Stress-testing retrieval self-check for retrieval-augmented answering under 32% distractor density: a paired bootstrap

ABSTRACT

We evaluated retrieval self-check for retrieval-augmented answering under 32% distractor density. A deterministic paired simulation generated 64 cases and preserved a low-signal stratum. Mean evidence faithfulness changed from 0.522 to 0.575; the paired difference was +0.053 (95% interval +0.051 to +0.056). 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 noisy-input inputs. 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 low-signal stratum. The operating setting is noisy-input; the stress level is fixed at 32% before analysis.

2. Methods

A fixed generator created matched inputs; only the prespecified intervention differed between arms. We generated 64 cases with seed 50114. 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-R05-114.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 Evaluating hallucination Mitigation in Retrieval-Augmented Generation Systems through Guarded Response Strategies, Embedding Comparison, and Retrieval Depth Analysis. Its evidence ledger records the specific descriptors selectively, unsupported, comparison, especially, measurable, mitigation, prevention, partially, involves, tradeoff, control, guarded, measure, refuses, depths, widely, judge, just; 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 Industrial AI Guardrails: Solving Field Service Reliability with Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors manufacturing, accelerating, architected, diagnostics, industrial, operations, equipment, examining, incidents, workforce, downtime, strained, verified, sectors, solving, damage, energy, sensor; 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 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 measurement, [6] supplies abstract-backed evidence on retrieval, rag, language model through the study CyberRerankBench: Benchmarking Rerankers for Cybersecurity Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors cyberrerankbench, explainability, cybersecurity, benchmarking, terminology, everywhere, resembling, securebert, cysecbert, direction, identical, intrusion, rerankers, revealing, encoders, headline, mismatch, parallel; 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 RAG (RETRIEVAL-AUGMENTED GENERATION) ЯК НОВА ПАРАДИГМА КОРПОРАТИВНОЇ АВТОМАТИЗАЦІЇ. Its evidence ledger records the specific descriptors inconsistencies, orchestration, adaptability, continuously, specificity, decoupling, extensions, interfaces, parameters, positioned, classical, considers, corporate, positions, primarily, business, customer, emerging; 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 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, [9] 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 baseline, [10] 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 transferable, automate, accu, quantization, extraction, employing, hydrides, hydrogen, obtained, extract, gemma2, llama3, metal, newly, give, prompting, datasets, protocol; 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 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, noisy-input 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 0.522; the intervention mean was 0.575. The paired change was +0.053, with a 95% interval from +0.051 to +0.056. In the prespecified adverse slice, the change was +0.044.

The machine-readable artifact `experiments/01-R05-114.json` contains all 64 paired records for retrieval-augmented answering. Consequently, the +0.053 result can be recomputed directly under the noisy-input 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 noisy-input, 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 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 low-signal stratum, 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 noisy-input 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). 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. Bellapukonda, J. (2026). Evaluating hallucination Mitigation in Retrieval-Augmented Generation Systems through Guarded Response Strategies, Embedding Comparison, and Retrieval Depth Analysis. https://doi.org/10.2139/ssrn.6702638

3. Fu, E. L. (2025). Industrial AI Guardrails: Solving Field Service Reliability with Retrieval-Augmented Generation. Global Business & Economics Journal. https://doi.org/10.70924/f83n6wqz/g91xh08t

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

6. Khan, N., Islam, M. S., & Saha, S. (2026). CyberRerankBench: Benchmarking Rerankers for Cybersecurity Retrieval-Augmented Generation. https://doi.org/10.2139/ssrn.7240159

7. НИЧ, & ПРАВОРСЬКА (2026). RAG (RETRIEVAL-AUGMENTED GENERATION) ЯК НОВА ПАРАДИГМА КОРПОРАТИВНОЇ АВТОМАТИЗАЦІЇ. Herald of Khmelnytskyi National University. Technical sciences, 361(1), 375-382. https://doi.org/10.31891/2307-5732-2026-361-53

8. 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

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

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