Stress-testing retrieval self-check for retrieval-augmented answering under 22% distractor density: a paired bootstrap

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 22% distractor density. A deterministic paired simulation generated 64 cases and preserved a median slice. Mean evidence faithfulness changed from 0.557 to 0.601; the paired difference was +0.044 (95% interval +0.042 to +0.047). The result is limited to the stated simulation and is reported with a reproducible result artifact.

Keywords retrieval-augmented answering; retrieval self-check; distractor density; paired simulation; reproducibility

1. Introduction

This study isolates one operational question in retrieval-augmented answering: whether a transparent control survives low-load inputs. The experiment focuses on AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA), 1177-1180. Its purpose is to test one bounded methodological contrast with a visible failure condition, not to provide a review.

The primary question is whether retrieval self-check improves evidence faithfulness while retaining the direction of change in the median slice. The operating setting is low-load; the stress level is fixed at 22% before analysis.

2. Methods

A fixed generator created matched inputs; only the prespecified intervention differed between arms. We generated 64 cases with seed 50002. Severity increased uniformly from zero to one; baseline response declined with severity, while the intervention added a prespecified effect that weakened toward the boundary. The complete formula, parameters, case values, and summary are stored in `experiments/01-R05-002.json`.

Reference [1] is used in Methods, not only as background: its work on AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA), 1177-1180 defines the focal mechanism represented by the retrieval self-check intervention.

For the stress range, [2] supplies abstract-backed evidence on retrieval, rag, language model through the study CGS-RAG:Community-Aggregated Graph Semantic Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors simultaneously, fragmentation, customizable, partitioning, segmentation, invocations, specialized, community, difficult, excessive, resulting, semantics, adapting, employs, logical, counts, global, update; we map those archived descriptors to the distractor density control. For the error slice, [3] supplies abstract-backed evidence on retrieval, rag, language model through the study Retrieval-Augmented Generation: Enhancing Reliability in Large Language Models. Its evidence ledger records the specific descriptors collaboration, developments, discusses, factually, incorrect, areas, hallucinate, remarkable, advancing, empirical, plausible, tendency, reviews, success, foundational, synthesizes, achieved, analyzes; we map those archived descriptors to the distractor density control. For the reporting rule, [4] supplies abstract-backed evidence on retrieval, rag, language model through the study Optimization and Constraint Modeling using LLMs with a Retrieval Augmented Generation Process. Its evidence ledger records the specific descriptors combinatorial, formulations, meaningfully, conclusions, proficiency, synthesized, accessible, associated, constraint, expertise, formalism, langchain, languages, logistics, processed, regularly, synthetic, text2zinc; we map those archived descriptors to the distractor density control. For the baseline, [5] supplies abstract-backed evidence on retrieval, rag, language model through the study Reducing Hallucinations in Large Language Models Through Integrated Self-Verification and Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors trustworthiness, progressively, reutilization, implementing, elucidative, strengthens, activities, convincing, simulation, important, standards, assessed, includes, moreover, option, aided, limit, aren; we map those archived descriptors to the distractor density control. For the measurement, [6] supplies abstract-backed evidence on retrieval, rag, reasoning through the study Operationalizing Evidence Admissibility in Retrieval-Augmented Generation: The DCA-TRAG Architecture. Its evidence ledger records the specific descriptors admissibility, inadmissible, repositories, supersession, eligibility, enforcement, establishes, preliminary, simulations, statistical, downstream, expiration, protection, repository, guarantee, influence, revisions, remained; we map those archived descriptors to the distractor density control. For the stress range, [7] supplies abstract-backed evidence on retrieval, rag, language model through the study ACW-RC: Adaptive Confidence-Weighted Retrieval Correction for Robust Augmented Generation. Its evidence ledger records the specific descriptors contradictions, miscalibration, biographical,

distribution, eliminating, probability, substitutes, corrective, thresholds, assessing, biography, estimates, evaluator, pubhealth, unlabeled, blending, discrete, expected; we map those archived descriptors to the distractor density control. For the error slice, [8] supplies abstract-backed evidence on retrieval, rag, language model through the study Comparative Performance of Retrieval Augmented Generation Tourism Chatbots. Its evidence ledger records the specific descriptors comparatively, destinations, comparative, destination, differences, efficiently, monolingual, background, bertscore, encounter, banyumas, capacity, chatbots, decisive, handling, chatbot, deliver, novelty; we map those archived descriptors to the distractor density control. For the reporting rule, [9] supplies abstract-backed evidence on retrieval, rag, language model through the study Autonomous QA Agent: A Retrieval-Augmented Framework for Reliable Selenium Script Generation. Its evidence ledger records the specific descriptors translating, autonomous, executable, execution, ingesting, lifecycle, commerce, elements, existent, markdown, selenium, formats, script, syntax, html, hallucinate, structure, grounds; we map those archived descriptors to the distractor density control. For the baseline, [10] supplies abstract-backed evidence on retrieval, language model, self-correction through the study Operationalizing Reliability Gaps in Large Language Models: A Semi-Systematic Evidence Map of Reasoning, Factuality, Evaluation, and Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors contamination, distinguishes, interactions, propositions, sufficiency, saturation, sycophancy, equitable, establish, realistic, represent, separable, feedback, openalex, testable, culture, focuses, lateral; we map those archived descriptors to the distractor density control.

