ISSN: 2320-5407

International Journal of Advanced Research

Publisher's Name: Jana Publication and Research LLP

www.journalijar.com

ISSN NO. 2320-5407

REVIEWER'S REPORT

Manuscript No.: 50699 Date: 19-03-2025

Title:

Optimizing LLaMA 3.2 1B Using Quantization Techniques using Bitsandbytesfor Efficient AI Deployment

Recommendation:
Accept … **Yes**
Accept after minor revision … …
Do not accept (*Reasons below*) …

Rating	Excel.	Good	Fair	Poor
Originality			**YES**	
Techn. Quality			**YES**	
Clarity			**YES**	
Significance		**YES**		

Reviewer Name: Gulnawaz Gani

Reviewer's Comment for Publication

The paper contributes by demonstrating how post-training quantization using the 'Bitsandbytes' library optimizes LLaMA 3.2-1B for efficient deployment on resource-constrained hardware.

Detailed Reviewer's Report

The paper explores post-training quantization techniques for optimizing LLaMA 3.2-1B using the 'Bitsandbytes' library, demonstrating significant improvements in memory efficiency and computational performance.
It provides a thorough comparison of BF16, FP16, INT8, and INT4, highlighting trade-offs between accuracy and resource utilization. However, the study lacks real-world deployment case studies, making practical implications less evident.
While the performance benchmarks are detailed, a deeper analysis of application-specific impacts would enhance relevance. Additionally, the study primarily focuses on synthetic benchmarks, limiting insights into real-world use cases. Future work could explore hybrid quantization strategies for better optimization.
Despite these limitations, the paper provides valuable insights into deploying LLMs efficiently on resource-constrained hardware.

Decision:
Accept