ISSN: 2320-5407

International Journal of Advanced Research

Publisher's Name: Jana Publication and Research LLP

www.journalijar.com

REVIEWER'S REPORT

Manuscript No.: IJAR-50699 Date: 18-03-2025

Title: Optimizing LLaMA 3.2 1B Using Quantization Techniques using Bitsandbytes for Efficient AI Deployment

Recommendation:
Accept as it is……**YES……**
Accept after minor revision……
Accept after major revision ……
Do not accept (*Reasons below*) ……

Rating	Excel.	Good	Fair	Poor
Originality	√			
Techn. Quality		√		
Clarity		√		
Significance			√	

Reviewer's Name: Mr Bilal Mir

Reviewer's Decision about Paper: **Recommended for Publication.**

Comments (*Use additional pages, if required*)
.

Reviewer's Comment / Report

Abstract
The abstract provides a well-structured overview of the study, clearly outlining the problem statement, methodology, and key findings. The focus on the trade-offs between various precision settings (BF16, FP16, INT8, and INT4) effectively sets the stage for the research. The mention of different hardware platforms for accuracy and performance benchmarking strengthens the practical relevance of the study. The conclusion succinctly highlights the benefits of quantization in reducing memory usage while maintaining model accuracy, making it a valuable contribution to AI model optimization.

Keywords
The chosen keywords accurately represent the core themes of the paper, covering technical terms such as quantization techniques, inference efficiency, and post-training quantization. These terms enhance the discoverability of the research within the domain of AI optimization.

Introduction
The introduction is well-written, providing a comprehensive background on LLM advancements and their computational challenges. It effectively establishes the need for optimization, particularly through

ISSN: 2320-5407
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

REVIEWER'S REPORT

quantization techniques. The discussion of various precision formats (BF16, FP16, INT8, and INT4) is informative and relevant, offering insight into their respective advantages and trade-offs. The explanation of post-training quantization (PTQ) and its benefits ensures clarity for readers unfamiliar with the technique. The mention of 'Bitsandbytes' as a key tool adds practical significance to the study. Furthermore, the gap in existing research is clearly identified, highlighting the novelty of this investigation.

Overall Assessment

The manuscript presents a well-structured and in-depth analysis of quantization techniques for optimizing LLaMA 3.2 1B. The research is well-motivated, addressing both theoretical and practical aspects of quantization for AI model deployment. The study's design, including controlled experiments on different hardware environments, enhances the reliability of the findings. The writing is clear and precise, ensuring accessibility to a broad audience in the AI and machine learning communities.