EnCoNeSyRAG: A Neuro-Symbolic Framework for Engineered Agency and Bidirectional Coordination in Lightweight RAG

Dilana Mathugama
Department of Computer Science and Engineering
University of Westminster
London, UK
dilanasasanka@gmail.com

Dilum De Silva
School of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
dilum.s@iit.ac.lk

ABSTRACT

Retrieval-Augmented Generation (RAG) systems mitigate Large Language Model (LLM) hallucinations by grounding generation in external knowledge. However, current architectures rely on passive, unidirectional pipelines where the generator blindly consumes retrieved context, leading to "shallow interaction" and parametric overconfidence when retrieving adversarial or insufficient data. This study introduces EnCoNeSyRAG, a neuro-symbolic framework designed to engineer agency within lightweight RAG systems (1.2B parameters). Utilizing Parameter-Efficient Fine-Tuning (LoRA), the generative backbone acts as a Causal Planner, emitting symbolic control tokens ([SEARCH]) to autonomously trigger hybrid retrieval (FAISS + BM25 via Reciprocal Rank Fusion) only upon detecting epistemic uncertainty. Crucially, the framework implements a Self-Correcting Feedback Loop governed by S-BERT semantic validation. If retrieved context fails quality thresholds, the system dynamically refines the query or executes a "Safe Abstention" guardrail to guarantee zero hallucinations on out-of-domain queries. Empirical evaluations on a domain-filtered HotpotQA subset demonstrate that EnCoNeSyRAG achieves an Acc† of 53.3% on complex multi-hop reasoning, rivaling 70B+ parameter models, and yielding an unprecedented Performance-to-Parameter Ratio (PPR) of 44.41. The results confirm that bidirectional, neuro-symbolic coordination effectively replaces raw parameter scale for reliable, hallucination-free enterprise AI.

CCS CONCEPTS

• Computing methodologies • Artificial intelligence • Natural language processing • Natural language generation • Information systems • Information retrieval • Retrieval models and ranking • Query representation

KEYWORDS

Neuro-Symbolic AI, Retrieval-Augmented Generation, Bidirectional Coordination, Hallucination Prevention, Instruction Tuning, Safe Abstention

1 Introduction

Retrieval-Augmented Generation (RAG) was introduced to resolve the factual inaccuracies inherent in the static parametric memory of Large Language Models (LLMs) [1]. However, a fundamental vulnerability persists across modern implementations: the "shallow interaction" between the retriever and the generator [2]. In standard pipelines, the retriever fetches documents based on a user's initial query, and the generator is forced to passively synthesize an answer from the concatenated text. Because the generative model does not understand why the context was provided, it remains highly susceptible to "parametric overconfidence." When confronted with noisy or adversarial context, the model will often hallucinate a plausible but incorrect answer rather than safely abstaining.

Recent works have attempted to bridge this semantic gap through active retrieval triggers [3] or iterative draft refinement [4]. However, these approaches often rely on rigid heuristics or require massive, computationally expensive models (e.g., 70B+ parameters) to orchestrate the feedback loops, rendering them impractical for resource-constrained edge deployments.).

This research proposes EnCoNeSyRAG, a framework that engineers true bidirectional agency into lightweight models. By adopting a neuro-symbolic approach, the system transforms the passive generator into an active "Causal Planner," enabling it to autonomously dictate retrieval timing, refine failing queries, and deterministically halt generation (Safe Abstention) when presented with insufficient evidence.

2 Methodology

The EnCoNeSyRAG framework utilizes a 3-tier Controller-Service architecture to separate probabilistic text generation from deterministic control logic. The core bidirectional coordination is governed by a rigorous self-correcting algorithm, detailed formally in Figure 1. The algorithm operates in three distinct phases:

Algorithm 1: Adaptive Self-Correcting Bidirectional RAG

REQUIRE: Q_{user} (user query)
ENSURE : FinalAnswer

```
// Phase 1: Assessment
PlanResponse ← G.generate("User: " || Q_user, model_type = "adaptive")
(NeedsSearch, q, DirectText) ← G.parse_output(PlanResponse)
if ¬NeedsSearch then
    RETURN DirectText

// Phase 2: Feedback Loop
MaxRetries ← 1
Attempt ← 0
C* ← []
τ ← 0.45
while Attempt ≤ MaxRetries do
    // Retrieval: hybrid search, top-k = 3
    C ← R.search(q, k = 3)
    // Grading: semantic relevance of the top chunk
    σ ← E.calculate_semantic(q, C[0])
    // Threshold check
    if σ ≥ τ then
        C* ← C
        break
    else
        // Refinement
        if Attempt < MaxRetries then
            q ← G.refine(q)
            Attempt ← Attempt + 1
        else
            C* ← C
            break

// Phase 3: Contextual Synthesis
Context ← format(C*)
RawAnswer ← G.generate(Context, Q_user)
// Algorithmic Guardrail (Safe Abstention)
(IsSearchingAgain, _, _) ← G.parse_output(RawAnswer)
if IsSearchingAgain then
    FinalAnswer ← "I apologize, insufficient information..."
else
    FinalAnswer ← RawAnswer
RETURN FinalAnswer
```

Figure 1: Formal pseudocode of the Adaptive Self-Correcting Bidirectional Algorithm, highlighting the neuro-symbolic feedback loop and the Safe Abstention guardrail

2.1 Cognitive Assessment

The process begins when a natural language query is passed zero-shot to the generative backbone (a Llama-3.2-1B model fine-tuned via LoRA). If the model determines its internal parametric memory is sufficient, it outputs a direct answer. If epistemic uncertainty is detected, it emits a strict symbolic control token (e.g., [SEARCH: <entity>]), initiating the retrieval protocol.

