Probing retrieval self-check for retrieval-augmented answering under 9% distractor density: a error-stratified estimate

ABSTRACT

We evaluated retrieval self-check for retrieval-augmented answering under 9% distractor density. A deterministic paired simulation generated 56 cases and preserved a upper-severity quartile. Mean evidence faithfulness changed from 0.544 to 0.590; the paired difference was +0.046 (95% interval +0.044 to +0.049). 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

A simple paired experiment provides a low-cost check of retrieval self-check under distractor density. 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 upper-severity quartile. The operating setting is shifted-distribution; the stress level is fixed at 9% before analysis.

2. Methods

Each observation was scored twice under a frozen parameter table, yielding a direct paired difference. We generated 56 cases with seed 50037. 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-R05-037.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 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 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, [4] 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 baseline, [5] supplies abstract-backed evidence on retrieval, rag, language model through the study Building a Security and Reliability Evaluation Suite for Retrieval-Augmented Generation (RAG) Systems. Its evidence ledger records the specific descriptors illustrative, instruments, sacrificing, avoidance, tradeoffs, fairness, monitors, template, toxicity, compute, utility, siloed, hence, suite, miss, conditioning, standardized, calibration; 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 UAMG-RAG: Adaptive Multi-Granularity Retrieval-Augmented Generation with Uncertainty-Guided Re-Retrieval. Its evidence ledger records the specific descriptors insufficiency, predetermined, complicated, granularity, responsible, iterations, retrievals, depending, obtaining, dropping, suggests, covered, designs, efforts, optimal, removal, wikihow, guided; 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 MedRAG: Retrieval-Augmented Generation for Medical QA-Comparing Base and RAG-Augmented LLM Performance on Evidence from Peer-Reviewed Clinical Research. Its evidence ledger records the specific descriptors evidencegrounded, substitution, transformers, counterpart, numerically, percentage, statistics, comparing, confident, consensus, academic, chromadb, deployed, failures, sentence, keyword, medical, reveals; 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 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. For the reporting rule, [9] 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 baseline, [10] supplies abstract-backed evidence on retrieval, rag, language model through the study A Privacy-First Architecture for Fully Local Retrieval-Augmented Generation in Secure Document Intelligence. Its evidence ledger records the specific descriptors sentencetransformers, dependencies, reproducible, accelerated, delivers, ragstack, choices, locally, private, pymupdf, gapped, hosted, ollama, stores, strict, genai, refer, trade; 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 upper-severity quartile. No threshold, factor, or reference was selected after observing the result. The error-stratified estimate 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, shifted-distribution setting, and parameter table; it does not claim observed field measurements. This keeps the distractor density experiment inexpensive while preserving a rerunnable error-stratified estimate.

4. Results

The baseline mean was 0.544; the intervention mean was 0.590. The paired change was +0.046, with a 95% interval from +0.044 to +0.049. In the prespecified adverse slice, the change was +0.040.

The machine-readable artifact `experiments/01-R05-037.json` contains all 56 paired records for retrieval-augmented answering. Consequently, the +0.046 result can be recomputed directly under the shifted-distribution setting rather than inferred from prose.

5. Discussion

Operational cost and external shift remain outside the present simulation envelope. For retrieval-augmented answering under shifted-distribution, the sign of the evidence faithfulness change remained positive in both the full set and the upper-severity quartile. This supports a focused follow-up on distractor density using measured inputs and the same error-stratified estimate 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 upper-severity quartile, and the error-stratified estimate 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 shifted-distribution generator. A real-data study should retain distractor density and the upper-severity quartile 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. 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. Köhler, A., Stein, M., & Vogt, E. (2026). Adaptive Context Reconstruction for Enhanced Retrieval Augmented Generation . https://doi.org/10.2139/ssrn.6101386

4. 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

5. Roy, P., & Roy, D. (2025). Building a Security and Reliability Evaluation Suite for Retrieval-Augmented Generation (RAG) Systems. https://doi.org/10.20944/preprints202510.0418.v2

6. S, S. B., & N, K. (2026). UAMG-RAG: Adaptive Multi-Granularity Retrieval-Augmented Generation with Uncertainty-Guided Re-Retrieval. https://doi.org/10.21203/rs.3.rs-10020935/v1

7. Dhanda, K. (2026). MedRAG: Retrieval-Augmented Generation for Medical QA-Comparing Base and RAG-Augmented LLM Performance on Evidence from Peer-Reviewed Clinical Research. https://doi.org/10.2139/ssrn.6608518

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

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

10. Srivastava, M. (2026). A Privacy-First Architecture for Fully Local Retrieval-Augmented Generation in Secure Document Intelligence. https://doi.org/10.36227/techrxiv.176800894.46972585/v1