

A Review of Hallucination Detection Techniques in Retrieval-Augmented Generation (RAG)

¹Divyanshu Garg, ²Priyansh Sharma

¹Student, Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, India

²Assistant Professor, Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, India

Abstract - Retrieval-augmented generation (RAG) anchors the generation of large language models (LLMs) to external documents, but models can still generate statements that are in conflict with or extend beyond the retrieved documents. To trustfully use a RAG system, it is essential to be able to detect such hallucinations. This paper reviews ten research papers that together span consistency-based, claim-level, retrieval-grounded, entailment-based, LLM-as-a-judge, black-box, white-box, mechanistic-interpretability and benchmark-oriented approaches, including work on long-context and noisy-label evaluation. The papers are classified in a taxonomy of nine categories and analyzed for their accuracy, faithfulness, grounding, explainability, computational cost, model accessibility, generalization, long-context capability and robustness to label noise. The review is qualitative, in that it does not apply or test a new method, and only summarizes the results as reported by the original authors. From the literature, three trends are apparent: output-only detectors have been replaced by detectors that are grounded and/or model-internal; RAG-specific benchmarks highlight challenges that general factuality benchmarks do not; and there is no single detector that is consistently effective across models, domains and context lengths. The paper ends with the research gaps, deployment challenges and future directions recommended by the literature reviewed.

Keywords - hallucination detection, retrieval-augmented generation, large language models, faithfulness, factuality, benchmarks, mechanistic interpretability, literature review.

I. INTRODUCTION

While large language models (LLMs) can now write, summarize and answer questions fluently, they can also fabricate, be unfaithful, or be inconsistent with the content. This behaviour is generally referred to as hallucination [5]. It is commonly employed to mitigate it by providing the model with retrieved documents, however, even with the inclusion of retrieved documents, LLMs can still make claims that are not supported by or contradict the retrieved documents [3]. However, research into the internal mechanism of RAG reveals that this is possible even if the retrieved content is correct and pertinent [4].

This is why it is essential to detect hallucinations, or determine whether and why a generated response deviates from its evidence, as a prerequisite for trustworthy RAG systems.

This paper is not an original study, but a review. It is not a model and it does not report any experiments that it has conducted. It examines ten research papers that together cover the main families of detection methods: SelfCheckGPT [1], FActScore [2], RAGTruth [3], ReDeEP [4], the benchmark study of Chen et al. [5], semantic entropy [6], Lookback Lens [7], RAGAs [8], AlignScore [9] and the LLM-as-a-judge study of Zheng et al. [10]. The set was designed to ensure that there is at least one paper for each approach category in Section IV, and that preference was given to papers which dealt with RAG directly. Foundational papers that are not RAG-specific are included where they supply a technique that RAG detectors build on; their relevance is explained in Section V.

Scope and evidence base. All bibliographic details were checked against proceedings pages, publisher records or the ACL Anthology. The full texts of RAGTruth [3] and ReDeEP [4] were analyzed, together with the definitional sections of FActScore [2]. In the other papers, the review is based on the published abstracts and bibliographic records. The quantitative results are therefore only reported when the original authors have reported them and when they were available to this review. If a detail is not mentioned here, the text will state that and direct the reader to the original paper. The reviewer's interpretation of the statements is indicated.

II. BACKGROUND

A. LARGE LANGUAGE MODELS

LLMs are large neural language models that are trained on a large body of text to predict the next token. They have performed well in text generation, translation and question answering, while one of the major challenges for their deployment in the real world is their tendency to hallucinate [3]. All the knowledge that an LLM has acquired without the help of humans is embedded in its parameters. This parametric knowledge is not always

complete or up-to-date, which is one reason for retrieval augmentation [3, 4].

B. Retrieval-Augmented Generation

A RAG system is a combination of retrieval module and generation module using LLM. The retrieval module searches a reference textual database and the LLM serves as a natural-language interface between the user and the database [8]. In a very simplified way, the pipeline consists of four steps: 1) The query is sent, 2) passages are retrieved, 3) the passages and query are combined into a prompt, and 4) the LLM generates a response. The quality of retrieval and faithfulness of generation are different issues, as in RAGAs, the relevance of retrieved passages, faithful use of retrieved passages, and the overall quality of generation are evaluated separately [8].

