SOVEREIGN: How do domain-agnostic question answering models trained on mixed-domain datasets (SQuAD 2.0, NewsQA, and Triv

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Finetuning language models on a collection of datasets phrased as instructions has been shown to improve model performance and generalization to unseen tasks. In this paper we explore instruction finetuning with a particular focus on (1) scaling the number of tasks, (2) scaling the model size, and (3) finetuning on chain-of-thought data. We find that instruction finetuning with the above aspects dramatically improves performance on a variety of model classes (PaLM, T5, U-PaLM), prompting setups (zero-shot, few-shot, CoT), and evaluation benchmarks (MMLU, BBH, TyDiQA, MGSM, open-ended generatio

1 Introduction

Analysis of: Scaling Instruction-Finetuned Language Models. Research goal: How do domain-agnostic question answering models trained on mixed-domain datasets (SQuAD 2.0, NewsQA, and TriviaQA) compare in performance degradation using BERT-based models on TPU hardware with batch size 16?.

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. 6 claims extracted, 5 verified. Tribunal: 8.0/10 → APPROVE (revision_round=1). Policy: AUTO_APPROVE.

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
Instruction finetuning with a collection of datasets phrased as instructions has been shown to improve model performance	✓	0.36
Flan-PaLM 540B instruction-finetuned on 1.8K tasks outperforms PaLM 540B by a large margin (+9.4% on average)	✓	0.39
Flan-PaLM 540B achieves 75.2% on five-shot MMLU	✓	0.28
Flan-PaLM 540B achieves state-of-the-art performance on several benchmarks	✓	0.31
	×	0.00
Instruction finetuning with the above aspects dramatically improves performance on a variety of model classes (PaLM, T5,	✓	0.53

References

- https://doi.org/10.18653/v1/2020.acl-main.503
- https://doi.org/10.18653/v1/2020.findings-emnlp.232
- https://doi.org/10.48550/arxiv.2210.11416