_length + (1|Model), data = data, family = binomial)`. The quantization method "BF16" at context length $\leq 75k$ was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. See Table 18 for post-hoc analysis.

CONTEXT	CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value
<75k	BF16 - (AWQ-int4)	-0.022	0.087	0.979	0.495	1.000
<75k	BF16 - (BNB-nf4)	0.091	0.086	1.095	0.523	1.000
<75k	BF16 - FP8	-0.126	0.088	0.881	0.468	1.000
<75k	BF16 - (GPTQ-int4)	0.027	0.086	1.028	0.507	1.000
<75k	BF16 - (GPTQ-int8)	-0.015	0.087	0.985	0.496	1.000
<75k	(AWQ-int4)- (BNB-nf4)	0.112	0.088	1.119	0.528	1.000
<75k	(AWQ-int4)- FP8	-0.105	0.090	0.900	0.474	1.000
<75k	(AWQ-int4)- (GPTQ-int4)	0.049	0.089	1.050	0.512	1.000
<75k	(AWQ-int4)- (GPTQ-int8)	0.006	0.089	1.006	0.502	1.000
<75k	(BNB-nf4) - FP8	-0.217	0.089	0.805	0.446	0.216
<75k	(BNB-nf4) - (GPTQ-int4)	-0.063	0.088	0.939	0.484	1.000
<75k	(BNB-nf4) - (GPTQ-int8)	-0.106	0.088	0.899	0.473	1.000
<75k	FP8 - (GPTQ-int4)	0.154	0.089	1.166	0.538	1.000
<75k	FP8 - (GPTQ-int8)	0.111	0.090	1.117	0.528	1.000
<75k	(GPTQ-int4) - (GPTQ-int8)	-0.043	0.089	0.958	0.489	1.000
>75k	BF16 - (AWQ-int4)	0.009	0.037	1.009	0.502	1.000
>75k	BF16 - (BNB-nf4)	0.038	0.037	1.039	0.509	1.000
>75k	BF16 - FP8	-0.035	0.037	0.966	0.491	1.000
>75k	BF16 - (GPTQ-int4)	-0.002	0.037	0.998	0.499	1.000
>75k	BF16 - (GPTQ-int8)	0.024	0.038	1.024	0.506	1.000
>75k	(AWQ-int4)- (BNB-nf4)	0.029	0.037	1.029	0.507	1.000
>75k	(AWQ-int4)- FP8	-0.044	0.037	0.957	0.489	1.000
>75k	(AWQ-int4)- (GPTQ-int4)	-0.012	0.037	0.989	0.497	1.000
>75k	(AWQ-int4)- (GPTQ-int8)	0.015	0.038	1.015	0.504	1.000
>75k	(BNB-nf4) - FP8	-0.073	0.037	0.930	0.482	0.758
>75k	(BNB-nf4) - (GPTQ-int4)	-0.040	0.037	0.961	0.490	1.000
>75k	(BNB-nf4) - (GPTQ-int8)	-0.014	0.038	0.986	0.496	1.000
>75k	FP8 - (GPTQ-int4)	0.033	0.037	1.033	0.508	1.000
>75k	FP8 - (GPTQ-int8)	0.059	0.038	1.061	0.515	1.000
>75k	(GPTQ-int4) - (GPTQ-int8)	0.026	0.038	1.026	0.507	1.000

Table 18: **NOCHA**: Post-hoc comparisons between quantization methods for **accuracy** (Table 17) across context lengths below and above 75k tokens using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).

lmer(log_veriscore ~ Quantization + (1 topic) + (1 Model), data = data)						
PREDICTORS	ESTIMATE	CI (LOWER)	CI (UPPER)	SE	p-value	
Intercept	2.183	1.754	2.611	0.219	0.000	***
AWQ-int4	-0.079	-0.181	0.023	0.052	0.129	
BNB-nf4	-0.228	-0.330	-0.126	0.052	0.000	***
FP8	0.074	-0.028	0.176	0.052	0.156	
GPTQ-int4	-0.127	-0.229	-0.025	0.052	0.015	*
GPTQ-int8	-0.003	-0.105	0.099	0.052	0.956	
RANDOM EFFECTS						
σ^2 (Residual)	0.679					
τ_{00} (Model)	0.118					
τ_{00} (Topic)	2.274					
ICC (Model)	0.039					
ICC (Topic)	0.740					
N (Model)	5					
N (Topic)	100					
Observations	3000					
R^2 (marginal)	0.003					
R^2 (conditional)	0.780					

Table 19: **FACTSCORE**: Summary of linear mixed-effects model with **quantization** as the predictor of **VERIScore**: `lmer(log_veriscore ~ Quantization + (1|topic) + (1|Model), data = data)`. The quantization method "BF16" was set as the reference level (intercept), with topic and Model set as grouping factors for random intercepts to account for variability. See Table 20 for post-hoc analysis.

CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
BF16 - (AWQ-int4)	0.079	0.052	1.082	0.520	1.000	
BF16 - (BNB-nf4)	0.228	0.052	1.256	0.557	0.000	***
BF16 - FP8	-0.074	0.052	0.929	0.482	1.000	
BF16 - (GPTQ-int4)	0.127	0.052	1.135	0.532	0.227	
BF16 - (GPTQ-int8)	0.003	0.052	1.003	0.501	1.000	
(AWQ-int4) - (BNB-nf4)	0.149	0.052	1.161	0.537	0.063	
(AWQ-int4) - FP8	-0.153	0.052	0.858	0.462	0.050	*
(AWQ-int4) - (GPTQ-int4)	0.048	0.052	1.049	0.512	1.000	
(AWQ-int4) - (GPTQ-int8)	-0.076	0.052	0.927	0.481	1.000	
(BNB-nf4) - FP8	-0.302	0.052	0.739	0.425	0.000	***
(BNB-nf4) - (GPTQ-int4)	-0.102	0.052	0.903	0.474	0.770	
(BNB-nf4) - (GPTQ-int8)	-0.225	0.052	0.798	0.444	0.000	***
FP8 - (GPTQ-int4)	0.201	0.052	1.222	0.550	0.002	**
FP8 - (GPTQ-int8)	0.077	0.052	1.080	0.519	1.000	
(GPTQ-int4) - (GPTQ-int8)	-0.124	0.052	0.884	0.469	0.263	

Table 20: **FACTSCORE**: Post-hoc comparisons between quantization methods for **VERIScore** (Table 19) using Bonferroni adjustments for multiple pairwise comparisons. The odds ratios and probabilities are derived from linear model estimates. A probability near 0.5 indicates that the two methods are comparably accurate (in **VERIScore**).

glmer(no_claims ~ Quantization + (1 Model) + (1 topic), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	-2.961	0.052	0.010	0.262	0.000	***
AWQ-int4	0.351	1.421	0.941	2.144	0.095	
BNB-nf4	0.831	2.295	1.522	3.459	0.000	***
FP8	-0.048	0.954	0.630	1.444	0.822	
GPTQ-int4	0.466	1.594	1.057	2.404	0.026	*
GPTQ-int8	-0.024	0.977	0.645	1.479	0.911	
RANDOM EFFECTS						
σ^2 (residual variance, logistic)	3.290					
τ_{00} (Model)	2.669					
τ_{00} (Topic)	11.401					
ICC (Model)	0.154					
ICC (Topic)	0.657					
N (Model)	5					
N (Topic)	100					
Observations	3000					
R^2 (marginal)	0.006					
R^2 (conditional)	0.812					

Table 21: **FACTSCORE**: Summary of generalized linear mixed model with **quantization** as the predictor of **no claim rate**: glmer(no_claims ~ Quantization + (1|Model) + (1|topic), data = data, family = binomial). The quantization method "BF16" was set as the reference level (intercept), with Model and topic set as grouping factors for random intercepts to account for variability across models and topics. See Table 22 for post-hoc analysis.

CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
BF16 - (AWQ-int4)	-0.351	0.210	0.704	0.413	1.000	
BF16 - (BNB-nf4)	-0.831	0.209	0.436	0.304	0.001	**
BF16 - FP8	0.048	0.212	1.049	0.512	1.000	
BF16 - (GPTQ-int4)	-0.466	0.210	0.627	0.385	0.392	
BF16 - (GPTQ-int8)	0.024	0.212	1.024	0.506	1.000	
(AWQ-int4) - (BNB-nf4)	-0.479	0.206	0.619	0.382	0.302	
(AWQ-int4) - FP8	0.399	0.210	1.490	0.598	0.871	
(AWQ-int4) - (GPTQ-int4)	-0.115	0.207	0.891	0.471	1.000	
(AWQ-int4) - (GPTQ-int8)	0.375	0.210	1.455	0.593	1.000	
(BNB-nf4) - FP8	0.878	0.210	2.406	0.706	0.000	***
(BNB-nf4) - (GPTQ-int4)	0.364	0.206	1.439	0.590	1.000	
(BNB-nf4) - (GPTQ-int8)	0.854	0.210	2.350	0.701	0.001	**
FP8 - (GPTQ-int4)	-0.514	0.210	0.598	0.374	0.217	
FP8 - (GPTQ-int8)	-0.024	0.212	0.976	0.494	1.000	
(GPTQ-int4) - (GPTQ-int8)	0.490	0.210	1.632	0.620	0.293	

