SOVEREIGN: To what extent does incorporating unanswerable questions through negative sampling techniques improve performa

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

We present the results of the Machine Reading for Question Answering (MRQA) 2019 shared task on evaluating the generalization capabilities of reading comprehension systems. 1 In this task, we adapted and unified 18 distinct question answering datasets into the same format. Among them, six datasets were made available for training, six datasets were made available for development, and the final six were hidden for final evaluation. Ten teams submitted systems, which explored various ideas including data sampling, multi-task learning, adversarial training, and ensembling. The best system achieve

1 Introduction

Analysis of: MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. Research goal: To what extent does incorporating unanswerable questions through negative sampling techniques improve performance on the MRQA datasets when evaluated on the SQuAD 2.0 and MRQA out-of-distribution splits?.

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

7 papers retrieved. 9 claims extracted, 9 verified. Tribunal: 7.3/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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 MRQA 2019 shared task focused on evaluating the generalization capabilities of reading comprehension systems.	✓	0.41
The MRQA 2019 task adapted and unified 18 distinct question answering datasets into the same format.	✓	0.36
Six datasets were made available for training in the MRQA 2019 task.	✓	0.23
Six datasets were made available for development in the MRQA 2019 task.	✓	0.23
Six datasets were hidden for final evaluation in the MRQA 2019 task.	✓	0.21
Ten teams submitted systems to the MRQA 2019 shared task.	✓	0.29
Submitted systems in the MRQA 2019 task explored ideas including data sampling, multi-task learning, adversarial trainin	✓	0.36
The best system in the MRQA 2019 task achieved an average F1 score of 72.5 on the 12 held-out datasets.	✓	0.32
The best system in the MRQA 2019 task outperformed the initial BERT-based baseline by 10.7 absolute points.	✓	0.19

References

- https://doi.org/10.18653/v1/d19-5801
- https://doi.org/10.18653/v1/2020.acl-main.503

- https://doi.org/10.18653/v1/2021.emnlp-main.696