Auditing retrieval self-check for retrieval-augmented answering under 22% distractor density: a sensitivity sweep

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 22% distractor density. A deterministic paired simulation generated 72 cases and preserved a rare-condition slice. Mean evidence faithfulness changed from 0.496 to 0.538; the paired difference was +0.041 (95% interval +0.039 to +0.044). 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

The design begins from a narrow failure mode in retrieval-augmented answering rather than a broad literature survey. The experiment focuses on Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation. 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), 397-400. 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 rare-condition slice. The operating setting is long-tail; the stress level is fixed at 22% before analysis.

2. Methods

We used a deterministic severity sweep and retained identical noise draws for the primary contrast. We generated 72 cases with seed 50131. 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-R05-131.json`.

Reference [1] is used in Methods, not only as background: its work on Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation. 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), 397-400 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 Privacy-Preserving Retrieval-Augmented Generation on Local Devices for Regenerative Medicine Applications. Its evidence ledger records the specific descriptors confidentiality, differentiation, commercially, consultation, institutions, regenerative, compensates, constructed, clinically, deployable, hepatocyte, inevitably, linguistic, unsuitable, computing, embryonic, inquiries, paramount; 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 A Survey of (Deep RAG) Deep Retrieval Augmented Generation and Reasoning in Large Language Models. Its evidence ledger records the specific descriptors instantiation, categorizing, supervision, performing, supervised, according, designing, paradigms, flexible, planning, referred, clarify, roadmap, reinforcement, multimodal, verifiable, landscape, workflows; 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 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 baseline, [5] supplies abstract-backed evidence on retrieval, rag, language model through the study Retrieval-Augmented Generation in Large Language Models through Selective Augmentation. Its evidence ledger records the specific descriptors preprocessing, rigorously, carefully, confirmed, exhibited, pertinent, selective, profound, elevate, rouge, bleu, augmentation, technologies, advancement, contributes, implemented, substantial, algorithm; we map those archived descriptors to the distractor density control. For the measurement, [6] 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 stress range, [7] supplies abstract-backed evidence on retrieval, rag, language model through the study Enhancing Context-Aware Search with Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors cryptography, enterprises, exploratory, experiment, government, industries, variations, functions, concepts, enriched, explored, increase, concept, digital, finance, matches,

utilize, faster; 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 Adaptive Context Reconstruction for Enhanced Retrieval Augmented Generation . Its evidence ledger records the specific descriptors reconstruction, restructuring, inferential, dependency, likelihood, reordering, statements, transforms, condensed, discourse, indicates, reduction, redundant, arranges, backbone, consists, rewrites, salience; 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 A Theoretical Analysis of Self-Contained Retrieval-Augmented Generation with Small Language Models. Its evidence ledger records the specific descriptors recommendations, compressing, theoretical, connection, contained, expensive, translate, concerns, concrete, internet, required, together, affects, bounded, develop, finding, helping, amount; 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 rare-condition slice. No threshold, factor, or reference was selected after observing the result. The sensitivity sweep 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, long-tail setting, and parameter table; it does not claim observed field measurements. This keeps the distractor density experiment inexpensive while preserving a rerunnable sensitivity sweep.

4. Results

The baseline mean was 0.496; the intervention mean was 0.538. The paired change was +0.041, with a 95% interval from +0.039 to +0.044. In the prespecified adverse slice, the change was +0.034.

The machine-readable artifact `experiments/01-R05-131.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.041 result can be recomputed directly under the long-tail setting rather than inferred from prose.

5. Discussion

Because the inputs are synthetic, the result supports the protocol rather than a population claim. For retrieval-augmented answering under long-tail, the sign of the evidence faithfulness change remained positive in both the full set and the rare-condition slice. This supports a focused follow-up on distractor density using measured inputs and the same sensitivity sweep 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 rare-condition slice, and the sensitivity sweep 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 long-tail generator. A real-data study should retain distractor density and the rare-condition 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). Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation. 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), 397-400. https://doi.org/10.1109/raiic65850.2025.11170170

2. Takamura, T., & Umezawa, A. (2025). Privacy-Preserving Retrieval-Augmented Generation on Local Devices for Regenerative Medicine Applications. https://doi.org/10.1101/2025.10.20.25337146

3. Liu, L. (2026). A Survey of (Deep RAG) Deep Retrieval Augmented Generation and Reasoning in Large Language Models. https://doi.org/10.36227/techrxiv.177272838.89432844/v1

4. 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

5. Quintela, J., & Sapateiro, M. (2024). Retrieval-Augmented Generation in Large Language Models through Selective Augmentation. https://doi.org/10.21203/rs.3.rs-4652959/v1

6. 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

7. Shetty, R. (2025). Enhancing Context-Aware Search with Retrieval-Augmented Generation. https://doi.org/10.36227/techrxiv.174060240.09460752/v1

8. Köhler, A., Stein, M., & Vogt, E. (2026). Adaptive Context Reconstruction for Enhanced Retrieval Augmented Generation . https://doi.org/10.2139/ssrn.6101386

9. Samal, M. P. (2026). A Theoretical Analysis of Self-Contained Retrieval-Augmented Generation with Small Language Models. https://doi.org/10.36227/techrxiv.177219989.96070478/v1

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