Critic-Based Self-Correction in Local Retrieval-Augmented Generation: An Empirical Evaluation in an Institutional Academic Chatbot

Gabriel de Souza Lima[1], Júlio César Mansano[1], Fernanda Diniz Marabin[1], João Emmanuel D'Alkmin Neves[1]

[1]Faculdade de Tecnologia de Americana Ministro Ralph Biasi – FATEC Americana
Americana – SP – Brasil

lima.gabriel94@outlook.com, julio.mansano@aluno.cps.sp.gov.br,
fernanda.marabin@aluno.cps.sp.gov.br, joao.neves@cps.gov.br

Abstract. *This paper evaluates critic-based self-correction in local Retrieval-Augmented Generation for an institutional academic chatbot. Five configurations were compared: a baseline without critic, a deterministic critic, two critic variants with a Large Language Model judge, and a conditional oracle policy. Evaluation used RAGAS, end-to-end latency, and paired Wilcoxon tests by query type. The central finding is that critic-based self-correction reduces Answer Relevancy, our proxy for response utility, in constrained local environments: uniform critic activation increases latency substantially and does not compensate for the synthesis limitations of small-scale models. The conditional policy improves the quality-cost trade-off over always-on critics, but does not outperform the baseline globally.*

1. Introduction

The adoption of chatbots supported by Large Language Models (LLMs) has expanded the scope of automated assistance in educational contexts. In public higher education institutions, such systems can support queries related to regulations, academic calendars, procedures, and frequently asked administrative questions, thereby reducing the workload of academic offices. However, the direct use of generative models without document grounding remains strongly associated with inaccurate responses, omission of regulatory exceptions, and factual hallucinations [Lewis et al., 2020; Gao et al., 2023].

Retrieval-Augmented Generation (RAG) combines text generation with the retrieval of relevant document passages, strengthening the link between answers and verifiable evidence [Lewis et al., 2020]. In institutional scenarios, this approach is particularly attractive because it enables operation over local corpora without mandatory reliance on cloud services, favoring privacy, data sovereignty, and operational control. Similar concerns regarding autonomous local decision-making, adaptive computational behavior, and operational efficiency have also been explored in intelligent multi-agent architectures applied to real-world environments [Neves et al., 2023]. Nevertheless, merely adopting RAG architecture does not guarantee reliable answers. System behavior is highly dependent on retrieval quality, context selection, the synthesis capacity of the generator, and the presence of additional verification mechanisms [Gao et al., 2023; Es et al., 2024].

Recent research has therefore explored adaptive strategies capable of deciding when retrieval, revision, or corrective mechanisms should be activated during generation. Notable examples include Self-RAG [Asai et al., 2024], FLARE [Jiang et al., 2023], Adaptive-RAG [Jeong et al., 2024], and CRAG [Yan et al., 2024]. Although these approaches are frequently associated with improvements in factuality and response quality, such results are predominantly obtained with large-scale models and substantial computational resources. It remains unclear whether these gains persist in fully local, resource-constrained environments, where limited model capacity and the additional overhead introduced by adaptive mechanisms may ultimately compromise overall system utility.

Against this backdrop, this paper investigates the impact of a critic-based self-evaluation mechanism on both response quality and latency in a fully local institutional academic chatbot built with small-scale language models. To guide this empirical investigation, the following Research Questions (RQs) are formulated:

- RQ1: Do critic-based self-evaluation mechanisms improve response quality in local RAG systems?
- RQ2: What is the impact of critic activation on latency and computational cost?
- RQ3: Do critic effects vary across different query types?
- RQ4: Do conditional activation policies provide a better quality–cost trade-off?

The initial hypothesis posited that critic activation would improve system reliability, particularly for sensitive regulatory queries, and that a conditional activation policy would yield a superior balance between quality and cost.

Contrary to prevalent assumptions in the literature, this work empirically demonstrates that critic mechanisms in local RAG systems can reduce Answer Relevancy, our automated proxy for response utility. Accordingly, the paper offers three main contributions:

(i) empirical evidence that uniform critic activation negatively impacts estimated response utility, as measured by Answer Relevancy, under hardware constraints, establishing a practical limit for these mechanisms.

