Auditing retrieval self-check for retrieval-augmented answering under 21% distractor density: a robust mean contrast

ABSTRACT
We evaluated retrieval self-check for retrieval-augmented answering under 21% distractor density. A deterministic paired simulation generated 72 cases and preserved a low-signal stratum. Mean evidence faithfulness changed from 0.576 to 0.617; the paired difference was +0.041 (95% interval +0.039 to +0.043). 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 low-signal stratum. The operating setting is mixed-load; the stress level is fixed at 21% 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 50051. 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-R05-051.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 Building a Security and Reliability Evaluation Suite for Retrieval-Augmented Generation (RAG) Systems. Its evidence ledger records the specific descriptors illustrative, instruments, sacrificing, avoidance, tradeoffs, fairness, monitors, template, toxicity, compute, utility, siloed, hence, suite, miss, conditioning, standardized, calibration; 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 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 reporting rule, [4] 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 baseline, [5] supplies abstract-backed evidence on retrieval, rag, language model through the study Graph-Native Intelligence: A Novel Architecture for Multi-Source Knowledge Synthesis Beyond Retrieval Augmented Generation. Its evidence ledger records the specific descriptors fundamentally, underexplored, perspectives, innovations, narratives, translator, traversing, connected, diversity, increases, ingestion, political, traversal, category, converts, narrator, becomes, vectors; 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 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 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 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, [9] supplies abstract-backed evidence on retrieval, rag, language model through the study Merging Mixture of Experts and Retrieval Augmented Generation for Enhanced Information Retrieval and Reasoning. Its evidence ledger records the specific descriptors applicability, acknowledges, investigates, contribute, overcoming, adaptable, extensive, intrinsic, showcases, analyses, already, avenues, experts, marking, merging, mixture, nuanced, ongoing; 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 Pool-Gated Retrieval: Beyond Retrieval-Augmented Generation Toward Accountable Evidential Admission. Its evidence ledger records the specific descriptors accountability, epistemically, justificatory, accountable, recognition, commitment, preventing, principled, propagated, registered, structural, widespread, admission, committed, declaring, enforcing, epistemic, exclusive; 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 robust mean contrast 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 robust mean contrast.

4. Results

The baseline mean was 0.576; the intervention mean was 0.617. The paired change was +0.041, with a 95% interval from +0.039 to +0.043. In the prespecified adverse slice, the change was +0.037.

The machine-readable artifact `experiments/01-R05-051.json` contains all 72 paired records for retrieval-augmented answering. Consequently, the +0.041 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 low-signal stratum. This supports a focused follow-up on distractor density using measured inputs and the same robust mean contrast 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 robust mean contrast 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 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). 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. Roy, P., & Roy, D. (2025). Building a Security and Reliability Evaluation Suite for Retrieval-Augmented Generation (RAG) Systems. https://doi.org/10.20944/preprints202510.0418.v2
3. 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
4. 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
5. Alvaro, A. (2026). Graph-Native Intelligence: A Novel Architecture for Multi-Source Knowledge Synthesis Beyond Retrieval Augmented Generation. https://doi.org/10.2139/ssrn.6335438
6. 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
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. 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
9. Xiong, X., & Zheng, M. (2024). Merging Mixture of Experts and Retrieval Augmented Generation for Enhanced Information Retrieval and Reasoning. https://doi.org/10.21203/rs.3.rs-3978298/v1
10. Kim, M. H. (2026). Pool-Gated Retrieval: Beyond Retrieval-Augmented Generation Toward Accountable Evidential Admission. https://doi.org/10.20944/preprints202606.0414.v1