Table 22: **FACTSCORE**: Post-hoc comparisons between quantization methods for **no claim rate** (Table 21) using Bonferroni adjustments for multiple pairwise comparisons. Results are reported on the log-odds ratio scale. The probability values refer to the likelihood that the quantization method listed first in each contrast abstains more than the one listed second (i.e., a value of 0.5 suggests comparable rate of empty claims).

lmer(log_veriscore ~ Quantization * cat + (1 Model) + (1 topic), data = data)						
PREDICTORS	ESTIMATE	CI (LOWER)	CI (UPPER)	SE	p-value	
Intercept	2.042	1.639	2.445	0.206	0.000	***
AWQ-int4	-0.142	-0.320	0.035	0.091	0.116	
BNB-nf4	-0.430	-0.607	-0.252	0.091	0.000	***
FP8	0.089	-0.089	0.267	0.091	0.326	
GPTQ-int4	-0.197	-0.374	-0.019	0.091	0.030	*
GPTQ-int8	0.023	-0.155	0.200	0.091	0.803	
very freq	1.831	1.456	2.205	0.191	0.000	***
very rare	-1.460	-1.837	-1.082	0.193	0.000	***
AWQ-int4 × very freq	0.014	-0.236	0.263	0.127	0.915	
BNB-nf4 × very freq	0.354	0.105	0.604	0.127	0.005	**
FP8 × very freq	-0.064	-0.314	0.185	0.127	0.612	
GPTQ-int4 × very freq	0.084	-0.166	0.333	0.127	0.511	
GPTQ-int8 × very freq	-0.060	-0.309	0.189	0.127	0.638	
AWQ-int4 × very rare	0.177	-0.074	0.428	0.128	0.166	
BNB-nf4 × very rare	0.245	-0.006	0.496	0.128	0.056	.
FP8 × very rare	0.021	-0.231	0.272	0.128	0.873	
GPTQ-int4 × very rare	0.126	-0.125	0.377	0.128	0.327	
GPTQ-int8 × very rare	-0.016	-0.267	0.236	0.128	0.903	
RANDOM EFFECTS						
σ^2 (Residual)	0.677					
τ_{00} (Model)	0.118					
τ_{00} (Topic)	0.477					
ICC (Model)	0.093					
ICC (Topic)	0.375					
N (Model)	5					
N (Topic)	100					
Observations	3000					
R^2 (marginal)	0.586					
R^2 (conditional)	0.780					

Table 23: **FACTSCORE**: Summary of linear mixed-effects model with **quantization** and **category** as predictors of **veriscore**: lmer(log_veriscore ~ Quantization * cat + (1|Model) + (1|topic), data = data). The quantization method "BF16" and category "very freq" were set as reference levels. Random intercepts were included for Model and topic to account for grouping variability. See Table 24 for post-hoc analysis.

