SOVEREIGN: What is the impact of domain adaptation fine-tuning on LLaMA-2-7B model performance degradation across held-ou

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

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 em

1 Introduction

Analysis of: DONOD: Efficient and Generalizable Instruction Fine-Tuning for LLMs via Model-Intrinsic Dataset Pruning. Research goal: What is the impact of domain adaptation fine-tuning on LLaMA-2-7B model performance degradation across held-out test sets?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

2 papers retrieved. 15 claims extracted, 4 verified. Tribunal: 4.8/10 $\rightarrow$ REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
DONOD achieves superior fine-tuning efficiency and improved robustness against noisy data.	✓	0.29
By filtering out 70% of the whole dataset, DONOD improves target-domain accuracy by 14.90% and cross-domain accuracy by	✓	0.27
Data pruned by smaller models (e.g., Llama 3.1-8B) generalize effectively on larger models (e.g., Llama 2-13B).	✓	0.26
DONOD demonstrates comparable or superior performance while remaining dataset-agnostic, enabling broader applicability.	✓	0.25
Code will be made publicly available.	×	0.08
The evaluation benchmark is constructed based on AGIEval and IFEval.	×	0.01
The benchmark assesses abilities in logical reasoning, mathematics, reading comprehension, and instruction following.	×	0.02
DONOD is evaluated on LLaMA-3.2-3B-Instruct, LLaMA-3.1-8B-Instruct, LLaMA-2-13B-Chat, and Qwen 2.5-7B-Instruct.	×	0.06
DONOD consistently outperforms others while using significantly less data across nearly all benchmarks, e.g., 20-30%.	×	0.03
DONOD achieves state-of-the-art performance in core reasoning tasks such as math and logic.	×	0.05
DONOD enables lossless training acceleration with only 20% of the data.	×	0.04
Traditional methods for dataset pruning often rely on external models as quality judges or employ reward models to ident	×	0.06
Dependence on auxiliary models incurs significant computational costs and limits scalability.	×	0.03
Recent studies question the effectiveness of the paradigm relying on auxiliary models.	×	0.06
Model-intrinsic pruning methods leverage the model's training dynamics to bypass explicit metric definitions.	×	0.09

References

- https://www.semanticscholar.org/paper/14f2992c9c509300c279f5daf59606848f5ff3c7
- https://arxiv.org/abs/2504.14810