DONOD Pruning vs. Random Pruning and Full Fine-Tuning on LLaMA-2-7B Cross-Domain Benchmarks

Assignee Research

May 29, 2026

Abstract

Recent work by Zellers et al. (2018) introduced a new task of commonsense natural language inference: given an event description such as "A woman sits at a piano," a machine must select the most likely followup: "She sets her fingers on the keys." With the introduction of BERT, near human-level performance was reached. Does this mean that machines can perform human level commonsense inference? In this paper, we show that commonsense inference still proves difficult for even state-of-the-art models, by presenting HellaSwag, a new challenge dataset. Though its questions are trivial for humans

1 Introduction

This paper examines: HellaSwag: Can a Machine Really Finish Your Sentence?. Research question: How does the DONOD pruning method compare to random data pruning and full-dataset fine-tuning on LLaMA-2-7B for cross-domain held-out benchmarks (MMLU, HellaSwag, WinoGrande) in terms of accuracy retention?.

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 4.0/10.

3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
BERT outperforms the best known ELMo NLI model (ESIM+ELMo) given only 64 examples, while the ELMo model requires the ent	×	0.02
BERT needs upwards of 16k examples to approach human performance, around which it plateaus.	×	0.03
BERT's performance drops by 11.9 points (86.7% to 74.8%) when context is omitted (Ending Only).	×	0.01
BERT's performance reduces by less than 10% when the words in each ending choice are randomly permuted (Shuffled).	×	0.02
BERT's performance drops to 60.4% when the context is removed and the words in each ending are shuffled.	×	0.03
ELMo's performance is around 60% from Zellers et al., 2018.	×	0.10
Adversarial Filtering (AF) results with BERT-Large as the discriminator show that GPT converges at random, while the LM	×	0.06
AF applied to WikiHow generations from GPT converges to random, 40%, and 50% for one, two, and three sentences respectiv	×	0.04
BERT-Large achieves 46.7% accuracy on the validation set and 47.3% on the test set.	×	0.03
Human performance on the HellaSwag dataset is 95.7% on the validation set and 95.6% on the test set.	×	0.08

References

- http://arxiv.org/abs/2406.01574v6

- http://arxiv.org/abs/1905.07830v1
- http://arxiv.org/abs/2503.20786v1