SOVEREIGN: How does negative sampling ratio affect zero-shot question answering performance across different LLM architec

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 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 ratio affect zero-shot question answering performance across different LLM architectures (7B vs 70B) on out-of-distribution benchmarks like MRQA 2019?.

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. 6 claims extracted, 0 verified. Tribunal: 2.3/10 $\rightarrow$ 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
SearchQA is the largest dataset by number of examples.	×	0.03
Our training procedure for each model involved fine-tuning the Transformer over two epochs, each with three validation c	×	0.07
The checkpoint with the highest Out-Domain Macro-Average was selected as the best for that training run.	×	0.04
We found this drastically outperformed the typical practice of excluding No Answer segments.	×	0.01
The improvement is exaggerated at the shorter max sequence length (MSL) of 200, where including NA segments increases Ou	×	0.02

References

- http://arxiv.org/abs/1910.09753v2
- http://arxiv.org/abs/1912.02145v1
- http://arxiv.org/abs/2404.14700v4