Hallucination Detection in Retrieval-Augmented Generation Systems via Self-Consistency Scoring—A Reference-Free, Retrieval-Aware Detection Framework

Tang Yuxin
Qingdao University of Technology, China

ABSTRACT

Large language models (LLMs) augmented with retrieval mechanisms have become a dominant paradigm for grounding generative output in external evidence, yet Retrieval-Augmented Generation (RAG) systems continue to produce hallucinated content even when relevant context is retrieved successfully. Existing hallucination detection methods either require access to token-level output probabilities unavailable in many deployed black-box systems, evaluate a single deterministic generation against retrieved passages, or were designed for open-ended generation without an explicit retrieval anchor. This paper proposes RAG-SC, a reference-free, claim-level hallucination detection framework tailored to the RAG setting. RAG-SC combines two complementary signals computed purely from black-box sampling access to the generator: a context-faithfulness score that checks whether each atomic claim is entailed by the retrieved passages, and a cross-sample consistency score that measures the semantic agreement of a claim across multiple stochastically resampled generations conditioned on the same retrieved context. We position RAG-SC against three established paradigms—sampling-based detection, atomic factuality scoring, and RAG-specific evaluation—and argue that none jointly exploits retrieval grounding and generation-time stochasticity. We present the method's design, a concrete evaluation protocol across three open-domain question-answering benchmarks, and a discussion of expected behavior, limitations, and threats to validity, providing a template for empirical validation in future work.

Keywords: Retrieval-Augmented Generation; Hallucination Detection; Self-Consistency; Large Language Models; Factuality Evaluation.

1. INTRODUCTION

Large language models produce fluent text that is not always grounded in fact, a phenomenon broadly termed hallucination (Ji et al., 2023) [4]. Retrieval-Augmented Generation (RAG) addresses this by conditioning generation on passages retrieved from an external corpus, typically via a dense retriever such as DPR (Karpukhin et al., 2020) [7], and was formalized as a general fine-tuning recipe by Lewis et al. (2020) [10]. By supplying the generator with relevant, up-to-date evidence, RAG reduces a model's reliance on parametric memory alone.

Critically, retrieval grounding reduces but does not eliminate hallucination. Shuster et al. (2021) [13] showed that neural-retrieval-in-the-loop dialogue models still produce unsupported statements, particularly when the retriever returns noisy or partially relevant passages, or when the generator disregards correct retrieved evidence in favor of its parametric prior. This leaves an open problem: how can a practitioner detect, at inference time and without ground-truth labels, which claims in a RAG system's output are unsupported by what was actually retrieved?

Several detection paradigms exist but none is fully suited to this setting. SelfCheckGPT (Manakul et al., 2023) [11] detects hallucination in a zero-resource, black-box manner by sampling multiple generations and measuring their mutual consistency, but it was designed for open-ended generation with no retrieval anchor, so it cannot distinguish a claim that is unsupported by the retrieved context from one that is simply phrased differently across samples. FActScore (Min et al., 2023) [12] decomposes long-form text into atomic facts and checks each against a knowledge source, but it assumes access to a comprehensive external knowledge base and evaluates a single generation rather than exploiting sampling variance. RAGAS (Es et al., 2024) [1] introduces a reference-free faithfulness metric specific to RAG pipelines, but computes it from one deterministic generation using LLM-prompted claim verification, leaving the additional signal available from repeated stochastic sampling unused.

This paper proposes RAG-SC (Retrieval-Augmented Self-Consistency Scoring), a framework that unifies retrieval-grounded entailment checking with sampling-based consistency. The design is motivated by two independent lines of evidence: semantic entropy research showing that clustering semantically equivalent generations yields a strong, label-free uncertainty signal correlated with factual error (Kuhn et al., 2023 [8]; Farquhar et al., 2024 [2]), and self-consistency decoding showing that repeated sampling under fixed conditioning reveals a model's effective answer distribution (Wang et al., 2023) [14]. RAG-SC adapts both ideas to a setting where the conditioning context is itself retrieved evidence, allowing the detector to separate retrieval-grounded claims from claims the generator introduces independently of that evidence.

The remainder of this paper reviews related work (Section 2), formalizes the detection problem (Section 3), describes the RAG-SC method (Section 4), specifies an evaluation protocol for future empirical validation (Section 5), and discusses expected behavior and limitations (Section 6).

2. RELATED WORK

2.1 Retrieval-Augmented Generation

RAG combines a parametric generator with a non-parametric retrieval index, using a query encoder to retrieve top-k passages that condition generation (Lewis et al., 2020) [10]. The retriever is commonly a dense bi-encoder trained with contrastive objectives, as in Dense Passage Retrieval (Karpukhin et al., 2020) [7]. Gao et al. (2024) [3] survey the rapid expansion of RAG architectures, noting a persistent gap between retrieval quality and generation faithfulness: even a perfect retriever does not guarantee that the generator will use the retrieved evidence correctly.