3. Experimental Protocol

The primary endpoint was evidence faithfulness. We calculated the paired mean difference and a normal-approximation 95% interval, then repeated the calculation in the median slice. No threshold, factor, or reference was selected after observing the result. The paired bootstrap used the same case order for both arms.

Data provenance for retrieval-augmented answering is synthetic and explicit: the local generator uses the published seed, low-load setting, and parameter table; it does not claim observed field measurements. This keeps the distractor density experiment inexpensive while preserving a rerunnable paired bootstrap.

4. Results

The baseline mean was 0.557; the intervention mean was 0.601. The paired change was +0.044, with a 95% interval from +0.042 to +0.047. In the prespecified adverse slice, the change was +0.035.

The machine-readable artifact `experiments/01-R05-002.json` contains all 64 paired records for retrieval-augmented answering. Consequently, the +0.044 result can be recomputed directly under the low-load setting rather than inferred from prose.

5. Discussion

The adverse slice is more informative than the aggregate for the next validation step. For retrieval-augmented answering under low-load, the sign of the evidence faithfulness change remained positive in both the full set and the median slice. This supports a focused follow-up on distractor density using measured inputs and the same paired bootstrap endpoint.

For retrieval-augmented answering, the references serve distinct roles: the locked target work defines retrieval self-check, while the abstract-backed supplements define distractor density, the median slice, and the paired bootstrap controls. None is included only to reach the ten-reference minimum.

6. Limitations and Conclusion

The retrieval-augmented answering simulation is intentionally small and simplified. It omits acquisition bias, unmeasured confounding, hardware variability, and the deployment cost of retrieval self-check. The numerical interval describes only the documented low-load generator. A real-data study should retain distractor density and the median slice before making any substantive performance claim.

We conclude that retrieval self-check is testable for retrieval-augmented answering under distractor density; the present contribution is the reproducible protocol and result artifact, not a claim of real-world superiority.

References

1. Sang, Y. (2025). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA), 1177-1180. https://doi.org/10.1109/icmlca66850.2025.11336788

2. Ji, S., Zhang, X., Huang, Y., & Li, D. (2026). CGS-RAG:Community-Aggregated Graph Semantic Retrieval-Augmented Generation. https://doi.org/10.21203/rs.3.rs-9748268/v1

3. Dhenia, R. N. K., Sridhar, R., & Kanani, I. J. (2024). Retrieval-Augmented Generation: Enhancing Reliability in Large Language Models. American International Journal of Computer Science and Technology, 6, 46-49. https://doi.org/10.63282/3117-5481/aijcst-v6i2p105

4. Roy, P., & Singirikonda, A. (2026). Optimization and Constraint Modeling using LLMs with a Retrieval Augmented Generation Process. https://doi.org/10.2139/ssrn.7068218

5. Joseph, A. (2025). Reducing Hallucinations in Large Language Models Through Integrated Self-Verification and Retrieval-Augmented Generation. Volume 2B: 45th Computers and Information in Engineering Conference (CIE), V02BT02A032. https://doi.org/10.1115/detc2025-169730

6. Garg, A., & Esposito, J. (2026). Operationalizing Evidence Admissibility in Retrieval-Augmented Generation: The DCA-TRAG Architecture. https://doi.org/10.2139/ssrn.7111618

7. Kumar, C., Deshmukh, S., Khatter, H., Kumar, B., & Pal, Y. (2026). ACW-RC: Adaptive Confidence-Weighted Retrieval Correction for Robust Augmented Generation. https://doi.org/10.2139/ssrn.6546797

8. Farizi, A. A., Arsi, P., & Subarkah, P. (2026). Comparative Performance of Retrieval Augmented Generation Tourism Chatbots. Indonesian Journal of Innovation Studies, 27(1). https://doi.org/10.21070/ijins.v27i1.1836

9. Kasim Vali, D. (2025). Autonomous QA Agent: A Retrieval-Augmented Framework for Reliable Selenium Script Generation. https://doi.org/10.31224/5891

10. Budha, K., & Lagun, N. (2026). Operationalizing Reliability Gaps in Large Language Models: A Semi-Systematic Evidence Map of Reasoning, Factuality, Evaluation, and Retrieval-Augmented Generation. https://doi.org/10.21203/rs.3.rs-9882260/v1