Multi-Step Retrieval Strategies Enhance BLEU Scores in Sub-10B Parameter Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: Does integrating multi-step retrieval strategies improve BLEU scores for sub-10B parameter models on complex query datasets like SQuAD. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Overcoming low-utility facets for complex answer retrieval. Research question: Does integrating multi-step retrieval strategies improve BLEU scores for sub-10B parameter models on complex query datasets like SQuAD?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

10 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Manual relevance judgments are graded on a scale from Must be mentioned (3) to Trash (-2).	×	0.01
Manual relevance judgments cover a subset of queries (702 of the 2,125 queries).	×	0.02
Two versions of the knowledge graph are generated: one using hyperlinks as entity mentions and one using entity mentions	×	0.07
Edge labels are limited to the top 1000 most frequently-used labels.	×	0.02
HolE embeddings are trained for 5,000 iterations.	×	0.03
Top 2 entity scores are included in the model.	×	0.05
Models are trained for 80 epochs with samples from train.fold1-2.	×	0.06
Automatic relevance judgments serve as a source of relevant documents.	×	0.01
Top non-relevant BM25 documents are used as negative training examples.	×	0.04
For each positive sample, 6 negative samples are included.	×	0.06
The training iteration that yields the highest R-Precision value on the validation dataset (test200) is selected for eva	×	0.02
Top 100 BM25 results for each query in benchmarkY1test are reranked.	×	0.03
Neural models have been shown to be competitive with conventional ranking techniques due to their ability to learn match	×	0.06
Preliminary work has shown neural ranking architectures to be more effective than conventional approaches out-of-the-box	×	0.05

References

- http://arxiv.org/abs/2605.02623v1
- http://arxiv.org/abs/1811.08772v1
- http://arxiv.org/abs/2404.14464v1