SOVEREIGN: How does the alignment between retriever robustness scores on BEIR and downstream LLM reasoning accuracy in mu

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Scaling laws provide important insights that can guide the design of large language models (LLMs). Existing work has primarily focused on studying scaling laws for pretraining (upstream) loss. However, in transfer learning settings, in which LLMs are pretrained on an unsupervised dataset and then finetuned on a downstream task, we often also care about the downstream performance. In this work, we study the scaling behavior in a transfer learning setting, where LLMs are finetuned for machine translation tasks. Specifically, we investigate how the choice of the pretraining data and its size affe

1 Introduction

Analysis of: Scaling Laws for Downstream Task Performance of Large Language Models. Research goal: How does the alignment between retriever robustness scores on BEIR and downstream LLM reasoning accuracy in multi-hop QA tasks vary when using dense retrievers (e.g., Contriever) versus sparse retrievers (e.g., BM25) across different model scales (e.g., 7B vs. 70B parameters)?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 5 claims extracted, 1 verified. Tribunal: 4.0/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The scaling laws fit well to the empirical results with prediction error at most 0.061 for the BLEU score ($\delta = 0.1$) and	✓	0.16
As the finetuning dataset size increases, the BLEU score increases and the cross-entropy loss decreases smoothly and mon	×	0.12
When the finetuning dataset is large enough, BLEU score is more or less constant regardless of the pretraining dataset s	×	0.11
The T5-3B model has an Embedding Dimension of 1024, 32 heads, and 24 encoder layers.	×	0.02
The T5-770M model has an Embedding Dimension of 1024, 16 heads, and 24 encoder layers.	×	0.02

References

- http://arxiv.org/abs/2604.18234v1
- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2402.04177v3