Retriever Portfolio Robustness to Domain Shifts in Specialized Reasoning Tasks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How robust are retriever portfolios to domain shifts, such as from general question-answering (AmbiEval) to specialized domains like legal or biomedical reasoning, as evaluated by accuracy on. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CPRet: A Dataset, Benchmark, and Model for Retrieval in Competitive Programming. Research question: How robust are retriever portfolios to domain shifts, such as from general question-answering (AmbiEval) to specialized domains like legal or biomedical reasoning, as evaluated by accuracy on domain-specific benchmarks like BRATS or JURASSIC?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

15 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 5.2/10.

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
NDCG@10 is used as the primary metric to measure ranking quality across tasks.	×	0.03
The retrieval pool was restricted to problems published before April 2023.	×	0.03
LiveCodeBench problems appeared starting May 2023.	×	0.04
Three retrieval settings were compared: no retrieval (baseline), using Qwen3-Embedding-4B as the retriever, and using CP	×	0.05
Three models from the Qwen-Coder series were used for code generation evaluation: qwen-coder-turbo, qwen-coder-plus, and	×	0.03
qwen-coder-turbo and qwen-coder-plus are non-reasoning models.	×	0.02
qwen3-coder-plus is a reasoning-capable model.	×	0.02
CPRetriever-Code-2B achieved scores of 70.40, 70.59, 38.68, 81.45, and 65.28 in different tasks.	×	0.03
CPRetriever-Prob-2B achieved scores of 56.50, 70.68, 60.06, 90.74, and 69.50 in different tasks.	×	0.03
CPRetriever-Code-Qwen3-4B achieved scores of 86.22, 86.70, 41.14, 88.10, and 75.54 in different tasks.	×	0.03
CPRetriever-Prob-Qwen3-4B achieved scores of 80.84, 87.10, 74.33, 96.15, and 84.60 in different tasks.	×	0.03
Four distinct retrieval tasks are defined: Text-to-Code Retrieval, Code-to-Code Retrieval, Problem-to-Duplicate Retrieva	✓	0.15
Text-to-Code Retrieval assesses a model's ability to align problem semantics with executable implementations.	×	0.07
Code-to-Code Retrieval evaluates a model's capacity to understand code functionality and identify semantically equivalen	×	0.06
Problem-to-Duplicate Retrieval is useful for identifying redundancy and measuring problem novelty across platforms.	×	0.08
Simplified-to-Full Retrieval examines cross-abstraction retrieval capabilities, bridging accessibility-oriented rewrites	×	0.05

References

- http://arxiv.org/abs/2605.31176v1
- http://arxiv.org/abs/2505.12925v2
- http://arxiv.org/abs/2308.13515v5