Performance-Efficiency Trade-offs in Multimodal Language Models Across Scales and Modalities

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: For multimodal language models (e.g., PaLI, IDEFICS), how does the PER trade-off between image-text and text-only tasks vary across different model sizes when evaluated on benchmarks like LAVIS or. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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: For multimodal language models (e.g., PaLI, IDEFICS), how does the PER trade-off between image-text and text-only tasks vary across different model sizes when evaluated on benchmarks like LAVIS or LLaVA?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 12 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 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.	✓	0.18
As the finetuning dataset size increases, the BLEU score increases and the cross-entropy loss decreases smoothly and mon	×	0.11
As the pretraining dataset size increases, there are improvements in both BLEU score and cross-entropy loss.	×	0.12
For large finetuning datasets, the BLEU score is more or less constant regardless of the pretraining dataset size.	×	0.07
There is little to no improvement of pretraining compared to non-pretrained models when the finetuning dataset is large.	×	0.07
The alignment score A is 0.7 when the pretraining dataset is 100% en-MC4.	×	0.06
The BLEU score is smaller and the cross-entropy loss is higher in Figure 3 compared to Figure 2 for the same finetuning	×	0.10
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 Huber loss is used to minimize overfitting to outliers when optimizing the scaling law coefficients.	×	0.04
The L-BFGS algorithm is used for optimization of the scaling law coefficients.	×	0.04

References

- `http://arxiv.org/abs/2402.04177v3`
- `http://arxiv.org/abs/2201.05729v3`
- `http://arxiv.org/abs/2601.08844v1`