SOVEREIGN: What is the impact of negative sampling versus domain-specific fine-tuning on exact match and F1 scores across

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large language models have recently been shown to attain reasonable zero-shot generalization on a diverse set of tasks (Brown et al., 2020). It has been hypothesized that this is a consequence of implicit multitask learning in language models' pretraining (Radford et al., 2019). Can zero-shot generalization instead be directly induced by explicit multitask learning? To test this question at scale, we develop a system for easily mapping any natural language tasks into a human-readable prompted form. We convert a large set of supervised datasets, each with multiple prompts with diverse wording.

1 Introduction

Analysis of: Multitask Prompted Training Enables Zero-Shot Task Generalization. Research goal: What is the impact of negative sampling versus domain-specific fine-tuning on exact match and F1 scores across the 12 in-domain and out-of-domain datasets in the MRQA 2019 benchmark?.

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, 3 verified. Tribunal: 7.0/10 $\rightarrow$ 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
Large language models attain reasonable zero-shot generalization on a diverse set of tasks	✓	0.31
The model attains strong zero-shot performance on several standard datasets	✓	0.28
The model often outperforms models up to 16x its size	×	0.12
The approach attains strong performance on a subset of tasks from the BIG-bench benchmark	✓	0.23
The approach often outperforms models up to 6x its size on BIG-bench benchmark tasks	×	0.15

References

- https://doi.org/10.48550/arxiv.2110.08207
- https://doi.org/10.18653/v1/2020.findings-emnlp.171
- https://doi.org/10.18653/v1/d19-58