Perplexity and Downstream Reasoning Performance in Language Models

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the relationship between language model perplexity and downstream reasoning task performance. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Scaling Laws for Downstream Task Performance of Large Language Models. Research question: What is the relationship between language model perplexity and downstream reasoning task performance.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

4 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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 BLEU score prediction error is at most 0.061 for the scaling laws in Figure 2.	×	0.08
The downstream cross-entropy loss prediction error is at most 5.95e-12 for the scaling laws in Figure 2.	✓	0.17
As the finetuning dataset size increases, the BLEU score increases and the cross-entropy loss decreases smoothly and mon	×	0.10
Improvements by an increase in the pretraining dataset size are more effective for smaller finetuning datasets.	×	0.05
When the finetuning dataset is large enough, the BLEU score is more or less constant regardless of the pretraining datas	×	0.09
There is little to no improvement of pretraining compared to the non-pretrained models when the finetuning dataset is la	×	0.08
The pretraining dataset in Figure 3 is 100% en-MC4, giving an alignment score of A = 0.7.	×	0.05
The BLEU score and cross-entropy loss are smaller and higher, respectively, in Figure 3 compared to Figure 2 for the sam	×	0.08
The T5-3B model has an embedding dimension of 1024, 32 heads, 24 encoder layers, 24 decoder layers, a head dimension of	×	0.01
The T5-770M model has an embedding dimension of 1024, 16 heads, 24 encoder layers, 24 decoder layers, a head dimension o	×	0.01
The Huber loss is used to minimize overfitting to the outliers when optimizing the scaling law coefficients.	×	0.03
The L-BFGS algorithm is used for optimization of the scaling law coefficients.	×	0.03

References

- http://arxiv.org/abs/2207.08179v1
- http://arxiv.org/abs/1002.1154v1
- http://arxiv.org/abs/2402.04177v3