2.2 Self-Correcting Feedback Loop

This phase implements the bidirectional coordination. A Hybrid Retriever fetches candidate chunks using Reciprocal Rank Fusion (RRF) across Sparse (BM25) and Dense (FAISS) indices. A Semantic Evaluator (S-BERT) then calculates a cosine similarity score (σ) for the top-ranked chunk.

- If $\sigma \geq \tau$ (where $\tau = 0.45$), the context is validated and passed to synthesis.
- If $\sigma < \tau$, the system intercepts the failure and sends a backward signal. The generator is prompted to refine and broaden the search query. This loop repeats until valid evidence is found or the retry limit is exhausted.

2.3 Contextual Synthesis and Safe Abstention

The validated context is fed back to the generator to synthesize the final response. Crucially, this phase incorporates an Algorithmic Guardrail. If the retrieved context remains objectively insufficient after all refinement retries, the system intercepts the generator's attempt to synthesize and executes a graceful halt, returning a deterministic Safe Abstention message. This guarantees a zero-hallucination rate against adversarial queries.

3 Experimental Setup & Results

To isolate architectural reasoning capabilities from scale-based knowledge gaps, system-level benchmarking was conducted on a domain-filtered corpus of 1,000 queries adapted from HotpotQA (multi-hop) and SQuAD (single-hop).

Because traditional exact-match metrics penalize conversational AI, a 70B-parameter model (Llama-3.3-70B) was utilized as an automated judge (Acc†) over a statistically significant sample (N=300).

3.1 Efficiency and Interaction Profiling

Analysis of the execution logs confirmed the framework's engineered agency. The system successfully bypassed retrieval entirely for 17.5% of queries (Direct Answer Efficiency). When retrieval was activated, the system self-corrected its own search query 21.9% of the time, intercepting poor context before generation.

3.2 SOTA Benchmarking and PPR

Evaluating a 1.2B-parameter model against massive State-of-the-Art (SOTA) frameworks requires standardizing for computational resources. A Performance-to-Parameter Ratio (PPR) was calculated (Main Metric Score / Model Parameters in Billions).

Architecture	Generative Backbone	Params	Multi-Hop Score	PPR
IRAGKR [5]	Llama2-13B-chat	13B	F1: 39.4%	3.03
R2AG [6]	Llama2-7B	7B	Acc: 66.7%	9.52
ITER-RETGEN [4]	text-davinci-003	~175B*	Acc†: 71.2%	~0.40
EnCoNeSyRAG	Llama-3.2-1B	1.2B	Acc†: 53.3%	44.41

Table 1: Cross-Architecture Benchmarking (PPR) on Multi-Hop QA

As shown in Table 1, EnCoNeSyRAG achieved a highly competitive multi-hop accuracy of 53.3%. By engineering bidirectional coordination at the system level, the 1.2B model achieved an unprecedented PPR of 44.41, vastly outperforming resource-heavy methods like R2AG.

4 Conclusion

The EnCoNeSyRAG framework successfully shifts the RAG paradigm from blind data consumption to active, epistemic reasoning. By integrating Parameter-Efficient Fine-Tuning, Hybrid Retrieval, and a mathematically gated Self-Correcting Feedback Loop, the prototype demonstrates that factual reliability and strict hallucination prevention can be achieved on lightweight, edge-deployable models, offering a robust blueprint for future enterprise AI systems.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Mr. Dilum De Silva for his invaluable supervision, guidance, and continuous support throughout the development of this research. Special thanks are also extended to the Informatics Institute of Technology (IIT), Sri Lanka, and the University of Westminster, UK, for providing the academic foundation and resources that made this project possible.

Positionality Statement: I declare that this work is original and adheres to the highest standards of academic integrity. This research is conducted in alignment with the British Computer Society (BCS) Code of Conduct, ensuring that AI development remains transparent, responsible, and serves the public interest.

REFERENCES

[1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, 2020, pp. 9459–9474.

[2] A. Asai et al., "Reliable, Adaptable, and Attributable Language Models with Retrieval," arXiv preprint arXiv:2403.03187, 2024.

[3] Z. Jiang et al., "Active Retrieval Augmented Generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 7969–7992.

[4] Z. Shao et al., "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy," in Findings of the Association for Computational Linguistics: EMNLP, 2023, pp. 9248–9274.

[5] K. Du et al., "IRAGKR: Iterative retrieval augmented generation with fine-grained knowledge refinement," Neurocomputing, vol. 655, p. 131282, 2025.

[6] F. Ye et al., "R^2AG: Incorporating Retrieval Information into Retrieval Augmented Generation," in Findings of the Association for Computational Linguistics: EMNLP, 2024, pp. 11584–11596