2.2 Hallucination: Taxonomy and Causes

Ji et al. (2023)[4] distinguish intrinsic hallucination, where output directly contradicts the source, from extrinsic hallucination, where output cannot be verified against the source at all. In the RAG setting, both types are possible: a generator may contradict a retrieved passage, or it may introduce a claim that the retrieved passages neither support nor contradict, typically drawn from parametric memory. Shuster et al. (2021) [13] provide direct empirical evidence that retrieval augmentation lowers but does not close this gap in dialogue generation.

2.3 Reference-Free Detection Methods

Three detection paradigms are most relevant. First, sampling-based detection: SelfCheckGPT (Manakul et al., 2023) [11] repeatedly samples a black-box model and checks whether a sentence in a reference generation is corroborated by the sampled set, using BERTScore, question-answering agreement, or natural language inference as the comparison function. Second, uncertainty-based detection: Kuhn et al. (2023) [8] propose semantic entropy, clustering sampled generations by bidirectional entailment and computing entropy over clusters rather than over raw token sequences, later validated at scale by Farquhar et al. (2024) [2] in Nature. Third, atomic factuality scoring: FActScore (Min et al., 2023) [12] decomposes generations into atomic facts and verifies each against a trusted knowledge source, reporting that even strong commercial systems achieve well under perfect factual precision on biography generation.

2.4 RAG-Specific Evaluation

RAGAS (Es et al., 2024) [1] is, to our knowledge, the most widely adopted reference-free evaluation suite designed specifically for RAG pipelines, offering faithfulness, answer relevancy, and context precision/recall metrics computed via LLM prompting rather than human annotation. Its faithfulness metric is conceptually closest to our goal, but it is computed from a single generation and does not incorporate the sampling-consistency signal that SelfCheckGPT [11] and semantic entropy methods [8] rely on. Table 1 summarizes how these paradigms compare along the dimensions most relevant to a black-box, retrieval-grounded deployment, and motivates the design of RAG-SC.

Table 1. Comparison of reference-free hallucination/faithfulness detection paradigms.

Method	Needs token probabilities	Needs reference answer	Retrieval-aware	Uses sampling variance
SelfCheckGPT (Manakul et al., 2023)	No	No	No	Yes
FActScore (Min et al., 2023)	No	No (uses KB)	Partial	No
RAGAS Faithfulness (Es et al., 2024)	No	No	Yes	No
Semantic Entropy (Kuhn et al., 2023)	Optional	No	No	Yes
RAG-SC (proposed)	No	No	Yes	Yes

3. PROBLEM FORMULATION

Consider a RAG pipeline in which a query q is passed to a retriever R, returning a context $C = \{d_1, \ldots, d_k\}$ of k passages, and a generator G samples an answer $y \sim p(y \mid q, C)$. Following Ji et al. (2023), we define a claim c within y as hallucinated, relative to C, if c is neither entailed by nor inferable from any passage in C. Our objective is to assign each claim c in a reference generation y_0 a hallucination-risk score $s(c) \in [0, 1]$, using only black-box sampling access to G — i.e., without access to token-level probabilities, model weights, or ground-truth labels — such that higher s(c) corresponds to higher probability that c is hallucinated.

4. PROPOSED METHOD: RAG-SC

RAG-SC produces a claim-level score by combining a context-faithfulness signal with a cross-sample consistency signal, both computed without reference labels.

4.1 Claim Decomposition

Given the reference generation y_0 (the model's greedy or top-1 output), we decompose it into a set of atomic claims $A = \{c_1, \ldots, c_m\}$ using the prompting-based decomposition strategy introduced by Min et al. (2023), in which an auxiliary LLM splits the passage into short, self-contained factual statements. Atomic decomposition is preferred over sentence-level checking because a single sentence frequently mixes a supported claim with an unsupported one, which sentence-level scoring would conflate.

4.2 Context-Faithfulness Score

For each claim c_i, we compute an entailment score $f(c_i)$ using a natural language inference (NLI) model applied between c_i and each passage $d_j \in C$, taking the maximum entailment probability across passages: $f(c_i) = \max_j \text{NLI}(d_j \rightarrow c_i)$. This directly measures whether the retrieved evidence, as opposed to the generator's other samples, supports the claim, addressing the gap identified in Section 2.3 where sampling-only methods cannot distinguish retrieval-grounded agreement from coincidental agreement across hallucinated samples.

4.3 Cross-Sample Consistency Score

Holding q and C fixed, we draw N additional stochastic samples $y_1, \ldots, y_n$ from $G(\cdot \mid q, C)$ at temperature $T > 0$, following the resampling procedure used by SelfCheckGPT (Manakul et al., 2023) and by semantic entropy methods (Kuhn et al., 2023). Each sample is decomposed into atomic claims, and for c_i we compute a consistency score $u(c_i)$ as the proportion of samples containing a semantically equivalent claim, where semantic equivalence is determined by bidirectional NLI entailment, mirroring the clustering criterion of Kuhn et al. (2023). Low $u(c_i)$ indicates that the generator does not reliably reproduce c_i when resampled, a signature associated with confabulation rather than evidence-grounded recall.