CATEGORY	CONTRAST	ESTIMATE	SE	OR	p-value	
medium	BF16 - (AWQ-int4)	0.142	0.091	1.153	1.000	
medium	BF16 - (BNB-nf4)	0.430	0.091	1.537	0.000	***
medium	BF16 - FP8	-0.089	0.091	0.915	1.000	
medium	BF16 - (GPTQ-int4)	0.197	0.091	1.217	0.451	
medium	BF16 - (GPTQ-int8)	-0.023	0.091	0.977	1.000	
medium	(AWQ-int4) - (BNB-nf4)	0.287	0.091	1.333	0.023	*
medium	(AWQ-int4) - FP8	-0.231	0.091	0.794	0.161	
medium	(AWQ-int4) - (GPTQ-int4)	0.054	0.091	1.056	1.000	
medium	(AWQ-int4) - (GPTQ-int8)	-0.165	0.091	0.848	1.000	
medium	(BNB-nf4) - FP8	-0.519	0.091	0.595	0.000	***
medium	(BNB-nf4) - (GPTQ-int4)	-0.233	0.091	0.792	0.153	
medium	(BNB-nf4) - (GPTQ-int8)	-0.452	0.091	0.636	0.000	***
medium	FP8 - (GPTQ-int4)	0.286	0.091	1.331	0.025	*
medium	FP8 - (GPTQ-int8)	0.066	0.091	1.069	1.000	
medium	(GPTQ-int4) - (GPTQ-int8)	-0.219	0.091	0.803	0.234	
very freq	BF16 - (AWQ-int4)	0.129	0.089	1.137	1.000	
very freq	BF16 - (BNB-nf4)	0.075	0.089	1.078	1.000	
very freq	BF16 - FP8	-0.025	0.089	0.975	1.000	
very freq	BF16 - (GPTQ-int4)	0.113	0.089	1.120	1.000	
very freq	BF16 - (GPTQ-int8)	0.037	0.089	1.038	1.000	
very freq	(AWQ-int4) - (BNB-nf4)	-0.053	0.089	0.948	1.000	
very freq	(AWQ-int4) - FP8	-0.153	0.089	0.858	1.000	
very freq	(AWQ-int4) - (GPTQ-int4)	-0.016	0.089	0.985	1.000	
very freq	(AWQ-int4) - (GPTQ-int8)	-0.091	0.089	0.913	1.000	
very freq	(BNB-nf4) - FP8	-0.100	0.089	0.905	1.000	
very freq	(BNB-nf4) - (GPTQ-int4)	0.038	0.089	1.038	1.000	
very freq	(BNB-nf4) - (GPTQ-int8)	-0.038	0.089	0.963	1.000	
very freq	FP8 - (GPTQ-int4)	0.138	0.089	1.148	1.000	
very freq	FP8 - (GPTQ-int8)	0.062	0.089	1.064	1.000	
very freq	(GPTQ-int4) - (GPTQ-int8)	-0.076	0.089	0.927	1.000	
very rare	BF16 - (AWQ-int4)	-0.035	0.091	0.966	1.000	
very rare	BF16 - (BNB-nf4)	0.185	0.091	1.203	0.621	
very rare	BF16 - FP8	-0.110	0.091	0.896	1.000	
very rare	BF16 - (GPTQ-int4)	0.071	0.091	1.073	1.000	
very rare	BF16 - (GPTQ-int8)	-0.007	0.091	0.993	1.000	
very rare	(AWQ-int4) - (BNB-nf4)	0.220	0.091	1.246	0.229	
very rare	(AWQ-int4) - FP8	-0.075	0.091	0.928	1.000	
very rare	(AWQ-int4) - (GPTQ-int4)	0.106	0.091	1.112	1.000	
very rare	(AWQ-int4) - (GPTQ-int8)	0.028	0.091	1.029	1.000	
very rare	(BNB-nf4) - FP8	-0.294	0.091	0.745	0.018	*
very rare	(BNB-nf4) - (GPTQ-int4)	-0.114	0.091	0.892	1.000	
very rare	(BNB-nf4) - (GPTQ-int8)	-0.192	0.091	0.826	0.514	
very rare	FP8 - (GPTQ-int4)	0.180	0.091	1.197	0.697	
very rare	FP8 - (GPTQ-int8)	0.103	0.091	1.109	1.000	
very rare	(GPTQ-int4) - (GPTQ-int8)	-0.078	0.091	0.925	1.000	

Table 24: **FACTSCORE**: Post-hoc comparisons between quantization methods for **VERIScore** across **category** using Bonferroni adjustments for multiple pairwise comparisons. The odds ratios are derived from linear mixed-effects estimates exponentiated for interpretability.

glmer(cbind(Success, Failures) ~ Quantization + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	1.573	4.822	3.473	6.694	0.000	***
AWQ-int4	0.022	1.022	0.966	1.080	0.450	
BNB-nf4	−0.085	0.918	0.869	0.970	0.002	**
FP8	−0.152	0.859	0.814	0.907	0.000	***
GPTQ-int4	0.051	1.052	0.995	1.113	0.075	.
GPTQ-int8	−0.016	0.984	0.931	1.040	0.571	
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	0.138					
ICC (Model)	0.121					
N (Model)	5					
Observations	4500					
R^2 (marginal)	0.016					
R^2 (conditional)	0.500					

Table 25: **CS4**: Summary of generalized linear mixed model with **quantization** as the predictor of **accuracy**: `glmer(cbind(Success, Failures) ~ Quantization + (1|Model), data = data, family = binomial)`. The quantization method "BF16" was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. Refer to Table 26 for post-hoc analysis.

CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value	
BF16 - (AWQ-int4)	−0.022	0.028	0.979	0.495	1.000	
BF16 - (BNB-nf4)	0.085	0.028	1.089	0.521	0.035	*
BF16 - FP8	0.152	0.028	1.164	0.538	0.000	***
BF16 - (GPTQ-int4)	−0.051	0.029	0.950	0.487	1.000	
BF16 - (GPTQ-int8)	0.016	0.028	1.016	0.504	1.000	
(AWQ-int4) - (BNB-nf4)	0.107	0.028	1.113	0.527	0.002	**
(AWQ-int4) - FP8	0.173	0.028	1.189	0.543	0.000	***
(AWQ-int4) - (GPTQ-int4)	−0.029	0.029	0.971	0.493	1.000	
(AWQ-int4) - (GPTQ-int8)	0.038	0.028	1.038	0.509	1.000	
(BNB-nf4) - FP8	0.067	0.027	1.069	0.517	0.222	
(BNB-nf4) - (GPTQ-int4)	−0.136	0.028	0.873	0.466	0.000	***
(BNB-nf4) - (GPTQ-int8)	−0.069	0.028	0.933	0.483	0.201	
FP8 - (GPTQ-int4)	−0.203	0.028	0.816	0.449	0.000	***
FP8 - (GPTQ-int8)	−0.136	0.028	0.873	0.466	0.000	***
(GPTQ-int4) - (GPTQ-int8)	0.067	0.029	1.069	0.517	0.286	

Table 26: **CS4**: Post-hoc comparisons between quantization methods for **accuracy** (Table 25) using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).

glmer(cbind(Success, Failures) ~ Quantization * Input_constraints + (1 Model), data = data, family = binomial)						
PREDICTORS	ESTIMATE	OR	CI (LOWER)	CI (UPPER)	p-value	
Intercept	2.546	12.762	8.774	18.562	0.000	***
AWQ-int4	0.076	1.079	0.836	1.391	0.560	
BNB-nf4	−0.116	0.890	0.697	1.137	0.350	
FP8	−0.032	0.968	0.755	1.242	0.800	
GPTQ-int4	0.204	1.226	0.944	1.593	0.126	
GPTQ-int8	−0.102	0.903	0.707	1.154	0.417	
23 constraints	−0.758	0.469	0.387	0.568	0.000	***
39 constraints	−1.201	0.301	0.251	0.362	0.000	***
AWQ-int4:23 constraints	−0.085	0.918	0.698	1.208	0.542	
BNB-nf4:23 constraints	0.098	1.103	0.846	1.439	0.467	
FP8:23 constraints	−0.171	0.843	0.645	1.102	0.212	
GPTQ-int4:23 constraints	−0.180	0.835	0.630	1.107	0.210	
GPTQ-int8:23 constraints	0.148	1.160	0.889	1.514	0.275	
AWQ-int4:39 constraints	−0.044	0.957	0.736	1.246	0.746	
BNB-nf4:39 constraints	0.002	1.002	0.778	1.292	0.987	
FP8:39 constraints	−0.109	0.897	0.693	1.160	0.407	
GPTQ-int4:39 constraints	−0.151	0.860	0.656	1.127	0.274	
GPTQ-int8:39 constraints	0.064	1.067	0.827	1.376	0.620	
RANDOM EFFECTS						
σ^2 (residual variance)	1.000					
τ_{00} (Model)	0.142					
ICC (Model)	0.124					
N (Model)	5					
Observations	4500					
R^2 (marginal)	0.487					
R^2 (conditional)	0.743					

Table 27: **CS4**: Summary of generalized linear mixed model with **quantization**, **input constraint level**, and their interaction as predictors of **accuracy**: glmer(cbind(Success, Failures) ~ Quantization * Input_constraints + (1|Model), data = data, family = binomial). The quantization method "BF16" at 7 constraints was set as the reference level (intercept), with Model set as the grouping factor for a random intercept to account for variability across models. See Table 28 for post-hoc analysis.

