Auditing retrieval self-check for retrieval-augmented answering under 35% distractor density: a factor ablation

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 35% distractor density. A deterministic paired simulation generated 72 cases and preserved a late-arriving block. Mean evidence faithfulness changed from 0.545 to 0.599; the paired difference was +0.055 (95% interval +0.052 to +0.057). 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 late-arriving block. The operating setting is noisy-input; the stress level is fixed at 35% 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 50139. 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-R05-139.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 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 error slice, [3] 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, [4] 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 baseline, [5] supplies abstract-backed evidence on retrieval, rag, language model through the study Enhancing Dietary Supplement Question Answer via Retrieval-Augmented Generation (RAG) with LLM. Its evidence ledger records the specific descriptors encompasses, underscores, attributes, conclusion, facilitate, supplement, ultimately, decisions, interface, materials, minimized, subgraphs, entities, friendly, informed, inherent, dietary, enabled; 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 In RAG We Trust? Measuring Robustness of Retrieval-Augmented Generation Under Document Poisoning. Its evidence ledger records the specific descriptors abstention, corruption, falsehoods, behaviour, corrupted, factorial, falsehood, measuring, poisoning, quantized, checking, damaging, fraction, majority, negation, poisoned, quantify, reaction; 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 Adaptive Optimization for Retrieval-Augmented Generation via Diagnostic-Driven Strategy Generation. Its evidence ledger records the specific descriptors recommendation, actionability, intervention, actionable, adjustment, experience, llmpowered, annotated, comprises, diagnosis, essential, heuristic, granular, wikieval, dynarag, extends, grained, mapping; 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 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 reporting rule, [9] 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, [10] 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.

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 late-arriving block. 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, noisy-input 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 0.545; the intervention mean was 0.599. The paired change was +0.055, with a 95% interval from +0.052 to +0.057. In the prespecified adverse slice, the change was +0.047.

The machine-readable artifact `experiments/01-R05-139.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.055 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 late-arriving block. 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 late-arriving block, 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 noisy-input generator. A real-data study should retain distractor density and the late-arriving block 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. 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

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

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

5. Hou, Y., & Zhang, R. (2024). Enhancing Dietary Supplement Question Answer via Retrieval-Augmented Generation (RAG) with LLM. https://doi.org/10.1101/2024.09.11.24313513

6. Fasolino, I. (2026). In RAG We Trust? Measuring Robustness of Retrieval-Augmented Generation Under Document Poisoning. https://doi.org/10.32388/7hxkqe

7. Gai, Z., & Kuang, L. (2026). Adaptive Optimization for Retrieval-Augmented Generation via Diagnostic-Driven Strategy Generation. https://doi.org/10.36227/techrxiv.177155637.75354706/v1

8. 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

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

10. 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