SOVEREIGN: An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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 different data sampling strategies including negative sampling affect the inference latency and throughput of large language models when deployed for real-time question answering, and what are the accuracy-efficiency trade-offs?.

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. 0 claims extracted, 0 verified. Tribunal: 5.0/10 → 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.

References

- http://arxiv.org/abs/2507.22352v1
- http://arxiv.org/abs/2504.11972v2
- http://arxiv.org/abs/1912.02145v1