4.4 Combined Score and Thresholding

The final hallucination-risk score is a convex combination $s(c_i) = \alpha \cdot (1 - f(c_i)) + (1 - \alpha) \cdot (1 - u(c_i))$, with $\alpha \in [0, 1]$ a tunable weight. Claims with $s(c_i)$ above a calibrated threshold are flagged for human review or automatic suppression. Two failure modes become separable under this formulation: claims with low f but high u indicate the generator is

consistently asserting something the retrieved context does not support (a strong hallucination signal independent of retrieval quality); claims with low f and low u indicate the generator is simply uncertain, which may instead reflect retrieval failure rather than generator confabulation. This decomposition is the principal methodological contribution relative to SelfCheckGPT, which has no $f(c_i)$ term, and RAGAS, which has no $u(c_i)$ term.

5. EVALUATION PROTOCOL

Because validating RAG-SC requires running multiple LLMs and retrieval pipelines at scale, we specify here a concrete, falsifiable evaluation protocol intended for empirical follow-up rather than reporting results obtained from an unexecuted study. Table 2 summarizes the protocol.

Table 2. Proposed evaluation protocol for RAG-SC.

Component	Choice	Rationale
Datasets	Natural Questions (Kwiatkowski et al., 2019) [9]; TriviaQA (Joshi et al., 2017) [5]; HotpotQA (Yang et al., 2018) [15]	Standard open-domain QA benchmarks with gold passages, enabling controlled retrieval and known answers for claim verification
Baselines	SelfCheckGPT, FActScore, RAGAS Faithfulness, semantic entropy	Covers sampling-based, atomic-fact-based, RAG-specific, and uncertainty-based paradigms
Metrics	AUC-PR for claim-level hallucination detection; Spearman correlation with human factuality ratings	Matches the evaluation protocol used by Manakul et al. (2023) for comparability
Sampling	N = 10 stochastic generations per query at temperature 0.8, fixed retrieved context	Balances variance estimation quality against inference cost

Ablations should isolate the contribution of $f(c_i)$ and $u(c_i)$ independently, vary $N \in \{3, 5, 10, 20\}$ to characterize the cost-accuracy trade-off, and test sensitivity to retriever quality by deliberately degrading top-k retrieval precision, following the diagnostic approach used in RAG evaluation surveys (Gao et al., 2024) [3].

6. DISCUSSION

We expect RAG-SC to outperform purely sampling-based detectors such as SelfCheckGPT specifically on claims where hallucinated content is reproduced consistently across samples-a case sampling-consistency alone misclassifies as faithful, but which the context-faithfulness term should catch. Conversely, we expect RAG-SC to outperform RAGAS's single-generation faithfulness metric on claims affected by sampling variance, since RAGAS cannot detect cases where the greedy output happens to be supported while the model's broader answer distribution is not.

Several limitations and threats to validity should be acknowledged. First, computational cost scales linearly with N, which may be prohibitive for latency-sensitive applications; this motivates the planned ablation over N. Second, the method inherits the error characteristics of its NLI component, and errors in entailment judgment propagate directly into both f and u. Third, RAG-SC's context-faithfulness term is only as reliable as retrieval quality itself: if the retriever returns irrelevant passages, low $f(c_i)$ may reflect retrieval failure rather than a hallucinated generator, and the method as currently formulated does not separately attribute blame between retriever and generator. Disentangling these two failure sources, perhaps via a counterfactual re-retrieval step, is a natural direction for future work.

7. CONCLUSION AND FUTURE WORK

This paper proposed RAG-SC, a reference-free, claim-level hallucination detection framework for retrieval-augmented generation that combines context-faithfulness entailment scoring with cross-sample consistency scoring. By situating the method against SelfCheckGPT, FActScore, and RAGAS, we argued that existing approaches each omit one of two complementary signals available in a black-box RAG deployment: grounding against retrieved evidence, and variance across repeated sampling. We provided a concrete evaluation protocol spanning three standard open-domain QA benchmarks to enable empirical validation. Future work includes implementing and running this protocol, extending the

framework to multi-hop retrieval settings, and exploring retriever-generator blame attribution to better localize the source of detected hallucinations.

REFERENCES

[1] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150-158). Association for Computational Linguistics.

[2] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625-630.

[3] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

[4] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248.

[5] Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1601-1611). Association for Computational Linguistics.

[6] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

[7] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). Association for Computational Linguistics.

[8] Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023).

[9] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466.

[10] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (pp. 9459-9474).

[11] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9004-9017). Association for Computational Linguistics.

[12] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 12076-12100). Association for Computational Linguistics.

[13] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3784-3803). Association for Computational Linguistics.

[14] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023).

[15] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2369-2380). Association for Computational Linguistics.