SOVEREIGN: What is the comparative evaluation of negative sampling versus domain-specific fine-tuning on MRQA 2019 benchm

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: What is the comparative evaluation of negative sampling versus domain-specific fine-tuning on MRQA 2019 benchmark scores for 7B and 70B parameter models?.

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

10 papers retrieved. 13 claims extracted, 0 verified. Tribunal: 3.3/10 $\rightarrow$ 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
Including No Answer segments in the training set drastically outperformed the typical practice of excluding these segmen	×	0.02
The improvement from including NA segments is particularly noticeable on datasets with longer sequences	×	0.04
Out-Domain EM increases from 43.78 to 50.04 on the XBC model at MSL of 200 when including NA segments	×	0.02
SearchQA is the largest dataset by number of examples	×	0.02
SearchQA generates 657K segments, double that of the next largest dataset	×	0.02
The training procedure involves fine-tuning the Transformer over two epochs, each with three validation checkpoints	×	0.08
Using only one detected answer per example prevents skewing multi-domain samples towards certain datasets	×	0.03
The original multi-domain dataset consisted of 75k examples from every training set	×	0.02
The effective batch size for BBC and XBC is 200	×	0.01
The effective batch size for XLC is 144	×	0.01
Lower learning rate and gradient accumulation are critical to achieve training stability	×	0.04
Negative samples designed to teach the model when to abstain from predictions prove highly effective out-domain	×	0.04

References

- http://arxiv.org/abs/2410.13187v3
- http://arxiv.org/abs/1912.02145v1
- http://arxiv.org/abs/1910.09753v2