Scaling Effects of Generative Pre-Training Models on GLUE Benchmark Performance

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does scaling the size of generative pre-training models influence the accuracy of self-supervised representations measured on the GLUE benchmark for natural language understanding. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Unified Language Model Pre-training for Natural Language Understanding and Generation. Research question: How does scaling the size of generative pre-training models influence the accuracy of self-supervised representations measured on the GLUE benchmark for natural language understanding?.

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 5.3/10.

3 Results

15 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 5.3/10.

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
UNILM achieves an RG-1 score of 43.33 on the CNN/DailyMail abstractive summarization task.	×	0.11
UNILM achieves an RG-2 score of 20.21 on the CNN/DailyMail abstractive summarization task.	×	0.11
UNILM achieves an RG-L score of 40.51 on the CNN/DailyMail abstractive summarization task.	✓	0.16
On the Gigaword dataset using only 10K training examples, UNILM achieves an RG-1 score of 32.96.	×	0.04
On the Gigaword dataset using the full training set (3.8M examples), UNILM achieves an RG-1 score of 38.45.	×	0.04
UNILM achieves a score of 61.1 on the CoLA task within the GLUE benchmark.	×	0.05
UNILM achieves an accuracy of 94.5 on the SST-2 task within the GLUE benchmark.	×	0.04
UNILM achieves an F1 score of 90.0 on the MRPC task within the GLUE benchmark.	×	0.08
UNILM achieves a Spearman correlation score of 87.7 on the STS-B task within the GLUE benchmark.	×	0.04
UNILM achieves an overall GLUE benchmark score of 80.8.	×	0.06
UNILM outperforms the best system in the DSTC7 shared task across all evaluation metrics.	×	0.08
The UNILM model was fine-tuned on the CNN/DailyMail training set for 30 epochs.	×	0.08
The masking probability used during UNILM fine-tuning for CNN/DailyMail was 0.7.	×	0.06
The batch size used for UNILM fine-tuning on the Gigaword dataset was 64.	×	0.05
Beam search with a beam size of 5 was used during decoding for the summarization tasks.	×	0.03
The GLUE benchmark consists of nine language understanding tasks.	×	0.13

References

- http://arxiv.org/abs/1905.03197v3
- http://arxiv.org/abs/2306.06371v1
- http://arxiv.org/abs/2308.10783v2