AdaptToken Scaling Laws and Human Preference Alignment Trade-offs

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the alignment of AdaptToken-8B with human preferences, as measured by benchmark scores on the HHH dataset, differ from AdaptToken-3B, and what trade-offs exist between model size and. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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: How does the alignment of AdaptToken-8B with human preferences, as measured by benchmark scores on the HHH dataset, differ from AdaptToken-3B, and what trade-offs exist between model size and alignment performance?.

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 4.5/10.

3 Results

15 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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 study analyzes models pretrained on a mixture of 50% en-MC4 and 50% de-MC4, then finetuned on en-de translation data	×	0.05
The study analyzes models pretrained on a mixture of 50% en-MC4 and 50% fr-MC4, then finetuned on en-fr translation data	×	0.04
The study analyzes models pretrained on a mixture of 50% en-MC4 and 50% ro-MC4, then finetuned on en-ro translation data	×	0.04
The scaling laws fit empirical results with a prediction error of at most 0.061 for the BLEU score using δ = 0.1.	×	0.08
The scaling laws fit empirical results with a prediction error of at most 5.95e-12 for the downstream cross-entropy loss	✓	0.17
As the finetuning dataset size increases, the BLEU score increases and the cross-entropy loss decreases smoothly and mon	×	0.12
Improvements from increasing pretraining dataset size are more effective for smaller finetuning datasets.	×	0.05
When the finetuning dataset is large enough, the BLEU score remains more or less constant regardless of the pretraining	×	0.08
There is little to no improvement from pretraining compared to non-pretrained models when the finetuning dataset is larg	×	0.07
Changing the pretraining dataset to 100% en-MC4 results in an alignment score of A = 0.7.	×	0.06
Models pretrained on 100% en-MC4 exhibit smaller BLEU scores and higher cross-entropy loss compared to multilingual pret	×	0.12
The T5-3B model has an embedding dimension of 1024, 32 heads, 24 encoder layers, 24 decoder layers, a head dimension of	×	0.02
The T5-770M model has an embedding dimension of 1024, 16 heads, 24 encoder layers, 24 decoder layers, a head dimension o	×	0.02
The study uses Huber loss to minimize overfitting to outliers during the optimization of scaling law coefficients.	×	0.04
The L-BFGS algorithm is used for optimization in this study.	×	0.04

References

- http://arxiv.org/abs/2402.04177v3
- http://arxiv.org/abs/2603.12895v1
- http://arxiv.org/abs/2509.09055v1