LoRA Orthogonality Effects on Codestral Inference Throughput in HumanEval

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the impact of imperfect orthogonality in LoRA-based fine-tuning on the inference throughput of Codestral when evaluated on the HumanEval benchmark. Pretraining Large Language Models (LLMs) on web-scale datasets becomes fundamental for advancing general-purpose AI. In contrast, enhancing their predictive performance on downstream tasks typically involves adapting their knowledge through fine-tuning. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Layer-wise LoRA fine-tuning: a similarity metric approach. Research question: What is the impact of imperfect orthogonality in LoRA-based fine-tuning on the inference throughput of Codestral when evaluated on the HumanEval benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

10 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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
The experiments utilize RoBERTabase and DeBERTa-v3base as encoder-only transformer architectures.	×	0.03
The experiments utilize LLaMA 2-7B, Mistral-7B-v0.1, and Gemma-7B as decoder-only transformer architectures.	×	0.03
The experiments utilize LLaVA-1.5-7B for evaluation in a multimodal context.	×	0.03
For RoBERTabase and DeBERTa-v3base on NLU tasks, the LoRA rank is set to 8 and alpha to 16.	×	0.03
For LLaMA 2-7B, Mistral-7B-v0.1, and Gemma-7B on NLG tasks, the LoRA rank and alpha are set to 128.	×	0.03
For the multimodal model LLaVA-1.5-7B, the LoRA rank and alpha are set to 128.	×	0.03
NLU evaluation is conducted on the GLUE benchmark using RoBERTabase and DeBERTa-v3base.	×	0.05
The method selects six layers (half of the architecture's layers) to fine-tune with LoRA for encoder-only models.	×	0.13
The first layer is excluded from fine-tuning for encoder-only models in this study.	×	0.09
Compared to regular LoRA, the proposed method achieves a 50% parameter reduction with DeBERTa-v3base.	×	0.07
With DeBERTa-v3base, the proposed method results in an average predictive performance drop of 0.27 percentage points com	×	0.08
The method by Lodha et al. (2023) results in an average predictive performance drop of 7.68 percentage points with DeBER	×	0.07
The method by Lodha et al. (2023) requires calculating the FIM score over training steps.	×	0.02
LLaMA 2-7B, Mistral-7B, and Gemma-7B were fine-tuned on Meta-MathQA and evaluated on GSM8K.	×	0.03
In the GSM8K evaluation, the proposed method increased accuracy compared to fine-tuning all layers with standard LoRA fo	×	0.05
LoRA adapts two matrices B and A such that the update ΔW = (α/r) * BA is added to the pre-trained weight matrix W0.	×	0.05

References

- http://arxiv.org/abs/2510.27364v1
- http://arxiv.org/abs/2501.19389v4
- http://arxiv.org/abs/2602.05988v1