SIIM
CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Benchmarking Quantization: A Comprehensive Comparison of Open-Source Large Language Models

Blake T. Passe, Mayo Clinic; Sanaz Vahdati, MD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction/Background

Large Language Models (LLMs) have seen a boom within the Artificial Intelligence community recently. As the parameter size of models has grown from millions to trillions, computational requirements present substantial deployment challenges. A potential approach to resolve these pressing difficulties is quantizing open-source LLMs - reducing the precision of model parameters - while aiming to preserve performance. In the current work, our aim is to compare how quantization of open-source LLMs impacts information extraction from radiology reports, latency, and computational demands.

Methods/Intervention

622 radiology reports were obtained in five categories: cervical spine fractures, glioma progression, liver metastases, pneumonia, and pulmonary embolism. Each glioma progression report was labeled “Improved”, “Progression”, “Stable”, “Pseudoprogression”, or “Pseudoresponse”, while the four remaining categories were labeled a binary “Yes" or “No”. Different ‘instruct’ versions of Llama3 and Phi-3 models were applied using Ollama and allotted one NVIDIA A100 80Gb GPU for inference. Prompting was conducted by a radiology artificial intelligence expert to describe the model’s task and criteria succinctly. The prompting structure consisted of four sequential steps: identity establishment, cognitive framework setup, report presentation, and contextual clarification. Finally, the JSON output, RAM usage, and latency for the extraction process was recorded for each model at each quantization level.

Results/Outcome

Findings indicate model size displays a positive correlation with both RAM and latency during inference (Fig. 1). Comparable accuracy (>92%) was observed between the 8, 5, and 4-bit quantized versions of Llama3:70b, Llama3:8b, and Phi3:14b (Table 1). This is intriguing due to the large gap between model sizes. The extreme 2-bit quantization demonstrated a prominent confabulation of answers. As shown in Table 1 and Fig. 2, response divergence (i.e. responses not within the defined structure, such as “Maybe” rather than the required “Yes” or “No”) is displayed before the performance degradation of the models. This suggests that LLMs may lose output obedience before specific performance metrics decline.

Conclusion

Our findings indicate that models can perform quite well even with substantial quantization for question-answering tasks applied to radiology reports.

Statement of Impact

Navigating the tradeoff between quantization and quality is largely unstudied in medicine, but it indicates significant potential to reduce computation load.

