Duplicate Problem Retrieval Accuracy and LLM Code Reasoning Benchmark Validity

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does duplicate problem retrieval accuracy in competitive programming datasets correlate with the evaluation validity of large language models on code reasoning benchmarks. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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 does duplicate problem retrieval accuracy in competitive programming datasets correlate with the evaluation validity of large language models on code reasoning benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

16 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
The study uses NDCG@10 as the primary metric to measure ranking quality across retrieval tasks.	×	0.04
The retrieval pool for the RAG experiments was restricted to problems published before April 2023.	×	0.04
The test set for the RAG experiments consisted of problems from LiveCodeBench appearing starting May 2023.	×	0.04
Three retrieval settings were compared: no retrieval (baseline), using Qwen3-Embedding-4B, and using CPRetriever-Prob.	×	0.04
CPRetriever-Prob is based on Qwen3-Embedding-4B.	×	0.05
The code generation evaluation used qwen-coder-turbo and qwen-coder-plus as non-reasoning models.	×	0.05
The code generation evaluation used qwen3-coder-plus as a reasoning-capable model.	×	0.07
In Table 8, CPRetriever-Prob-Qwen3-4B achieved a Pass@1 score of 84.60 on the average of the four tasks.	×	0.05
In Table 8, CPRetriever-Code-Qwen3-4B achieved a Pass@1 score of 75.54 on the average of the four tasks.	×	0.06
Text-to-Code Retrieval is defined as retrieving correct solution codes given a natural language problem description.	×	0.11
Code-to-Code Retrieval is defined as retrieving alternative correct solutions given one accepted solution.	×	0.06
Problem-to-Duplicate Retrieval is defined as retrieving duplicated or closely related problems given a problem descripti	×	0.08
Simplified-to-Full Retrieval is defined as retrieving the full problem description given a simplified version.	×	0.05
The benchmark covers four dimensions: problem-code alignment, solution diversity, duplication detection, and abstraction	×	0.09

References

- http://arxiv.org/abs/2407.08716v1
- http://arxiv.org/abs/2505.12925v2
- http://arxiv.org/abs/2604.18234v1