Scaling Retriever Portfolios: Accuracy-Latency Trade-offs in Complex Question Answering

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the trade-off between answer accuracy and retrieval latency when scaling the number of candidates in the retriever portfolio for complex question-answering tasks. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: RealTime QA: What's the Answer Right Now?. Research question: What is the trade-off between answer accuracy and retrieval latency when scaling the number of candidates in the retriever portfolio for complex question-answering tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

14 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
REALTIME QA is a multiple-choice question dataset.	×	0.08
The NOTA (none of the above) setting resulted in performance degradation across the board.	×	0.03
NOTA choices can be found in other multiple-choice QA or reading comprehension datasets (Richardson et al., 2013; Lai et	×	0.07
Under the generation setting, performance is evaluated with exact matching and token-based F1 scores.	×	0.05
Human performance on REALTIME QA resulted in an accuracy of 96.7%.	×	0.06
REALTIME QA executes six baselines in real time that are based on strong pretrained models: four open-book and two close	×	0.12
Open-book QA models follow a two-step pipeline: document retrieval and answer prediction.	×	0.06
Open-book systems have the advantage of being capable of updating the external knowledge source at test time.	×	0.06
For the retrieval step, two configurations are experimented with: top-5 Wikipedia documents from dense passage retrieval	×	0.03
REALTIME QA is a dynamic benchmark based on newly-published news articles.	×	0.13
REALTIME QA provides a regularly-updated (weekly in the current version) evaluation platform for the research community.	×	0.10
Every week, REALTIME QA retrieves news articles and human-written, multiple-choice questions from news websites (CNN, TH	×	0.04

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2207.13332v2
- http://arxiv.org/abs/cs/0107006v1