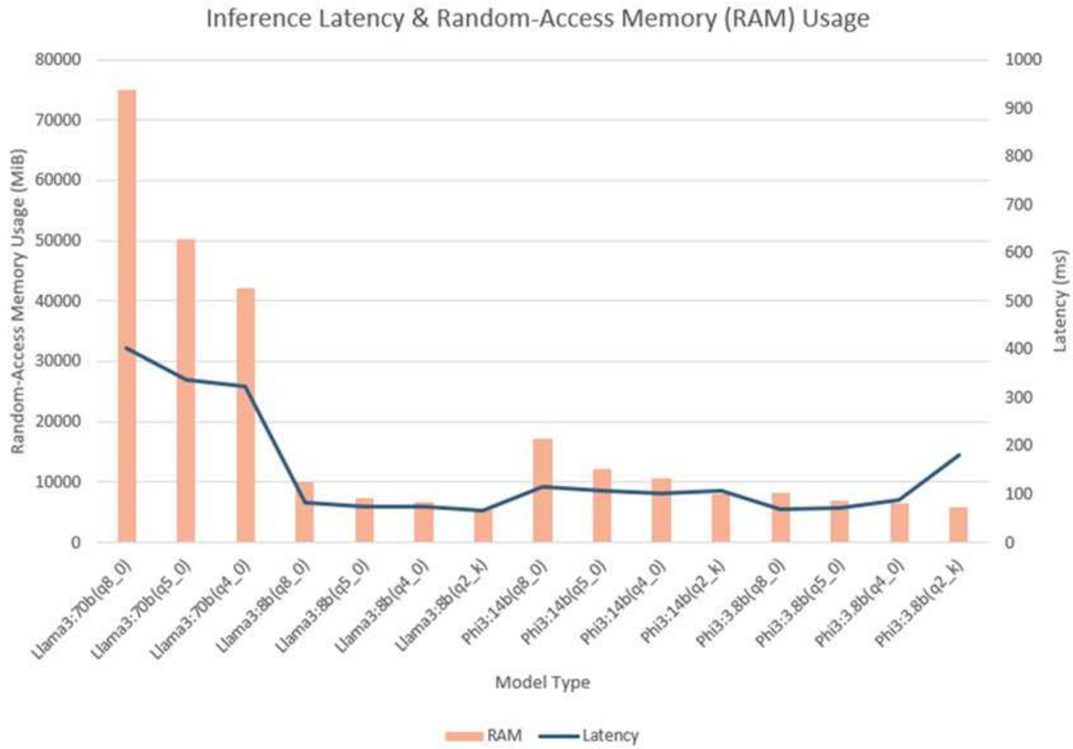
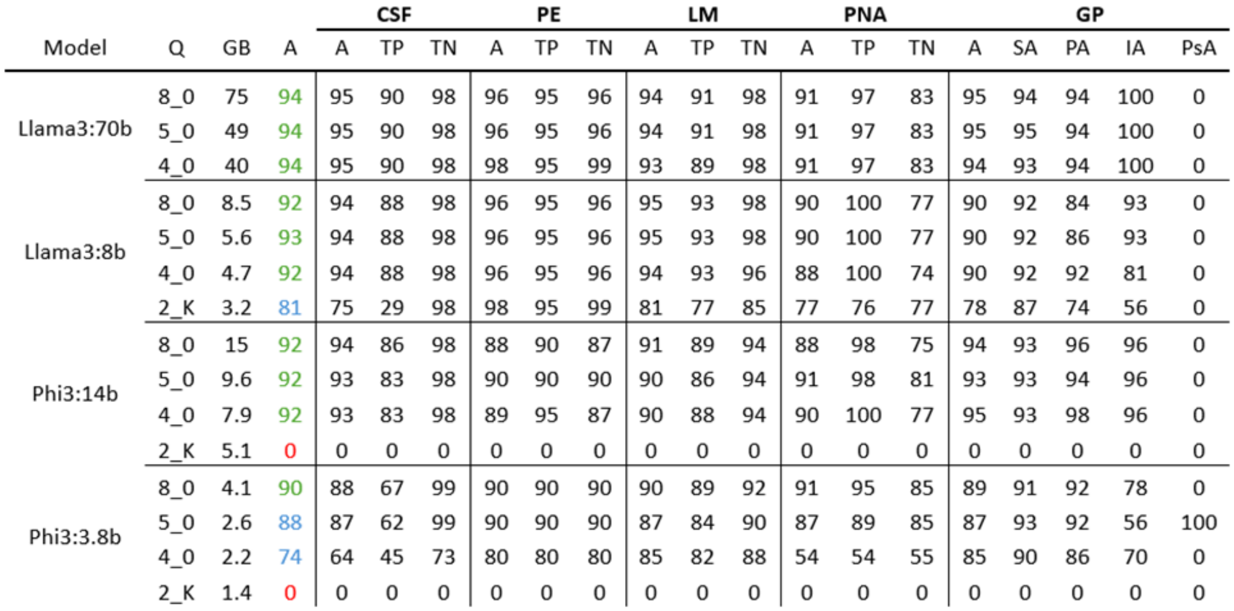


Fig. 1: Dual-axis chart plotting RAM usage and latency during inference by model type.



				CSF			PE			LM			PNA			GP				
Model	Q	GB	A	A	TP	TN	A	TP	TN	A	TP	TN	A	TP	TN	A	SA	PA	IA	PsA
Llama3:70b	8_0	75	94	95	90	98	96	95	96	94	91	98	91	97	83	95	94	94	100	0
	5_0	49	94	95	90	98	96	95	96	94	91	98	91	97	83	95	95	94	100	0
	4_0	40	94	95	90	98	98	95	99	93	89	98	91	97	83	94	93	94	100	0
Llama3:8b	8_0	8.5	92	94	88	98	96	95	96	95	93	98	90	100	77	90	92	84	93	0
	5_0	5.6	93	94	88	98	96	95	96	95	93	98	90	100	77	90	92	86	93	0
	4_0	4.7	92	94	88	98	96	95	96	94	93	96	88	100	74	90	92	92	81	0
	2_K	3.2	81	75	29	98	98	95	99	81	77	85	77	76	77	78	87	74	56	0
Phi3:14b	8_0	15	92	94	86	98	88	90	87	91	89	94	88	98	75	94	93	96	96	0
	5_0	9.6	92	93	83	98	90	90	90	90	86	94	91	98	81	93	93	94	96	0
	4_0	7.9	92	93	83	98	89	95	87	90	88	94	90	100	77	95	93	98	96	0
	2_K	5.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Phi3:3.8b	8_0	4.1	90	88	67	99	90	90	90	90	89	92	91	95	85	89	91	92	78	0
	5_0	2.6	88	87	62	99	90	90	90	87	84	90	87	89	85	87	93	92	56	100
	4_0	2.2	74	64	45	73	80	80	80	85	82	88	54	54	55	85	90	86	70	0
	2_K	1.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1: Table presents general model statistics as well as performance on all five categories for each model tested. Both Phi-3 models shown are the 4k context window versions. [Q, 'Bit Quantization of Model'], [GB, 'Gb Size of Model'], [A, 'Overall Accuracy on All Categories' or 'Overall Accuracy on Specified Category'], [TP, 'True Positives Accuracy'], [TN, 'True Negatives Accuracy'], [SA, 'Accuracy of Ground Truth Stables Correctly Labeled'], [PA, 'Accuracy of Ground Truth Progressions Correctly Labeled'], [IA, 'Accuracy of Ground Truth Improved Correctly Labeled'], [PsA, 'Accuracy of Ground Truth Pseudoprogressions Correctly Labeled']. There is no column for "Pseudoresponse" as there existed no cases in the dataset. Green color denotes accuracy at or above 90%, blue color denotes accuracy at or above 70%, and red color denotes accuracy below 70%.

