Domain-Specific Adaptations (E.G., Biomedical, Legal) Of Quantized Slms Under 10B Parameters Impact Their Performance

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do domain-specific adaptations (e.g., biomedical, legal) of quantized SLMs under 10B parameters impact their performance on SLM-Bench's 14 domains compared to general-purpose SLMs. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models. Research question: How do domain-specific adaptations (e.g., biomedical, legal) of quantized SLMs under 10B parameters impact their performance on SLM-Bench's 14 domains compared to general-purpose SLMs?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

4 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
ESFT surpasses LoRA significantly and is competitive with FFT in customization ability evaluation.	×	0.02
ESFT-Token and ESFT-Gate achieve near-best results in model enhancement tasks like Math.	×	0.03
ESFT-Gate achieves the best performance in the Humaneval task.	×	0.04
ESFT-Gate achieves near-best performance in 3 out of 4 model adaptation tasks.	×	0.04
ESFT-Gate's average score is 50.2, competitive compared to FFT's 51.0, slightly better than ESFT-Token's 49.4, and signi	×	0.02
ESFT consistently outperforms FFT and LoRA by showing less performance degradation in general ability evaluation.	×	0.01
ESFT-Token performs better than ESFT-Gate, with average scores of 61.5 and 60.6, respectively.	×	0.01
ESFT retains effectiveness in maintaining general task performance, surpassing FFT's 58.8 and LoRA's 59.1.	×	0.02
The average training time for ESFT-Token and ESFT-Gate is 19.8 minutes and 20.9 minutes, respectively.	×	0.02
The FFT method takes significantly longer at 28.5 minutes.	×	0.03
LoRA achieves a shorter training time of 16.5 minutes.	×	0.02
The average storage space of parameters trained is 2.57 GB for ESFT-Token and 3.20 GB for ESFT-Gate.	×	0.02
FFT demands a substantial 28.6 GB of storage space.	×	0.00
Most tasks and layers train 5-15% of experts, demonstrating ESFT's effectiveness in selecting task-related experts.	×	0.07
ESFT demonstrates excellent performance in training time and storage space, significantly outperforming FFT.	×	0.02
FFT and LoRA exhibit even more severe degradation in Math and Code performance, while ESFT shows a minimal performance d	×	0.03
The routing distribution for a specific task tends to be highly concentrated, while the distribution of activated expert	✓	0.32
MoE models with finer-grained experts are more advantageous in selecting the combination of experts that are most releva	✓	0.39

References

- http://arxiv.org/abs/2508.17624v1
- http://arxiv.org/abs/2407.01906v2
- http://arxiv.org/abs/2507.23104v1