Instruction-Tuned LLaMA 3.2 Variants and Base Models in Zero-Shot Defect Detection Under INT8 Quantization

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do instruction-tuned variants of LLaMA 3.2 compare to base models in zero-shot defect identification robustness under INT8 quantization constraints. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Shadow-FT: Tuning Instruct Model via Training on Paired Base Model. Research question: How do instruction-tuned variants of LLaMA 3.2 compare to base models in zero-shot defect identification robustness under INT8 quantization constraints?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

15 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
BAAI-2k dataset was built by extracting 2000 samples from BAAI-Infinity-Instruct Dataset.	×	0.03
Samples with high rewards were selected to ensure data quality and uniform sampling among all categories for data divers	×	0.07
Qwen 3 series, Llama 3 series, Gemma-3 series, Yi series, and Falcon series were tuned using Shadow-FT.	×	0.11
LLaMA-Factory was employed for the code base and two tuning strategies: full-parameter and LoRA were applied.	×	0.07
All experiments were conducted on 8 A100 GPUs.	×	0.03
OpenCompass framework and lmdeploy were used as the acceleration framework for evaluation.	×	0.03
During inference, the cutoff length was set as 4096 and the batch size as 512.	×	0.01
Math-7, Code-3, and Knowledge-9 were used to evaluate mathematical, coding, and commonsense reasoning abilities respecti	×	0.04
Math-7 includes AIME24 MAA, GSM8K, MATH, MATH-500, Minerva_Math, SVAMP.	×	0.01
Code-3 includes HumanEval, HumanEval+, LiveCodeBench.	×	0.02
Knowledge-9 includes ARC-challenge, BBH, DROP, GPQA, MMLU, MMLU Pro, Winogrande, TheoremQA.	×	0.05
Evaluations were mainly conducted under a zero-shot setting.	×	0.09
For Qwen-3 series, enable_thinking was set as false for universal evaluations.	×	0.02
Shadow-FT introduces no extra training costs compared to traditional tuning methods.	×	0.13
Shadow-FT tunes the BASE model first and then grafts the weight updates to the INSTRUCT model.	✓	0.23
Traditional tuning methods rely on the INSTRUCT model while Shadow-FT relies on the BASE model.	✓	0.19
The weight updates from the BASE model are believed to be more suitable for modeling knowledge with less priority.	×	0.09
Task Vectors aim to represent the ability on tasks as vectors and are used for arithmetic operations on these tasks.	×	0.03

References

- http://arxiv.org/abs/2304.03277v1
- http://arxiv.org/abs/2304.15010v1
- http://arxiv.org/abs/2505.12716v3