(ii) a stratified analysis showing that critic effects are strictly query-type dependent, rendering aggregated interpretations insufficient.

(iii) the proposal and analysis of a simulated conditional policy that mitigates degradation and offers a more favorable trade-off than always-on approaches.

2. Related Work

The conceptual foundation of this work lies in Retrieval-Augmented Generation as introduced by Lewis et al. [2020], later consolidated by extensive surveys covering architectures, retrieval variants, and limitations of document-grounded generative systems [Gao et al., 2023]. On the evaluation side, the RAGAS framework has become a reference for operationalizing metrics such as Faithfulness, Answer Relevancy, and Context Precision without exclusive reliance on large-scale human annotation [Es et al., 2024]. In parallel, the LLM-as-a-Judge paradigm has been increasingly adopted for output assessment, although its reliability is highly task- and model-dependent [Zheng et al., 2024].

Regarding adaptive strategies, Self-RAG [Asai et al., 2024] enables models to generate self-reflective signals guiding retrieval and correction. FLARE [Jiang et al., 2023] extends active retrieval throughout the generation process, while Adaptive-RAG [Jeong et al., 2024] selects policies based on question complexity. CRAG [Yan et al., 2024] introduces a lightweight evaluator of retrieved document quality, triggering corrective actions when retrieval is deemed insufficient. While sharing the premise that adaptive decisions enhance factuality and robustness, these approaches are mostly evaluated on general-purpose benchmarks, in English, and with architectures not directly comparable to local institutional chatbots.

Parallel work also highlights how contextual noise and evidence selection can degrade retrieval-based systems. Cuconasu et al. [2024] show that response quality may decline even when retrieval performance improves, as increased context often amplifies semantic noise. This phenomenon is especially critical in regulatory domains, where correct answers demand high normative precision and low tolerance for free inference.

Within local and institutional settings, prior empirical analyses have demonstrated that response quality and system cost in RAG pipelines are highly query-dependent and that retrieval or correction mechanisms do not produce uniform gains across categories. Building on these insights, the present study focuses specifically on critic-based self-evaluation, comparing deterministic critics, critics augmented with LLM judges, and a simulated conditional policy under identical local constraints.

3. Methodology

3.1. System Architecture

The evaluated system is an institutional academic chatbot executed entirely in a local environment. Its architecture integrates vector retrieval, lexical retrieval, rank fusion, generation with a small-scale LLM, and an optional critic-based self-evaluation component. The corpus consists of institutional documents used for academic assistance in a public higher education institution.

The critic operates in two modes. The first is deterministic, checking for signals such as insufficient evidence coverage, unsupported numerical or normative claims, and low lexical alignment with retrieved context. The second mode augments these checks with an LLM-based judge that evaluates consistency between the generated answer and the retrieved evidence. When the critic rejects a response, the system may trigger a broader re-retrieval strategy. Specifically, the deterministic critic rejects a response when at least one of the following conditions is met: (i) lexical overlap between the generated answer and the retrieved chunks falls below a minimum threshold, measured by token-level intersection; (ii) the answer contains numerical or normative claims absent from the retrieved passages; or (iii) there is a citation mismatch between generated claims and retrievable context. Upon rejection, the re-retrieval strategy expands the query scope by increasing k and relaxing semantic similarity thresholds, injecting additional retrieved passages into the context window before a second generation attempt.

3.2. Experimental Design

Five configurations are evaluated. Configuration C0 corresponds to the baseline without critic. C1 activates only the deterministic critic. C2 combines the deterministic critic with a same-family LLM judge. C2′ combines the deterministic critic with a cross-family LLM judge. C3 represents a simulated oracle-based conditional policy that applies C1 to regulation queries and C0 to procedure and complex queries.

