Tuning retrieval self-check for retrieval-augmented answering under 12% distractor density: a error-stratified estimate

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 12% distractor density. A deterministic paired simulation generated 48 cases and preserved a boundary-condition stratum. Mean evidence faithfulness changed from 0.479 to 0.530; the paired difference was +0.051 (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

Reproducible stress tests can expose weaknesses in retrieval-augmented answering before expensive field collection. 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 boundary-condition stratum. The operating setting is mixed-load; the stress level is fixed at 12% before analysis.

2. Methods

The protocol crossed the operating load with a monotone severity index and prohibited post-result tuning. We generated 48 cases with seed 50148. 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-R05-148.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 Dynamic Multi-Hop Retrieval-Augmented Generation Framework for Professional Domain Question Answering. Its evidence ledger records the specific descriptors completeness, structuring, critically, inadequate, surpassing, belief, upon, coherence, financial, incurring, validate, towards, stakes, professional, consistency, evaluations, benchmarks, factuality; we map those archived descriptors to the distractor density control. For the error slice, [3] supplies abstract-backed evidence on retrieval, rag, reasoning through the study Operationalizing Evidence Admissibility in Retrieval-Augmented Generation: The DCA-TRAG Architecture. Its evidence ledger records the specific descriptors admissibility, inadmissible, repositories, supersession, eligibility, enforcement, establishes, preliminary, simulations, statistical, downstream, expiration, protection, repository, guarantee, influence, revisions, remained; 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 GIANT: Leveraging Retrieval-Augmented Generation for Automated Bug Reproduction Test Generation. Its evidence ledger records the specific descriptors construction, reproduction, comparisons, conventions, procedures, reproduced, reproduces, submission, underlying, assurance, defects4j, exhausted, examined, isolated, produced, projects, reported, defects; 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 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 measurement, [6] 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 stress range, [7] 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, [8] 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, [9] 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, [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 boundary-condition stratum. 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, mixed-load 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 0.479; the intervention mean was 0.530. The paired change was +0.051, 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-R05-148.json` contains all 48 paired records for retrieval-augmented answering. Consequently, the +0.051 result can be recomputed directly under the mixed-load setting rather than inferred from prose.

5. Discussion

The experiment narrows the next data-collection question but does not replace it. For retrieval-augmented answering under mixed-load, the sign of the evidence faithfulness change remained positive in both the full set and the boundary-condition stratum. 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 boundary-condition stratum, 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 mixed-load generator. A real-data study should retain distractor density and the boundary-condition 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). 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. Zhu, J., & Tong, Y. (2026). Dynamic Multi-Hop Retrieval-Augmented Generation Framework for Professional Domain Question Answering. https://doi.org/10.22541/au.177499050.00368942/v1

3. Garg, A., & Esposito, J. (2026). Operationalizing Evidence Admissibility in Retrieval-Augmented Generation: The DCA-TRAG Architecture. https://doi.org/10.2139/ssrn.7111618

4. Bo, L., Duan, R., Shi, S., Sun, X., Lin, Z., & Ji, W. (2026). GIANT: Leveraging Retrieval-Augmented Generation for Automated Bug Reproduction Test Generation. https://doi.org/10.2139/ssrn.7269657

5. 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

6. 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

7. 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

8. 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

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

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