C. Hallucination in LLMs

There is a variation in definitions between papers, and this is another finding of this review. In [5] Chen et al. take a wide approach to include unfaithful, fabricated or inconsistent content. Farquhar et al. focus on a smaller subset, namely, arbitrary and incorrect generations, called confabulations [6]. SelfCheckGPT sees the problem as statements that are not factual, which can lead to mistrust of fluent outputs [1]. FActScore observes that long generations mix supported and unsupported pieces of information, so a single true/false label is inadequate [2].

D. Hallucination in RAG

In the RAG setting, the natural reference for judging a response is the retrieved context. RAGTruth describes RAG hallucination as the presentation of claims that are unsupported by, or contradictory to, the retrieved contents [3]. Sun et al. use the term RAG hallucination for exactly this situation and distinguish it from knowledge conflict: a conflict between context and parametric knowledge can lead to hallucination but does not always do so [4]. Their study attributes RAG hallucination to two internal factors, feed-forward layers that add too much parametric knowledge and attention heads that fail to retain the retrieved information [4].

E. Types of Hallucinations

RAGTruth defines four RAG-specific types: evident conflict, subtle conflict, evident introduction of baseless information, and subtle introduction of baseless information [3]. The first two contradict the context; the last two add content that the context does not support. In that corpus, baseless information was substantially more common than conflict, especially in question answering [3]. Other papers employ various units: sentences [1]; atomic facts [2]; confabulations [6] and “contextual

hallucinations” with respect to an input passage [7]. These can be arranged on two dimensions in our synthesis: against what is the response judged (world knowledge or provided context) and on what level of granularity is the judgment made (passage, sentence, atomic fact or word span).

III. HALLUCINATION DETECTION: PROBLEM FORMULATION

A. What Is Being Detected

In our formulation, the query q is presented to a detector, along with an optional retrieved context C , and a generated response r . It provides a judgement or a score for the entire response or labels for sections of the response. There are different granularities: RAGTruth scores at the response level and span level [3], SelfCheckGPT scores at the sentence level [1] and FActScore scores at the atomic fact level [2]. Detectors can also be different in the type of input they can take: sampled responses from the same model [1], [6], a knowledge source [2], the retrieved context [3], [8], [9] or the internal states of the model [4], [7].

B. Factuality versus Faithfulness

Factuality is related to the agreement with the real world, faithfulness is related to the agreement with a provided source. FActScore falls between the two by establishing factual precision with respect to a selected knowledge source, instead of a single global truth [2]. This is important for RAG as a response may be correct but still unfaithful. RAGTruth makes this clear: information that is not mentioned in the reference, but which might be true is also marked as hallucination in the strict RAG setting, with a separate implicit-truth annotation, allowing the user to choose other policies [3]. The reason cited is that LLM's tend to make mistakes when it comes to the details of facts, and RAG applications generally tell the models not to say things that aren't in the references [3].

C. Incorrect versus Unsupported

If the statement is incorrect, it is not supported by the retrieved evidence; if the statement is unsupported, it is not supported by the retrieved evidence. A response to a question about restaurant amenities that claims credit cards are accepted, when the retrieved data is silent, is unsupported and could turn out to be correct. RAGTruth's labeling of such content as hallucination follows from this reasoning [3]. Consequently, a detector tuned for factual correctness (for example one based on consistency across samples [1], [6]) and a detector tuned for context support [3], [9] answer different questions, and reported scores are not interchangeable.

D. Why RAG Changes the Task

Five differences follow from the reviewed papers. First, an explicit reference exists, so detection can be framed as a grounding check, whereas SelfCheckGPT was designed for the case in which no external database is used [1]. Second, the LLM balances external context against parametric knowledge, and detectors that ignore this balance can be confounded [4]. Third, RAG inputs can be long, and long-context grounded benchmarks were found to be scarce [5]. Fourth, hallucinations are relatively rare in RAG, which means that a significant amount of resources is needed for annotation [3]. Fifth, labels are sometimes controversial; RAGTruth offers additional annotations for cases where there is ambiguity, like null values in structured data [3]; and Chen et al. point out that in real deployments, there is label noise [5].

