SOVEREIGN: How do different prompting strategies affect the calibration of uncertainty estimates in retrieval-augmented l

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 do different prompting strategies affect the calibration of uncertainty estimates in retrieval-augmented language models during out-of-distribution inference?.

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

11 papers retrieved. 6 claims extracted, 0 verified. Tribunal: 1.0/10 $\rightarrow$ 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 CodeLlama showing improvements of MPSC, ExeDec, a	×	0.10
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, confirming that domain knowledge is	×	0.07
EVOR achieves 16.1% and 100.2% absolute gain with ChatGPT and CodeLlama respectively on top of DocPrompting.	×	0.05
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.	×	0.12

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2310.14025v1
- http://arxiv.org/abs/2411.18583v1