SOVEREIGN: How does RAG performance vary across different external knowledge bases when evaluated on the HotPotQA benchma

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

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 Language Models (LLMs) to domains they have insufficient knowledge of. In this work, we develop a novel pipeline, EVOR, that employs the synchronous evolution of both queries and diverse knowledge bases. On two realistic settings where the external knowledge is required to solve code generation tasks, we compile four new data

1 Introduction

Analysis of: EVOR: Evolving Retrieval for Code Generation. Research goal: How does RAG performance vary across different external knowledge bases when evaluated on the HotPotQA benchmark using Llama-2-7B as the backbone model?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

3 papers retrieved. 5 claims extracted, 0 verified. Tribunal: 3.0/10 $\rightarrow$ REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Existing code generation approaches perform poorly on EVOR-BENCH, with improvements of less than 2% when using CodeLlama	×	0.11
Execution accuracy remains 0 for the Ring benchmark across MPSC, ExeDec, and Reflexion methods.	×	0.05
DocPrompting significantly outperforms MPSC, ExeDec, and Reflexion methods on EVOR-BENCH.	×	0.06
EVOR achieves a 16.1% absolute gain with ChatGPT and 16.2% with CodeLlama over DocPrompting.	×	0.05
DocPrompting uses documentation as a single retrieval source without evolution in queries and knowledge.	×	0.13

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2411.18583v1
- http://arxiv.org/abs/2506.06962v3