EVOR's Multi-Source Knowledge Adaptation and Pass@k Gains on Out-of-Domain Programming Tasks

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does EVOR's diverse knowledge base adaptation improve pass@k scores on out-of-domain programming tasks relative to single-source retrieval methods. Recently the retrieval-augmented generation (RAG) has been successfully applied in code generation. However, existing pipelines for retrieval-augmented code generation (RACG) employ static knowledge bases with a single source, limiting the adaptation capabilities of Large. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: EVOR: Evolving Retrieval for Code Generation. Research question: To what extent does EVOR's diverse knowledge base adaptation improve pass@k scores on out-of-domain programming tasks relative to single-source retrieval methods?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

13 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
With CodeLlama, the improvements of MPSC, ExeDec, and Reflexion on EVOR-BENCH are smaller than 2% on average compared to	×	0.06
The execution accuracy of MPSC, ExeDec, and Reflexion remains 0 on the Ring subset of EVOR-BENCH.	×	0.08
DocPrompting significantly surpasses MPSC, ExeDec, and Reflexion on EVOR-BENCH.	×	0.07
EVOR achieves a 16.1% absolute gain over DocPrompting when using ChatGPT.	×	0.06
EVOR achieves a 16.2% absolute gain over DocPrompting when using CodeLlama.	×	0.06
DocPrompting uses documentation as a single retrieval source without evolution in queries and knowledge.	×	0.13
The paper uses execution accuracy (pass@1) as the default metric for evaluation.	×	0.04
MPSC prompts LLMs to generate diverse outputs from three perspectives: Solution, Specification, and Test case.	×	0.03
MPSC constructs a 3-partite graph and selects the optimal solution based on confidence scores.	×	0.03
ExeDec employs a subgoal model to predict the subgoal of the desired program state for the next part of the program.	×	0.02
ExeDec uses a synthesizer model to generate the corresponding subprogram to achieve the predicted subgoal.	×	0.02
Retrieval-augmented code generation carries a risk of retrieving biased or incorrect information that could propagate er	×	0.09
Retrieval-augmented code generation poses privacy and security risks if sensitive code snippets are inadvertently includ	×	0.09

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2412.21199v2

- http://arxiv.org/abs/2510.08325v2