Cover Letter

Dear area chair and reviewers,

In this document, we explain the main changes that we made in the present resubmission. We address the comments in the order they were presented in the system.

We use black color for the reviewer comments, blue color for our replies.

Best regards,
The authors

Summary of our additions:

1. We have added three additional models and datasets in our analysis to improve generalization and better discussion in our findings.
2. We have better motivated our research objectives in the introduction.
3. We improved the discussion on calibration analysis by examining whether model size or activation function choice contributes to the observed effects, using comparisons between models with similar architectures and activation functions.
4. We expanded the related work section to include additional papers addressing interpretation and quantization.

Meta Review (Area Chair csx5):

Comment MR1: The motivation of the paper should be elaborated. While interpreting the effects of quantization on LLMs is important, the authors have not explained why they chose to investigate the specific four research questions raised in the paper. Furthermore, the four research questions appear to be unrelated to each other, which may limit the coherence of the study. As a result, although the individual findings may be of interest to the community, the overall contribution could be more clearly defined.

Ans: We have improved the motivation of our work in the introduction. The common theme across our research questions is the analysis of model and neurons behavior under quantization. Since the objective is to establish a foundational understanding that can motivate further research, we began by formulating broad questions that examine the model from multiple perspectives. While these questions may initially appear disjointed, each is intended to highlight a specific foundational aspect that can serve as a springboard for deeper investigation in its respective area.

Comment MR2: The study is currently limited to two small models and two datasets, which raises concerns about the generalizability of the findings. The authors are encouraged to conduct experiments on a wider range of models and datasets to strengthen their conclusions. Additionally, providing an in-depth analysis, such as explaining why 4-bit quantization improves calibration, could enhance the overall contribution of the paper.

Ans: We extended our analysis by including three additional models: Qwen 3B and Qwen 7B from the same model family to assess intra-family effects, and Mistral 7B, which matches

LLaMA 2 7B and Qwen 7B in parameter count and activation function. We also incorporated three new datasets: PIQA, HellaSwag, and a sentiment task to broaden task diversity.
It is also important to note that adding each new model entails analysis under three additional quantization settings, making the experiments increasingly computationally expensive with each added model or task.
Our findings indicate that although quantization results vary, the differences are generally not significant enough to undermine its practicality for model compression.
The inclusion of models with similar architectures and activation functions revealed that calibration behavior can differ across tasks, likely due to differences in pretraining objectives, and does not exhibit consistent patterns across tasks.

Reviewer DPfK (R1)

Comment (DPfK) 1: The findings in this paper are too fragmented.
Ans: We have better motivated our research objectives in the introduction. In this research, our aim is to include a broad range of diverse interpretation techniques to explore the effects of quantization on model interpretability from multiple perspectives, which can provide a foundation for further research into specific topics to analyze in detail.

Comment (DPfK) 2: The authors should focus on some points to discuss more about 'what we can do with' their findings instead of just 'what we can know'.
Ans: Given the positive results, our work suggests that quantization minimally impacts learned knowledge of models that provides strong evidence that quantized models as viable and reliable solutions for resource-constraint environment.

Comment (DPfK) 3: In related work, the authors should tell us what's the difference between this paper and existing quantization analysis.
Ans: We have expanded related work to include research work exploring quantization and interpretability. Existing quantization analyses mainly evaluate the effect of quantization on a model's end-to-end task performance. A few works that interpret quantize models only interpret a specific part such as confidence and calibration of model but uses different models and quantization techniques to make direct comparison possible. To the best of our knowledge, our work presents the first comprehensive analysis on LLMs measuring effect on model's internal representation using various interpretability techniques, that provides evidence on reliability of quantized models.

Comment (DPfK) 4: The Conclusion takes too much space which can be used to give more results and analysis.
Ans: We made the suggested change.

Comment (DPfK) 5: More models should be tested to provide more regular results, and more explanations should be given in sections such as 4.2.2 to analyze why the two models behave differently.
Ans: We have added 3 additional models and 3 additional datasets.

We have extended the discussion on effect on calibration as we have added additional models sharing similar architecture or model family

Reviewer H9To (R2)

Comment (H9To) 1: The study is restricted to two relatively small models, Phi-2 (2.7B) and Llama-2-7b. This raises questions about the generalizability of the findings to larger models, such as LLaMa 3 70B Dense or Deepseek v3 671B MoE, which may exhibit different behaviors under quantization.

Ans: We have incorporated three additional models and datasets to enhance the generalizability of our analysis. Our experiments involve computationally intensive techniques such as attribution analysis and neuron redundancy estimation. Since each additional model or task must be evaluated under three different quantization settings, scaling to larger LMs like LLaMA 3 70B (Dense) or DeepSeek V3 671B (MoE) is currently infeasible.

Comment (H9To) 2: Narrow Dataset and Layer Analysis: Only two datasets were analyzed, limiting the breadth of the conclusions. The findings may not apply to diverse tasks (e.g., generative reasoning, code, math).

Ans: We incorporated three additional datasets spanning multiple tasks. However, due to the computational cost of experiments such as attribution analysis, where longer output sequences significantly increase complexity, we selected tasks that primarily require single token predictions.

Comment (H9To) 3: The paper does not deeply explore why 4-bit quantization improves calibration or how architectural differences mediate quantization effects. This leaves a gap in understanding the underlying mechanisms.

Ans: We have expanded the discussion on the impact of quantization on calibration. Overall, the effect is not substantial enough to discourage the use of quantization as a practical model compression technique. Its impact varies depending on the model and task, with calibration sometimes showing a slight reduction and, in some cases, even improving compared to the full precision model.

Reviewer obpp (R3)

Comment (obpp) 1: The main weaknesses are the coverage of datasets (only 2) and models (only 2). This is unfortunately too little to expect generalization of the findings – especially as the results are along the lines that results are not affected by quantization. The authors do acknowledge this weakness in limitations and phrase their claims carefully, with that weakness taken into account. The paper does not present a methodological contribution but uses existing interpretability tools to answer an original research question.

Ans: We have added three additional models and three datasets in our analysis to improve the generalizability of our findings

Comment (obpp) 2: Table styling (vert./horizontal lines) could be harmonized.

Ans: We have made the suggested change.

Comment (obpp) 3: Related work could be expanded by more techniques other analysis touching quantization and interpretability, e.g., Q-SENN (Sade & Soriano, 2024)
Ans: We have added more research work exploring quantization and interpretation. However, direct comparison with existing work is not feasible, as similar questions have not been widely explored by others, or the analyses have been conducted using different models or quantization methods.