Scaling Pretraining Data and Few-Shot Learning in Self-Supervised Sequence Models

Assignee Research

June 11, 2026

Abstract

Prior work on language models (LMs) shows that training on a large number of diverse tasks improves few-shot learning (FSL) performance on new tasks. We take this to the extreme, automatically extracting 413,299 tasks from internet tables - orders of magnitude more than the next-largest public datasets. Finetuning on the resulting dataset leads to improved FSL performance on Natural Language Processing (NLP) tasks, but not proportionally to dataset scale. In fact, we find that narrow subsets of our dataset sometimes outperform more diverse datasets. For example, finetuning on software document

1 Introduction

This paper examines: Few-shot Adaptation Works with UnpredicTable Data. Research question: What is the impact of scaling pretraining dataset size on the few-shot learning capabilities of self-supervised sequence models across diverse modalities?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

13 papers retrieved. 12 claims extracted; 8 independently verified. Quality review score: 7.5/10.

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
UnpredicTable is a dataset containing 413,299 few-shot tasks derived from web tables.	✓	0.15
Fine-tuning on narrow subsets of UnpredicTable outperforms fine-tuning on the diverse UnpredicTable dataset and on curat	✓	0.21
Handpicked datasets expected to be helpful are not strongly correlated with model performance.	×	0.08
Training datasets leading to strong improvements often cover trivia content unrelated to downstream test tasks, such as	✓	0.20
Fine-tuning on narrow datasets containing unrelated trivia content causes broad improvements similar to fine-tuning on c	✓	0.17
The method described outperforms multi-task training with 40 NLP datasets in few-shot task transfer.	✓	0.16
GPT2 achieves a 0-shot score of 34.9 on LR tasks, 34.2 on Class tasks, 40.4 on QA tasks, 25.5 on NLI tasks, and 34.2 on	✓	0.15
GPT2 achieves a k-shot score of 38.2 on LR tasks, 37.4 on Class tasks, 40.2 on QA tasks, 34.0 on NLI tasks, and 33.7 on	✓	0.17
MetaICL trained with NLP (IID) data achieves scores of 43.2 (LR), 43.4 (Class), 45.9 (QA), and 33.1 (Para).	×	0.15
MetaICL trained with NLP (OOD) data achieves scores of 38.2 (LR), 38.7 (Class), 49.0 (QA), and 33.1 (Para).	×	0.14
UnpredicTable-5k achieves scores of 43.7 (LR), 46.1 (Class), 42.3 (QA), 36.3 (NLI), and 45.7 (Para).	×	0.14
The UnpredicTable dataset was constructed using tables from the English-language Relational Subset of the WDC Web Table	✓	0.28

References

- http://arxiv.org/abs/1910.03560v2
- http://arxiv.org/abs/2409.03868v1

- http://arxiv.org/abs/2208.01009v2