Auditing retrieval self-check for retrieval-augmented answering under 13% distractor density: a hold-out check

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 13% distractor density. A deterministic paired simulation generated 72 cases and preserved a upper-severity quartile. Mean evidence faithfulness changed from 0.580 to 0.632; the paired difference was +0.053 (95% interval +0.051 to +0.055). 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 AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA), 1177-1180. 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 mixed-load; the stress level is fixed at 13% 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 50099. 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-R05-099.json`.

Reference [1] is used in Methods, not only as background: its work on AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA), 1177-1180 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 RAGulate: Retrieval-Augmented Generation for Post-hoc Literature-Grounded Regulatory Assessment. Its evidence ledger records the specific descriptors interpretations, transcriptional, classification, prioritization, normalization, transcription, availability, explanations, classifying, identifiers, predictions, accelerate, biologists, faithfully, hypothesis, motivation, regulation, streamline; 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 MULTI-STAGE VECTOR RETRIEVAL FOR RETRIEVAL-AUGMENTED GENERATION. Its evidence ledger records the specific descriptors prioritising, composition, additively, reciprocal, resistant, nfcorpus, sizefits, compose, default, matters, noisier, release, sampled, scifact, targets, uniform, adding, always; 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 ACW-RC: Adaptive Confidence-Weighted Retrieval Correction for Robust Augmented Generation. Its evidence ledger records the specific descriptors contradictions, miscalibration, biographical, distribution, eliminating, probability, substitutes, corrective, thresholds, assessing, biography, estimates, evaluator, pubhealth, unlabeled, blending, discrete, expected; 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 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 measurement, [6] 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 stress range, [7] supplies abstract-backed evidence on retrieval, rag, language model through the study Large Language Model with Federated Retrieval-Augmented Generation for Improved Knowledge Retrieval. Its evidence ledger records the specific descriptors outperformed, contributed, distributed, facilitated, detailed, leverage, revealed, nodes, lots, mmlu, upto, substantial,

federated, mistral, nature, introducing, achieving, potential; 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 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 reporting rule, [9] 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 baseline, [10] supplies abstract-backed evidence on retrieval, rag, language model through the study Research on Citizen Privacy Risk Assessment Method Based on Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors automatically, effectiveness, infringement, assessments, maintaining, mitigating, monitoring, validating, contracts, indicator, judgments, citizens, gradient, personal, severity, changes, citizen, factors; 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 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, mixed-load 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 0.580; the intervention mean was 0.632. The paired change was +0.053, with a 95% interval from +0.051 to +0.055. In the prespecified adverse slice, the change was +0.044.

The machine-readable artifact `experiments/01-R05-099.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.053 result can be recomputed directly under the mixed-load 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 mixed-load, 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 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 upper-severity quartile, 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 mixed-load 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). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA), 1177-1180. https://doi.org/10.1109/icmlca66850.2025.11336788

2. Zandigohar, M., & Dai, Y. (2026). RAGulate: Retrieval-Augmented Generation for Post-hoc Literature-Grounded Regulatory Assessment. https://doi.org/10.64898/2026.01.20.700704

3. Alabi Bankole, S., & Kayode Saheed, Y. (2026). ADAPTIVE MULTI-STAGE VECTOR RETRIEVAL FOR RETRIEVAL-AUGMENTED GENERATION. NLP & Big Data, 79-98. https://doi.org/10.5121/csit.2026.1601307

4. Kumar, C., Deshmukh, S., Khatter, H., Kumar, B., & Pal, Y. (2026). ACW-RC: Adaptive Confidence-Weighted Retrieval Correction for Robust Augmented Generation. https://doi.org/10.2139/ssrn.6546797

5. 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

6. Aryani, A. (2024). A Brief Introduction to Retrieval Augmented Generation (RAG). https://doi.org/10.59350/8mwjx-hry76

7. Hou, X., & Wang, X. (2024). Large Language Model with Federated Retrieval-Augmented Generation for Improved Knowledge Retrieval. https://doi.org/10.36227/techrxiv.171837853.31531482/v1

8. Ibiloye, A. C. (2025). Long Text Language Ensemble Models: Extension of Retrieval-Augmented Generation (RAG). https://doi.org/10.31219/osf.io/b94ng_v1

9. 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

10. Qu, J., & Luo, F. (2026). Research on Citizen Privacy Risk Assessment Method Based on Retrieval-Augmented Generation. https://doi.org/10.21203/rs.3.rs-9015963/v1