DONOD Threshold Optimization for Efficient LLaMA-2-7B Instruction Fine-Tuning and Code Generation Performance

Assignee Research

May 29, 2026

Abstract

Ad-hoc instruction fine-tuning of large language models (LLMs) is widely adopted for domain-specific adaptation. While domain-specific supervised fine-tuning (SFT) is effective and efficient, it often weakens cross-domain generalization and struggles with noisy training data. To address these challenges, we propose DONOD, a lightweight model-intrinsic data pruning method. Our approach evaluates data using two model-parameter-based metrics: Delta of Norm (DON), which captures the cumulative influence on model weights, and Norm of Delta (NOD), which quantifies weight instability. Moreover, by

1 Introduction

This paper examines: DONOD: Efficient and Generalizable Instruction Fine-Tuning for LLMs via Model-Intrinsic Dataset Pruning. Research question: What is the effect of varying the DON threshold in DONOD on LLaMA-2-7B instruction fine-tuning efficiency (training tokens reduced) versus performance on code generation benchmarks (HumanEval, MBPP)?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

13 papers retrieved. 15 claims extracted; 3 independently verified. Quality review score: 6.0/10.

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
By filtering out 70% of the whole dataset, we improve target-domain accuracy by 14.90% and cross-domain accuracy by 5.67	✓	0.30
Data pruned by smaller models (e.g., Llama 3.1-8B) generalize effectively on larger models (e.g., Llama 2-13B).	✓	0.25
DONOD demonstrates comparable or superior performance while remaining dataset-agnostic, enabling broader applicability.	✓	0.26
The benchmark is designed to be comprehensive and domain-orthogonal, assessing abilities in logical reasoning, mathemati	×	0.05
The benchmark reflects real-world ad-hoc SFT scenarios, where the objective is to strengthen targeted model abilities ra	×	0.05
We assess DONOD across the following settings: SAT Math Chain-of-Thought (COT), LogiQA-Train, IFEval-like Data, GSM8K.	×	0.03
We evaluate DONOD on a diverse set of instruction-tuned models: LLaMA-3.2-3B-Instruct, LLaMA-3.1-8B-Instruct, LLaMA-2-13	×	0.15
DONOD consistently outperforms others while using significantly less data across nearly all benchmarks, e.g., 20-30%, an	×	0.04
DONOD achieves state-of-the-art performance in core reasoning tasks such as math and logic, outperforming full-data base	×	0.09
DONOD achieves lossless training acceleration with only 20% of the data, highlighting the scalability and robustness of	×	0.06
Traditional methods often rely on external models as quality judges or employ reward models to identify high-quality dat	×	0.04
Recent studies question the effectiveness of the paradigm relying on auxiliary models for dataset pruning.	×	0.09
Alternative approaches focus on intrinsic data metrics such as length, naturalness, and coherence, but there is no conse	×	0.05
Model-intrinsic methods leverage the model's training dynamics to bypass explicit metric definitions, enabling automated	×	0.08
Model-intrinsic pruning methods assume access to a target data distribution, often via validation or development sets.	×	0.07

References

- http://arxiv.org/abs/2504.14810v2
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2312.10793v3