The evaluation dataset contains 149 queries across five categories; the main analysis focuses on regulation, procedure, and complex queries, while chitchat and out-of-scope queries are retained only for operational execution. The dataset was manually constructed by the authors from representative academic assistance interactions. Regulation queries demand normative precision (deadlines, requirements, rules); procedure queries require actionable step-by-step guidance; complex queries combine multiple categories or cross-reference documents. Distribution: 60 regulation, 49 procedure, 20 complex, 10 chitchat, and 10 out-of-scope queries, all reflecting authentic institutional needs with no synthetic or external-benchmark items. Category assignments were performed jointly by two annotators through discussion and consensus rather than independent classification, so inter-annotator agreement was not quantitatively recorded. The RAG corpus comprises four institutional documents totaling 105 pages (249,421 characters, ~139 chunks under the configuration of Section 3.3): the general regulation (31p), internship manual (63p), disciplinary code (7p), and glossary (4p). All 149 queries were elaborated from these documents, ensuring a supported answer exists in the corpus for each.

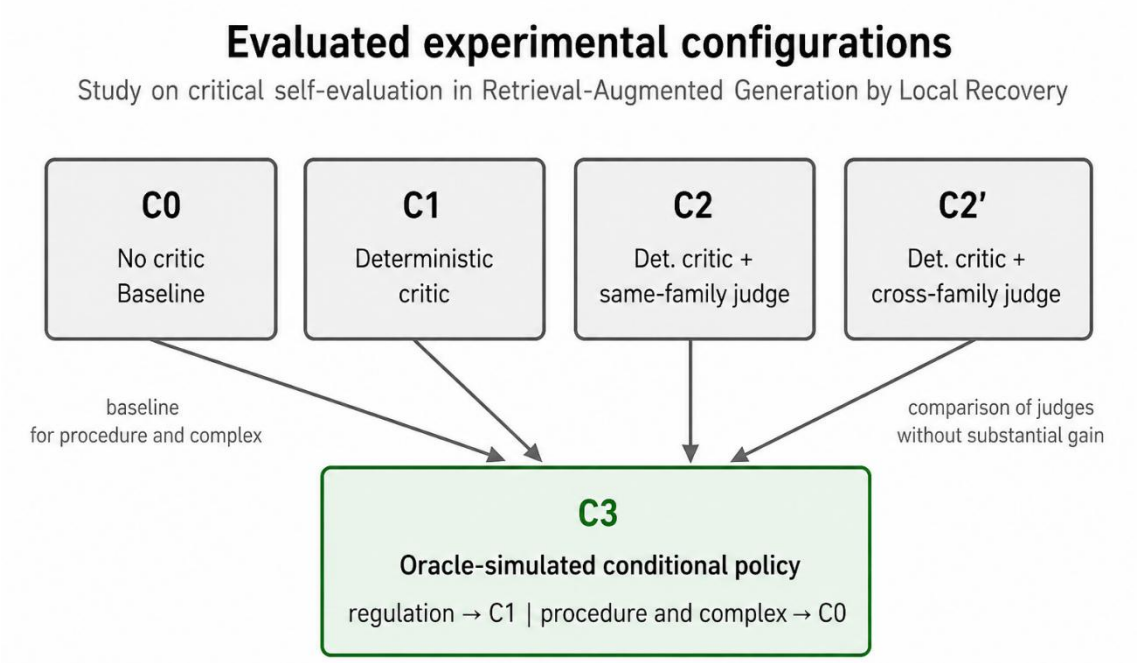


Figure 1. Overview of the experimental configurations evaluated in the study.

While Figure 1 illustrates the conceptual flow and the relationship between the experimental variants, the specific technical parameters, such as the model family used as the judge and the respective activation policies are detailed in Table 1.

Table 1. Evaluated experimental configurations.

Configuration	Critic Mechanism	Additional Judge	Activation Policy	Execution
C0	Disabled	Not applicable	No critic	Real
C1	Deterministic	Not applicable	All RAG queries	Real
C2	Deterministic	Same-family model (Gemma 3:4B)	All RAG queries	Real
C2′	Deterministic	Cross-family model (Llama 3.2:3B)	All RAG queries	Real
C3	Oracle-conditional	Not applicable	Regulation uses C1; procedure and complex use C0	Simulated post hoc

