Scaling Diverse Knowledge Bases in Retrieval-Augmented Code Generation and Human Alignment

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the scaling of diverse knowledge bases in retrieval-augmented code generation affect the alignment of generated code with human intentions, as measured by functional correctness on MBPP and. 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: EVOR: Evolving Retrieval for Code Generation. Research question: How does the scaling of diverse knowledge bases in retrieval-augmented code generation affect the alignment of generated code with human intentions, as measured by functional correctness on MBPP and HumanEval?.

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 3.7/10.

3 Results

10 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
Existing code generation approaches perform poorly on EVOR-BENCH.	×	0.12
With CodeLlama, the improvements of MPSC, ExeDec, and Reflexion are smaller than 2% on average, compared to the vanilla	×	0.03
The execution accuracy remains 0 in Ring across three methods (MPSC, ExeDec, Reflexion).	×	0.06
DocPrompting significantly surpasses MPSC, ExeDec, and Reflexion by a large margin.	×	0.03
EVOR achieves 16.1% and 16.2% absolute gain with ChatGPT and CodeLlama respectively on top of DocPrompting.	×	0.05
DocPrompting only uses the documentation as a single retrieval source, without evolution in both queries and knowledge.	×	0.13
The use of retrieval-augmented code generation with large language models introduces several potential risks, primarily	×	0.10
There is a risk of biased or incorrect information being retrieved, which could propagate errors or introduce vulnerabil	×	0.03
There are concerns about privacy and security if sensitive code snippets are inadvertently included in the retrieval pro	×	0.02
Addressing these risks requires careful curation of retrieval sources, robust validation mechanisms, and continuous moni	×	0.03
We use the execution accuracy (pass@1) as the metric throughout the paper.	×	0.04
We compare EVOR to the vanilla generation and four recent methods that demonstrate significant performance improvement i	×	0.11
MPSC incorporates both inter- and intra consistency.	×	0.00
ExeDec employs a subgoal model to predict the subgoal of the desired program state for the next part of the program and	×	0.02

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2504.19754v1
- http://arxiv.org/abs/2412.21199v2