Auditing retrieval self-check for retrieval-augmented answering under 8% distractor density: a hold-out check

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 8% distractor density. A deterministic paired simulation generated 72 cases and preserved a rare-condition slice. Mean evidence faithfulness changed from 0.539 to 0.588; the paired difference was +0.049 (95% interval +0.046 to +0.051). 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 Robustness of Fine-Tuned Llms Under Noisy Retrieval Inputs. 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), 417-420. 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 noisy-input; the stress level is fixed at 8% 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 50043. 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-R05-043.json`.

Reference [1] is used in Methods, not only as background: its work on Robustness of Fine-Tuned Llms Under Noisy Retrieval Inputs. 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), 417-420 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 A2JRAG: Process-Aware Retrieval-Augmented Generation for Public Legal Information Systems. Its evidence ledger records the specific descriptors entityrelationship, representation, twodimensional, demonstration, substantively, procedurally, conditioned, proceedings, transitions, escalation, misaligned, relational, formalize, mandatory, traverses, typically, authored, carolina; 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 for building datasets from scientific literature. Its evidence ledger records the specific descriptors necessitates, publications, accordingly, abstracts, adjusted, predict, closed, growth, remove, ready, h2wt, hyst, quantization, automation, extraction, employing, hydrides, hydrogen; 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 Graph Retrieval-Augmented Generation for Large Language Models: A Survey. Its evidence ledger records the specific descriptors ameliorating, representing, applicable, imperative, obfuscates, scientists, vectorized, generally, intending, engender, possible, concise, contain, desired, insofar, massive, prompts, snippet; 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 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 measurement, [6] 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 stress range, [7] 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 error slice, [8] 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 reporting rule, [9] supplies abstract-backed evidence on retrieval, rag, language model through the study GIANT: Leveraging Retrieval-Augmented Generation for Automated Bug Reproduction Test Generation. Its evidence ledger records the specific descriptors construction, reproduction, comparisons, conventions, procedures, reproduced, reproduces, submission, underlying, assurance, defects4j, exhausted, examined, isolated, produced, projects, reported, defects; we map those archived descriptors to the distractor density control. For the baseline, [10] supplies abstract-backed evidence on retrieval, rag, language model through the study Retrieval-Augmented Text Generation: Methods, Challenges, and Applications. Its evidence ledger records the specific descriptors vulnerabilities, democratizing, interpretable, opportunities, practitioners, summarization, demonstrated, obsolescence, synthesizing, unstructured, furthermore, researchers, inherently, principles, continual, continues, outlining, bridging; 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 hold-out check 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, noisy-input setting, and parameter table; it does not claim observed field measurements. This keeps the distractor density experiment inexpensive while preserving a rerunnable hold-out check.

4. Results

The baseline mean was 0.539; the intervention mean was 0.588. The paired change was +0.049, with a 95% interval from +0.046 to +0.051. In the prespecified adverse slice, the change was +0.042.

The machine-readable artifact `experiments/01-R05-043.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.049 result can be recomputed directly under the noisy-input 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 noisy-input, 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 hold-out check 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 hold-out check 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 noisy-input 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). Robustness of Fine-Tuned Llms Under Noisy Retrieval Inputs. 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), 417-420. https://doi.org/10.1109/aiea66061.2025.11160531

2. Martin, T., & Gilchrist, S. (2026). A2JRAG: Process-Aware Retrieval-Augmented Generation for Public Legal Information Systems. https://doi.org/10.2139/ssrn.7114139

3. Maharana, P. R., Verma, A., & Joshi, K. (2025). Retrieval augmented generation for building datasets from scientific literature. https://doi.org/10.26434/chemrxiv-2024-qjx32-v2

4. Procko, T., & Ochoa, O. (2024). Graph Retrieval-Augmented Generation for Large Language Models: A Survey. https://doi.org/10.2139/ssrn.4895062

5. 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

6. 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

7. 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

8. 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

9. Bo, L., Duan, R., Shi, S., Sun, X., Lin, Z., & Ji, W. (2026). GIANT: Leveraging Retrieval-Augmented Generation for Automated Bug Reproduction Test Generation. https://doi.org/10.2139/ssrn.7269657

10. Genesis, J. (2025). Retrieval-Augmented Text Generation: Methods, Challenges, and Applications. https://doi.org/10.20944/preprints202504.0443.v1