3.3 Experimental Setup

To ensure reproducibility, all experiments were executed under fixed hardware and software conditions. Experiments were conducted on a workstation equipped with an AMD Ryzen 5 5500G processor, 16 GB RAM, and an NVIDIA GeForce GTX 1050 GPU with 4 GB VRAM, running Ubuntu Server 24.04 LTS. The embedding model used was nomic-embed-text-v1.5. Documents were segmented into chunks of 512 tokens with 64-token overlap. Top-k retrieval was configured with k=5. All experiments used

temperature=0.2 and a maximum output length of 512 tokens. These parameters were held constant across all five configurations (C0–C3) to ensure that observed differences in quality and latency are attributable exclusively to the presence or absence of the critic mechanism. Response generation was performed using Gemma3:4B, executed locally through Ollama. The institutional corpus, evaluation queries, and generated responses were originally written in Brazilian Portuguese. RAGAS evaluation was performed using the same Gemma3:4B model as evaluator, with nomic-embed-text embeddings (via OllamaEmbeddings), using Portuguese (pt-BR) evaluation prompts. These settings were kept fixed across all evaluated configurations.

3.4 Evaluation Metrics

The quality of the responses was measured using the RAGAS framework across three metrics: Faithfulness, Answer Relevancy, and Context Precision. Answer Relevancy was adopted as the primary metric because it best proxies practical usefulness to the end user. Faithfulness alone can be achieved through overly conservative responses that, although technically grounded, fail to resolve the user's question. In institutional settings, users require actionable guidance rather than extractive reproduction of policy fragments. A response that scores high on Faithfulness by quoting normative text verbatim may still be operationally useless. Priority is therefore given to direct, actionable alignment with user needs, where an unhelpful answer carries the same practical cost as no answer at all.

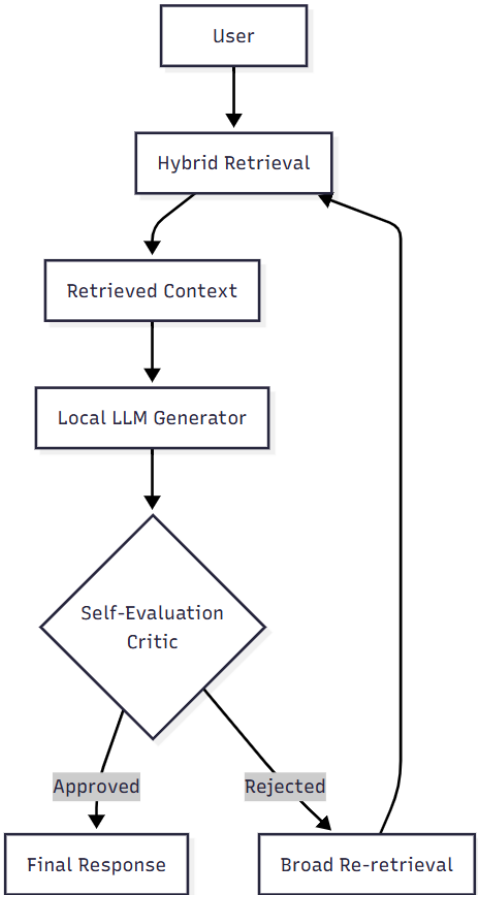


Figure 2. Architecture of the local RAG pipeline with a critical self-evaluation and re-retrieval cycle.

3.5 Statistical Analysis

In addition to quality metrics, end-to-end latency metrics were collected, especially the 95th percentile (p95) of total latency, as well as operational signals associated with the critic, such as the re-retrieval rate. Statistical analysis was conducted using the Wilcoxon signed-rank test, which is appropriate for paired samples without assuming normality, allowing stratified comparisons by query type. In addition to statistical significance, effect sizes were computed using the r metric ($r = Z / \sqrt{N}$) for each Wilcoxon signed-rank test, in order to assess the practical magnitude of observed differences beyond what p-values alone can indicate.

It is essential to note that configuration C3 is treated methodologically as an oracle-simulated conditional policy, consisting of a post hoc analytical construct. C3 does not represent a native end-to-end experiment executable in the backend, but rather an exploratory policy analysis intended to estimate the ideal balance between performance and latency. This distinction is crucial to ensure the internal validity of the study and to delimit the scope of the simulated results in relation to the actual executions (C0 to C2').

