

# SOVEREIGN: How do simple negative sampling techniques compare to advanced data augmentation methods for improving out-of-

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 do simple negative sampling techniques compare to advanced data augmentation methods for improving out-of-distribution performance in machine reading question answering benchmarks?.

## 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. 4 claims extracted, 0 verified. Tribunal: 3.3/10 → 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.	×	0.10
Models with multi-domain pre-fine-tuning NewsQA, SearchQA, and TriviaQA performed the worst on the out-domain (O) Macro-	×	0.05
SearchQA is the largest dataset by number of examples and generates 657K segments, double that of the next largest datas	×	0.02
For our later experiments we chose not to use all instances of a detected answer, as this would further skew our multi-d	×	0.05

### References

- <http://arxiv.org/abs/2506.09781v2>
- <http://arxiv.org/abs/1912.02145v1>
- <http://arxiv.org/abs/1910.09753v2>