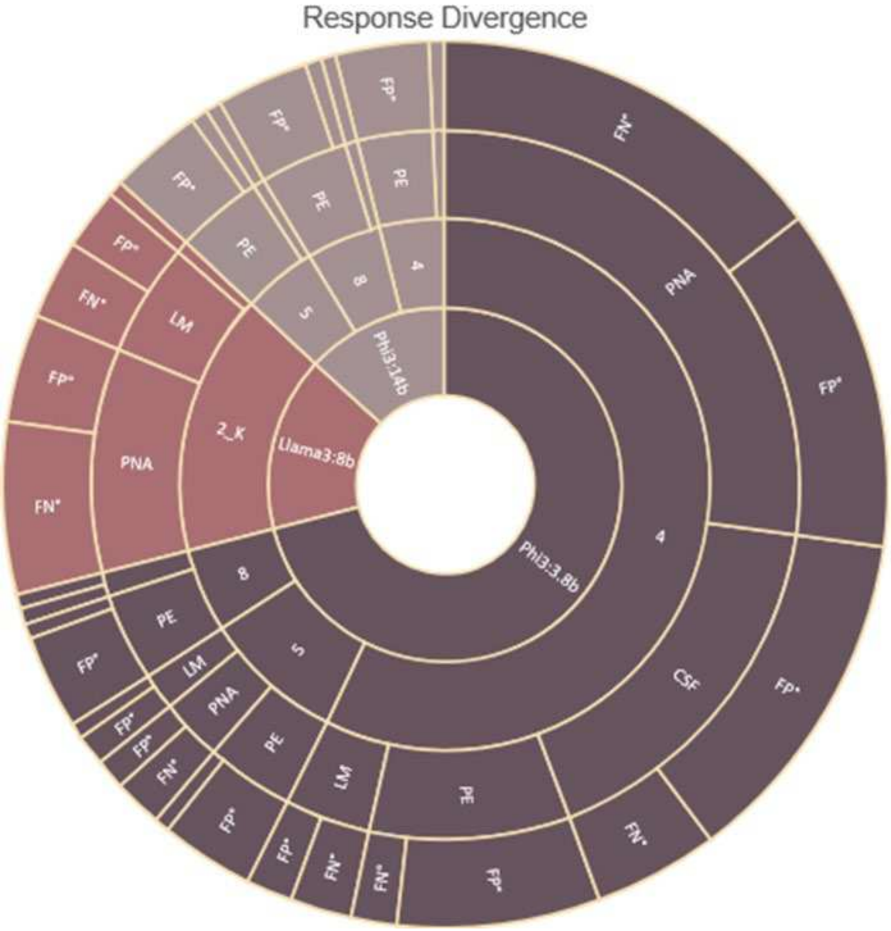


Fig. 2: Sunburst chart depicting response divergence by model type, quantization, category, and response type. FP° denotes a ground truth positive that is falsely labeled and not as “No.” Similarly, FN° denotes a ground truth negative that is falsely labeled and not as “Yes”. Llama3:80b models are not shown due to complete in-frame responses. The 2_K quantized versions of both Phi-3 models tested were excluded due to their complete inability to perform the desired task.

Keywords

Artificial Intelligence; Large Language Model; Quantization; Radiology Report Data Extraction