IV. TAXONOMY OF HALLUCINATION DETECTION METHODS

The literature is divided into nine categories as summarized in Tab. I. There are overlaps between the categories: a method can be black-box and consistency-based, or benchmark-oriented and long-context. The papers are then classified in each case into a main category and a secondary category (where applicable) listed in Section VI.

TABLE I MAPPING OF THE REVIEWED PAPERS TO THE TAXONOMY

Ref.	Primary category	Access
[1]	Consistency-based (4.1)	Black-box
[2]	Fact/claim-level (4.2)	Black-box
[3]	Benchmark (4.6)	Model-agnostic
[4]	Mechanistic interpretability (4.8)	White-box
[5]	Robust evaluation (4.9)	Model-agnostic
[6]	Consistency-based (4.1)	Sampling
[7]	White-box / model-internal (4.7)	White-box
[8]	LLM-as-a-judge (4.5)	Black-box
[9]	NLI/entailment (4.4)	Trained checker
[10]	LLM-as-a-judge (4.5)	Black-box

A. Consistency-Based Methods

The basic idea and workflow is to sample a few responses to the same input, and consider disagreement as a sign of hallucination. SelfCheckGPT compares the

sentence with the sampled responses [1]. Semantic entropy measures uncertainty in meaning, as opposed to words, because the same meaning can be expressed in numerous ways [6]. Advantages: There is no external database or task-specific training data required [1, 6]. Limitations: multiple generations needed (our inference) and consistency is not correctness or grounding; semantic entropy is against confabulations, a subset of hallucinations [6].

B. Fact/Claim-Level Methods

The basic concept and process: divide a response into small verifiable parts and verify each part with evidence. FactScore is based on atomic facts and reports the percentage that is supported by a knowledge source [2]. RAGAs uses the following decomposition of faithfulness, which is related to the above. Strengths: partially supported claims can be measured, and unsupported claims can be found [2]. Limitations: Results are highly dependent on the knowledge source, and human evaluation is expensive, so this led to the development of an automated estimator for FactScore [2].

C. Retrieval-Grounded Methods

Basic idea and workflow: compare the response with retrieved evidence, either own RAG system's context or evidence retrieved for checking. FactScore extracts passages to check atomic facts [2]; RAGAs assesses faithful use of retrieved passages [8]; RAGTruth marks passages in the response with the context provided to the model [3]. Strengths: judgements are based on clear evidence. Weaknesses: The quality of the evidence reduces the quality of detection; the quality of detection is based on the support, not on the truth.

D. NLI/Entailment-Based Methods

Basic idea and workflow: consider the retrieved context as a premise and the response as a hypothesis and make an estimation whether the premise supports the response. AlignScore generalizes this by training one alignment function on data from seven tasks, including natural language inference (NLI), question answering and fact verification [9]. Lookback Lens compares its attention-based detector with a text-based entailment model [7]. Strengths: a small trained checker can be used [9]. Weaknesses: the checker must be trained and maintained separately, and the unit of comparison (chunk, sentence) must be chosen for RAG use (our inference).

E. LLM-as-a-Judge Methods

Basic idea and workflow: prompt a strong LLM to assess a response, optionally against the context. Zheng et al. study this paradigm for chat assistants and document

position, verbosity and self-enhancement biases and limited reasoning ability [10]. RAGTruth's prompt baselines and RAGAs-style faithfulness scoring are applications to RAG [3], [4], [8]. Strengths: flexible and requires little labeled data. Weaknesses: cost, bias, and dependence on the judge; Chen et al. nevertheless report that a basic LLM-as-a-judge baseline performs competitively on RAG-based benchmarks [5].

F. Dataset/Benchmark-Based Evaluation

Basic idea and workflow: build annotated corpora so detectors can be compared. RAGTruth provides word-level annotations of natural RAG responses [3]; Chen et al. set out desiderata for such benchmarks and contribute TRIVIA+ [5]. Strengths: reproducible comparison and analysis of where hallucinations occur [3]. Weaknesses: annotation is costly and partly subjective, and a benchmark covers only its tasks, models and context lengths [3], [5].

G. White-Box / Model-Internal Methods

Basic idea and workflow: read signals from the model's internal computation, such as hidden states or attention. Lookback Lens uses the ratio of attention on the context versus newly generated tokens as features for a linear classifier [7]. Strengths: a compact detector that, according to its authors, transfers across tasks and models [7]. Weaknesses: requires access to model internals, which excludes closed APIs (our inference from the method).

