SOVEREIGN: How does negative sampling performance scale across different LLM architectures (7B vs 70B) when evaluated on

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

To produce a domain-agnostic question answering model for the Machine Reading Question Answering (MRQA) 2019 Shared Task, we investigate the relative benefits of large pre-trained language models, various data sampling strategies, as well as query and context paraphrases generated by back-translation. We find a simple negative sampling technique to be particularly effective, even though it is typically used for datasets that include unanswerable questions, such as SQuAD 2.0. When applied in conjunction with per-domain sampling, our XLNet (Yang et al., 2019)-based submission achieved the second

1 Introduction

Analysis of: An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering. Research goal: How does negative sampling performance scale across different LLM architectures (7B vs 70B) when evaluated on out-of-distribution benchmarks like MRQA 2019, and what is the optimal balance between negative sampling ratio and model scale for domain-agnostic QA performance?.

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

12 papers retrieved. 8 claims extracted, 0 verified. Tribunal: 1.7/10 → 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 SQuAD fine-tuned model achieves the best results on both in and out-domain 'Macro-Average' Exact Match.	×	0.09
NewsQA, SearchQA, and TriviaQA performed the worst on the out-domain Macro-Average among models with multi-domain pre-fi	×	0.06
SearchQA is the largest dataset by number of examples.	×	0.02
SearchQA's long contexts generate 657K segments, double that of the next largest dataset.	×	0.03
Including No Answer segments in the training set drastically outperformed the typical practice of excluding these segmen	×	0.03
Including NA segments increases Out-Domain EM from 43.78 to 50.04 on the XBC model at max sequence length of 200.	×	0.03
Back-translated augmentations yield no noticeable improvement.	×	0.00
Negative samples designed to teach the model when to abstain from predictions prove highly effective out-domain.	×	0.05

References

- http://arxiv.org/abs/2605.11209v1
- http://arxiv.org/abs/1912.02145v1
- http://arxiv.org/abs/2410.13187v3