Auditing retrieval self-check for retrieval-augmented answering under 9% distractor density: a blocked comparison

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 9% distractor density. A deterministic paired simulation generated 72 cases and preserved a median slice. Mean evidence faithfulness changed from 0.585 to 0.636; the paired difference was +0.050 (95% interval +0.048 to +0.053). 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 median slice. The operating setting is mixed-load; the stress level is fixed at 9% 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 50123. 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-R05-123.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 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 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 reporting rule, [4] supplies abstract-backed evidence on retrieval, rag, language model through the study Improving Retrieval-Augmented Generation: A Systematic Literature Review of Retrieval-Enhancement Techniques. Its evidence ledger records the specific descriptors representative, bibliographic, generalisable, categorised, communities, enhancement, comparable, fragmented, identifies, vulnerable, databases, dispersed, practices, reporting, staleness, analysed, included, advance; 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 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. 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 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 error slice, [8] supplies abstract-backed evidence on retrieval, rag, language model through the study Autonomous QA Agent: A Retrieval-Augmented Framework for Reliable Selenium Script Generation. Its evidence ledger records the specific descriptors translating, autonomous, executable, execution, ingesting, lifecycle, commerce, elements, existent, markdown, selenium, formats, script, syntax, html, hallucinate, structure, grounds; 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 Retrieval-Augmented Generation. Its evidence ledger records the specific descriptors comprehensively, personalization, distortion, generators, mainstream, emergence, filtering, freshness, although, outdated, follow, timely, basic, power, transparency, retrievers, highlight, discuss; 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 A2JRAG: Process-Aware Retrieval-Augmented Generation for Public Legal Information Systems. Its evidence ledger records the specific descriptors entityrelationship, representation, twodimensional, demonstration, substantively, procedurally, conditioned, proceedings, transitions, escalation, misaligned, relational, formalize, mandatory, traverses, typically, authored, carolina; 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 median slice. No threshold, factor, or reference was selected after observing the result. The blocked comparison 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 blocked comparison.

4. Results

The baseline mean was 0.585; the intervention mean was 0.636. The paired change was +0.050, with a 95% interval from +0.048 to +0.053. In the prespecified adverse slice, the change was +0.040.

The machine-readable artifact `experiments/01-R05-123.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.050 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 median slice. This supports a focused follow-up on distractor density using measured inputs and the same blocked comparison 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 median slice, and the blocked comparison 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 median slice 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. 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. 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

4. Karasan, A. (2026). Improving Retrieval-Augmented Generation: A Systematic Literature Review of Retrieval-Enhancement Techniques. https://doi.org/10.21203/rs.3.rs-10368581/v1

5. 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

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

7. 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

8. Kasim Vali, D. (2025). Autonomous QA Agent: A Retrieval-Augmented Framework for Reliable Selenium Script Generation. https://doi.org/10.31224/5891

9. Liu, C. (2025). Retrieval-Augmented Generation. https://doi.org/10.31224/5781

10. Martin, T., & Gilchrist, S. (2026). A2JRAG: Process-Aware Retrieval-Augmented Generation for Public Legal Information Systems. https://doi.org/10.2139/ssrn.7114139