4. Results

Table 2 summarizes the overall results of the five configurations. The baseline C0 achieved Faithfulness of 0.3529, Answer Relevancy of 0.1688, Context Precision of 0.4767, and p95 latency of 35,050.36 ms. Configuration C1 increased Faithfulness to 0.4443, but reduced Answer Relevancy to 0.0841 and raised p95 latency to 48,232.16 ms. Configurations C2 and C2' followed the same overall pattern, with no global gain in the primary metric and an even greater increase in latency, reaching approximately 70 to 72 seconds at the 95th percentile. The conditional policy C3 fell between the extremes, with Answer Relevancy of 0.1129 and p95 latency of 41,342.33 ms.

These figures show that the always-on critic did not produce a global quality gain within the evaluated scope. The improvement observed in Faithfulness under the critic-enabled configurations did not translate into an aggregate improvement in Answer Relevancy. In practice, the system became more tightly grounded in the retrieved context in some cases but produced answers that were less useful for the user's question and incurred a substantially higher time cost.

Table 2. Overall quality and latency results by configuration.

Configuration	Faithfulness	Answer Relevancy	Context Precision	p95 Latency
C0	0.3529	0.1688	0.4767	35,050.36
C1	0.4443	0.0841	0.4684	48,232.16
C2	0.3993	0.0765	0.4686	70,155.57
C2′	0.4175	0.0748	0.4671	71,967.77
C3	0.3910	0.1129	0.4258	41,342.33

The analysis by query type clarifies why the aggregate view is insufficient. In regulation, C0 achieved an Answer Relevancy of 0.2410, whereas C1, C2, C2', and C3 remained close to 0.12. In procedure, C0 and C3 reached 0.0646, outperforming C1, C2,

and C2’. In complex, C0 and C3 achieved 0.2077, while the other variants ranged from 0.0550 to 0.0810. Thus, the degradation caused by the critic was particularly visible in complex and regulatory queries, which are precisely the most sensitive in the institutional context.

The Wilcoxon tests reinforce this interpretation. In regulation, the Answer Relevancy of C0 was significantly higher than that of C1, C2, C2’, and C3, surviving Holm correction. In procedure, gains in Faithfulness were observed for C1, C2, and C2’ compared with C0 and C3, as well as an increase in Context Precision for C2 and C2’, though after Holm correction only C1 versus C0 remains significant. In complex, C0 and C3 outperformed C2 and C2’ in Answer Relevancy, while the comparison between C0 and C1 fell into a borderline zone and should not be treated as strong evidence. These findings suggest that the critic changed the nature of the generated response, favoring a type of conservatism that does not always translate into greater perceived usefulness.

