Auditing retrieval self-check for retrieval-augmented answering under 30% distractor density: a factor ablation

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 30% distractor density. A deterministic paired simulation generated 72 cases and preserved a low-signal stratum. Mean evidence faithfulness changed from 0.497 to 0.545; the paired difference was +0.048 (95% interval +0.046 to +0.050). 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 low-signal stratum. The operating setting is long-tail; the stress level is fixed at 30% 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 50083. 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-R05-083.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 Long Text Language Ensemble Models: Extension of Retrieval-Augmented Generation (RAG). Its evidence ledger records the specific descriptors philosophical, alternatives, centricity, education, extending, extension, segmented, transform, ensemble, explores, networks, ethical, perform, reality, virtual, agency, agents, extend; 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 Dynamic Multi-Hop Retrieval-Augmented Generation Framework for Professional Domain Question Answering. Its evidence ledger records the specific descriptors completeness, structuring, critically, inadequate, surpassing, belief, upon, coherence, financial, incurring, validate, towards, stakes, professional, consistency, evaluations, benchmarks, factuality; 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 Pool-Gated Retrieval: Beyond Retrieval-Augmented Generation Toward Accountable Evidential Admission. Its evidence ledger records the specific descriptors accountability, epistemically, justificatory, accountable, recognition, commitment, preventing, principled, propagated, registered, structural, widespread, admission, committed, declaring, enforcing, epistemic, exclusive; 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 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 measurement, [6] supplies abstract-backed evidence on retrieval, rag, language model through the study Is Retrieval-Augmented Generation Fair Use? . Its evidence ledger records the specific descriptors technologically, copyrights, summarizes, ubiquitous, unexamined, companies, copyright, describes, impactful, infringes, similarly, therefore, conducts, doctrine, employed, lawsuits, legality, powering; 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 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 error slice, [8] supplies abstract-backed evidence on retrieval, rag, language model through the study A Brief Introduction to Retrieval Augmented Generation (RAG). Its evidence ledger records the specific descriptors introduction, omnipotent, dialogues, realize, shuster, aryani, brief, amir, inaccurate, increasing, people, generating, leading, various, application, scenarios, answers, hallucinations; 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 FaithGate: A Proposed Architecture for Real-Time Hallucination Detection and Correction in Retrieval-Augmented Generation Systems. Its evidence ledger records the specific descriptors verificationcorrection, pseudoalgorithmic, evaluationmetric, implementability, interdependent, contradictory, independently, specification, disagreement, disconnected, functionally, groundedness, deployments, identifying, regenerates, decomposer, dependence, entailment; 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 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.

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 low-signal stratum. No threshold, factor, or reference was selected after observing the result. The factor ablation 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 factor ablation.

4. Results

The baseline mean was 0.497; the intervention mean was 0.545. The paired change was +0.048, with a 95% interval from +0.046 to +0.050. In the prespecified adverse slice, the change was +0.042.

The machine-readable artifact `experiments/01-R05-083.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.048 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 low-signal stratum. This supports a focused follow-up on distractor density using measured inputs and the same factor ablation 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 low-signal stratum, and the factor ablation 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 low-signal stratum 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. Ibiloye, A. C. (2025). Long Text Language Ensemble Models: Extension of Retrieval-Augmented Generation (RAG). https://doi.org/10.31219/osf.io/b94ng_v1

3. Zhu, J., & Tong, Y. (2026). Dynamic Multi-Hop Retrieval-Augmented Generation Framework for Professional Domain Question Answering. https://doi.org/10.22541/au.177499050.00368942/v1

4. Kim, M. H. (2026). Pool-Gated Retrieval: Beyond Retrieval-Augmented Generation Toward Accountable Evidential Admission. https://doi.org/10.20944/preprints202606.0414.v1

5. 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

6. Glendenning, J. (2026). Is Retrieval-Augmented Generation Fair Use?. https://doi.org/10.2139/ssrn.6165167

7. 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

8. Aryani, A. (2024). A Brief Introduction to Retrieval Augmented Generation (RAG). https://doi.org/10.59350/ayxyj-h9751

9. Madhurima, M. (2026). FaithGate: A Proposed Architecture for Real-Time Hallucination Detection and Correction in Retrieval-Augmented Generation Systems. https://doi.org/10.2139/ssrn.7242858

10. 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