Retrieval-Augmented Generation Impact on Sub-10B Model Pass@k Accuracy in HumanEval-Java

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the integration of retrieval-augmented generation affect the pass@k accuracy of sub-10B parameter models on the HumanEval-Java benchmark compared to fine-tuned baselines. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. 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 integration of retrieval-augmented generation affect the pass@k accuracy of sub-10B parameter models on the HumanEval-Java benchmark compared to fine-tuned baselines?.

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 4.0/10.

3 Results

14 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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.	×	0.13
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).	×	0.05
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 default configuration of EVOR uses execution accuracy (pass@1) as the metric.	×	0.06
Vanilla generation directly gets outputs from LLMs based on the coding question without augmenting external knowledge.	×	0.07
MPSC incorporates both inter- and intra consistency and prompts LLMs to generate diverse outputs from three perspectives	×	0.03
ExeDec employs a subgoal model to predict the subgoal of the desired program state for the next part of the program and	×	0.03

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2504.16584v1

- http://arxiv.org/abs/2503.16581v1