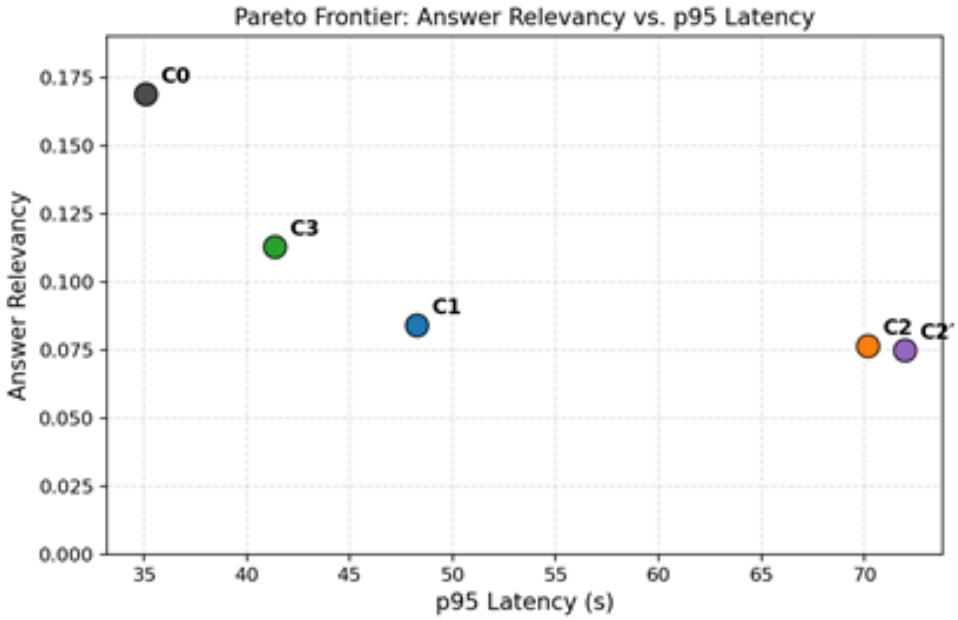


Figure 3. Global Pareto frontier between Answer Relevancy and p95 latency.

Figure 3 summarizes the global trade-off between Answer Relevancy and p95 latency across all configurations. In this space, the only point on the Pareto frontier is C0: no critic-enabled configuration simultaneously outperforms the baseline in both quality and cost. C3 occupies an intermediate position, offering a better compromise than the always-on variants without recovering the baseline's global superiority.

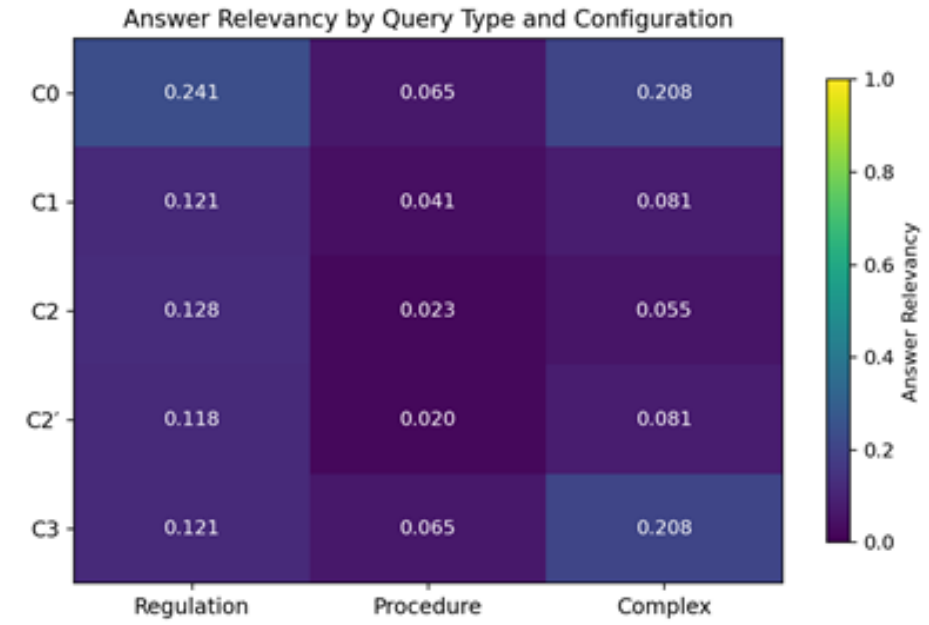


Figure 4. Answer Relevancy by query type and configuration.

Figure 4 reveals the heterogeneity of the critic's effect across query types. C3 inherits the behavior of C1 in regulation queries and of C0 in procedure and complex queries, which improves the trade-off relative to always-on configurations but does not restore the global superiority of the baseline.

As an operational complement, Figure 5 relates the critic's average cost and the re-retrieval rate by query type. The chart shows that the judge-based variants substantially increase the cost of the review mechanism without producing a proportional gain in the study's primary metric.