CONSTRAINTS	CONTRAST	ESTIMATE	SE	OR	PROBABILITY	p-value
7	BF16 - (AWQ-int4)	-0.076	0.130	0.927	0.481	1.000
7	BF16 - (BNB-nf4)	0.116	0.125	1.123	0.529	1.000
7	BF16 - FP8	0.032	0.127	1.033	0.508	1.000
7	BF16 - (GPTQ-int4)	-0.204	0.133	0.815	0.449	1.000
7	BF16 - (GPTQ-int8)	0.102	0.125	1.107	0.525	1.000
7	(AWQ-int4) - (BNB-nf4)	0.192	0.127	1.212	0.548	1.000
7	(AWQ-int4) - FP8	0.108	0.129	1.114	0.527	1.000
7	(AWQ-int4) - (GPTQ-int4)	-0.128	0.136	0.880	0.468	1.000
7	(AWQ-int4) - (GPTQ-int8)	0.177	0.127	1.194	0.544	1.000
7	(BNB-nf4) - FP8	-0.084	0.124	0.919	0.479	1.000
7	(BNB-nf4) - (GPTQ-int4)	-0.320	0.131	0.726	0.421	0.213
7	(BNB-nf4) - (GPTQ-int8)	-0.015	0.122	0.985	0.496	1.000
7	FP8 - (GPTQ-int4)	-0.236	0.133	0.790	0.441	1.000
7	FP8 - (GPTQ-int8)	0.069	0.124	1.072	0.517	1.000
7	(GPTQ-int4) - (GPTQ-int8)	0.305	0.131	1.357	0.576	0.296
23	BF16 - (AWQ-int4)	0.010	0.053	1.010	0.502	1.000
23	BF16 - (BNB-nf4)	0.018	0.053	1.018	0.504	1.000
23	BF16 - FP8	0.203	0.051	1.225	0.551	0.001 **
23	BF16 - (GPTQ-int4)	-0.024	0.053	0.976	0.494	1.000
23	BF16 - (GPTQ-int8)	-0.047	0.053	0.954	0.488	1.000
23	(AWQ-int4) - (BNB-nf4)	0.008	0.052	1.008	0.502	1.000
23	(AWQ-int4) - FP8	0.193	0.051	1.213	0.548	0.002 **
23	(AWQ-int4) - (GPTQ-int4)	-0.034	0.053	0.967	0.492	1.000
23	(AWQ-int4) - (GPTQ-int8)	-0.056	0.053	0.945	0.486	1.000
23	(BNB-nf4) - FP8	0.185	0.051	1.203	0.546	0.004 **
23	(BNB-nf4) - (GPTQ-int4)	-0.042	0.053	0.959	0.490	1.000
23	(BNB-nf4) - (GPTQ-int8)	-0.065	0.053	0.937	0.484	1.000
23	FP8 - (GPTQ-int4)	-0.227	0.051	0.797	0.444	0.000 ***
23	FP8 - (GPTQ-int8)	-0.249	0.052	0.779	0.438	0.000 ***
23	(GPTQ-int4) - (GPTQ-int8)	-0.023	0.053	0.977	0.494	1.000
39	BF16 - (AWQ-int4)	-0.032	0.035	0.968	0.492	1.000
39	BF16 - (BNB-nf4)	0.114	0.035	1.121	0.529	0.015 *
39	BF16 - FP8	0.141	0.035	1.152	0.535	0.001 **
39	BF16 - (GPTQ-int4)	-0.053	0.036	0.948	0.487	1.000
39	BF16 - (GPTQ-int8)	0.037	0.035	1.038	0.509	1.000
39	(AWQ-int4) - (BNB-nf4)	0.146	0.035	1.158	0.537	0.000 ***
39	(AWQ-int4) - FP8	0.173	0.035	1.189	0.543	0.000 ***
39	(AWQ-int4) - (GPTQ-int4)	-0.021	0.036	0.979	0.495	1.000
39	(AWQ-int4) - (GPTQ-int8)	0.069	0.035	1.072	0.517	0.752
39	(BNB-nf4) - FP8	0.027	0.034	1.027	0.507	1.000
39	(BNB-nf4) - (GPTQ-int4)	-0.167	0.035	0.846	0.458	0.000 ***
39	(BNB-nf4) - (GPTQ-int8)	-0.077	0.035	0.926	0.481	0.384
39	FP8 - (GPTQ-int4)	-0.194	0.035	0.824	0.452	0.000 ***
39	FP8 - (GPTQ-int8)	-0.104	0.035	0.901	0.474	0.038 *
39	(GPTQ-int4) - (GPTQ-int8)	0.090	0.035	1.094	0.522	0.167

Table 28: **CS4**: Post-hoc comparisons between quantization methods for **accuracy** (Table 27) across constraint levels using Bonferroni adjustments for multiple pairwise comparisons. The probability values refer to the likelihood that the quantization method listed first in each contrast is more accurate than the one listed second (i.e., a value of 0.5 suggests that both methods are comparable in terms of accuracy).