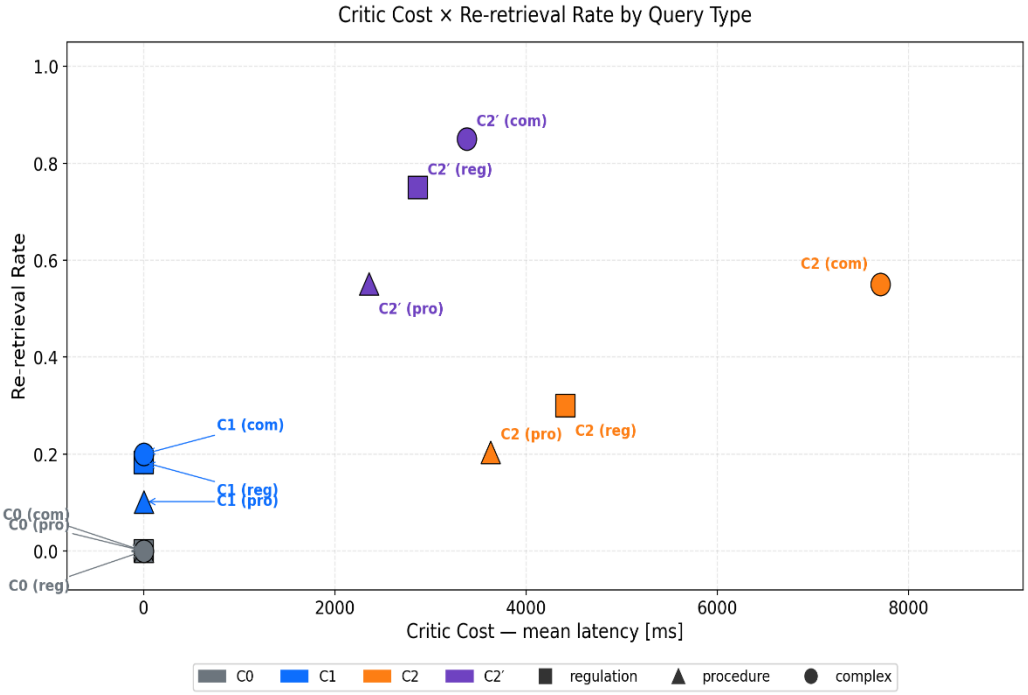


Figure 5. Critic cost and re-retrieval rate by query type.

Statistical validation is needed to confirm whether the disparities observed in Figure 5 are significant; Table 3 details the paired Wilcoxon comparisons, highlighting where the critic produced statistically significant effects by query category.

Table 3. Main statistical comparisons stratified by query type.

Query Type	Metric	Comparison	n	p-value*	r	Magnitude
Regulation	Answer Relevancy	C0 > C1	60	0.0096	0.43	medium
Regulation	Answer Relevancy	C0 > C2	60	0.0100	0.43	medium
Regulation	Answer Relevancy	C0 > C2′	60	0.0153	0.41	medium
Regulation	Answer Relevancy	C0 > C3	60	0.0096	0.43	medium
Regulation	Faithfulness	C1 > C2	60	0.1176	0.26	small
Procedure	Faithfulness	C1 > C0	48	0.0448	0.40	medium
Procedure	Faithfulness	C2 > C0	48	0.1176	0.34	medium
Procedure	Faithfulness	C2′ > C0	48	0.1176	0.31	medium
Procedure	Context Precision	C2 > C0	49	0.1176	0.30	medium
Procedure	Context Precision	C2′ > C0	49	0.1176	0.33	medium
Complex	Answer Relevancy	C0 > C2	20	0.0693	0.58	large
Complex	Answer Relevancy	C0 > C2′	20	0.1176	0.46	medium

Note. C0-C1 in complex queries showed borderline significance ($p = 0.0505$) and is discussed only in the text. Effect size $r = Z/\sqrt{N}$; magnitude thresholds follow Cohen (1992): small ≥ 0.1, medium ≥ 0.3, large ≥ 0.5. *p-values adjusted via Holm correction across the 12 comparisons; only Regulation Answer Relevancy and Procedure Faithfulness (C1 vs. C0) remain significant at $\alpha=0.05$. One Procedure query was excluded from Faithfulness (48 vs. 49 observations) due to an off-topic response (Context Precision = 0.0).

5. Discussion

The findings reveal three structural effects. First, Faithfulness and Answer Relevancy diverged: critic-enabled configurations increased Faithfulness while reducing Answer Relevancy, our automated proxy for response utility. This indicates over-conservatism: acting as a rigid filter, the critic suppresses useful inference and shifts the system toward

forced literalism, severely harming practical relevance in precisely the regulatory contexts where reliability matters most.

The second effect concerns the failure of LLM-based judges (C2 and C2′) to resolve the problem: the bottleneck lies in the corrective mechanics itself. Broader re-retrieval injects additional documents, amplifying semantic noise that the small-scale local LLM, operating near its capacity limit, cannot separate from signal in the already dense normative context, rendering correction counterproductive and often worse than C0. This is consistent with the noise-amplification effect reported by Cuconasu et al. [2024], and is especially acute in normative corpora, where retrieved chunks are structurally similar and semantically dense.

Third, C3 remains scientifically meaningful without outperforming the baseline globally, showing that selective routing mitigates the damage of uniform critic activation and confirms query-type-aware policies as a viable design direction; this is illustrated concretely by individual query outcomes, where in regulation queries C0 produces direct procedural answers while C1 and C2 generate evasive literal quotations that omit the interpretive step, and in complex queries C0 synthesizes coherent answers while critic-enabled configurations produce fragmented chunk-level responses.

Finally, within constrained local settings, the problems of over-conservatism and noise amplification are likely to recur in other normative domains (legal, financial, corporate compliance) where document bases are similarly dense and tolerance for error is low. For such environments, RAG pipelines should incorporate prior routing policies rather than universal critic layers.

5.1. Threats to Validity

Three threats to validity should be noted. Internally, C3 is an oracle-simulated conditional policy rather than a native end-to-end experiment, so its results estimate potential rather than proven production performance; it should be read as an upper bound for query-type-aware routing, since a deployable classifier would introduce classification errors and latency overhead not modeled here. For construct validity, automated RAGAS metrics carry evaluator-inherent bias, and Faithfulness may misalign with actual user-perceived utility under over-conservative behavior, an effect potentially amplified by using the same model (Gemma3:4B) for generation, the C2 judge, and RAGAS evaluation. Externally, findings are bounded to a single institutional chatbot and normative corpus, and the observed degradation may partly reflect the limited synthesis capacity of small-scale models; results should therefore not be read as evidence that critic-based correction is universally ineffective, but that its benefits may not transfer to constrained local deployments with lightweight generators.

6. Conclusion

This paper empirically evaluated critic-based self-correction in a local RAG institutional chatbot across five configurations: a baseline (C0), a deterministic critic (C1), two LLM-judge variants (C2 and C2′), and an oracle-simulated conditional policy (C3). In constrained-hardware environments with small-scale models, uniform critic activation did not function as a universal quality layer: it induced over-conservatism, raising Faithfulness at the cost of reduced practical utility and substantially higher latency. That C0 remained the best overall configuration challenges common assumptions in the

adaptive RAG literature, showing that pipeline sophistication must be balanced against the local generator's synthesis capacity.

As future work, we recommend implementing the conditional policy natively for end-to-end evaluation, refining activation logic based on retrieval-uncertainty signals, and investigating how larger local models (>7B parameters) respond to the same critical review mechanism.

7. References

Asai, A., Wu, Z., Wang, Y., Sil, A. and Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the International Conference on Learning Representations.

Cohen, J. (1992). A Power Primer. Psychological Bulletin, 112(1), 155-159.

Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N. and Silvestri, F. (2024). The Power of Noise: Redefining Retrieval for RAG Systems. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Es, S., James, J., Anke, L. E. and Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 150-158.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. and Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Jeong, S., Baek, J., Cho, S., Hwang, S. J. and Park, J. C. (2024). Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Yu, J. D., Yang, Y., Callan, J. and Neubig, G. (2023). Active Retrieval Augmented Generation. arXiv:2305.06983.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S. and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems 33.

Neves, J. E. D. A.; Pedro, P. S. M.; Hernandez, M. F. G.; Fabri Junior, L. A. Simulation of the Implementation of Domestic Solar Systems Using Multi-agent Systems from Web Scraping. Smart Innovation, Systems and Technologies. 1ed.: Springer International Publishing, 2023, v. 1, p. 88-96.

Yan, S.-Q., Gu, J.-C., Zhu, Y. and Ling, Z.-H. (2024). Corrective Retrieval Augmented Generation. arXiv:2401.15884.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E. and Stoica, I. (2024). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems 37.