H. Mechanistic Interpretability Methods

Basic idea and workflow: identify specific internal components responsible for hallucination and build detectors from them. ReDeEP finds that Knowledge FFNs and Copying Heads are implicated in RAG hallucination and regresses two decoupled scores to detect it [4]. Strengths: the method offers an explanation as well as a score. Weaknesses: white-box access is needed and the analysis in [4] was conducted on the LLaMA family.

I. Robust/Noise-Aware Evaluation

Basic idea and workflow: test detectors under conditions closer to deployment, such as very long contexts and imperfect labels. Chen et al. supply noisy label sets for stress-testing and report that label noise hinders detection performance [5]. Strengths: exposes fragility hidden by clean benchmarks. Weaknesses: the noise schemes are constructed by design, and this line of work is at an early stage; among the ten papers, only [5] addresses it directly.

V. DETAILED REVIEW OF THE TEN PAPERS

The papers are presented in the order of the references. Each entry follows a common template. Results are those that are reported by the original authors, otherwise it is considered as interpretation.

A. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection [1]

Authors: P. Manakul, A. Liusie, M. J. F. Gales. Year: 2023. Venue: EMNLP 2023 (main conference). DOI: 10.18653/v1/2023.emnlp-main.557.

Research problem. Existing fact-checking approaches need access to output probabilities, which some systems do not expose, or depend on external databases reached through separate modules; the authors seek a method for black-box models that needs neither [1].

Methodology. A sampling-based approach: several responses are drawn for the same prompt and the response under test is checked for consistency with them. The intuition is that facts the model knows tend to be reproduced across samples, whereas fabricated details vary [1]. RAGTruth's replication samples three extra responses and uses GPT-3.5-turbo to verify sentence-level consistency [3].

Results. The authors report that the approach detects non-factual and factual sentences and ranks passages by factuality, with considerably higher AUC-PR and correlation than grey-box baselines [1]. Exact values are not reproduced here. On RAGTruth, its authors report an overall response-level F1 of 58.8% with GPT-3.5-turbo and describe the performance as unsatisfactory [3].

Limitations. Consistency across samples is not the same as agreement with a source (our interpretation), and multiple samples are required. It does not use the retrieved context.

Contribution and relevance to RAG. It established sampling-based black-box detection as a baseline used by later RAG work [3], [4]. It is not RAG-specific; its value for RAG is as a context-free baseline.

B. FACTScore: Fine-grained Atomic Evaluation of Factual Precision [2]

Authors: S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi. Year: 2023. Venue: EMNLP 2023. DOI: 10.18653/v1/2023.emnlp-main.741.

Research problem. Long-form generations combine supported and unsupported information, which is not possible to be judged by binary judgment and human evaluation is slow and expensive [2].

Methodology. The generation is divided into atomic facts, each of which is a short sentence that conveys one piece of information, and the score is the percentage that is backed by a reliable knowledge source. The authors consider factual precision as a function of the knowledge source provided and not a single truth in the world. An automated estimator is based on retrieval and a powerful language model [2]. Even single sentences can contain supported and unsupported facts, as the authors point out, in 40% of the cases they analyzed, ChatGPT did that [2].

Results. The human evaluation resulted in 58% for ChatGPT, whereas the automated estimator had an error rate of less than 2%, and ChatGPT and GPT-4 were more factual than the public models, with Vicuna and Alpaca being among the best public models [2]. According to the authors, it would have cost approximately \$26K to have evaluated the 6,500 generations by humans [2].

Limitations The human evaluation is expensive [2]. Paper evaluation will be on biographies and will be based on the knowledge source covered (our interpretation).

Contribution and relevance to RAG. The atomic-fact unit and source-relative definition of precision naturally extend to context-relative faithfulness in RAG. The atomic-fact unit and source-relative definition of precision extend naturally to context-relative faithfulness in RAG. The paper is not specific to RAG but it assesses a retrieval-augmented system and verifies claims using retrieval [2].

C. RAGTruth: A Hallucination Corpus for Trustworthy RAG Models [3]

Authors: C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, T. Zhang. Year: 2024. Venue: ACL 2024 (Volume 1: Long Papers). DOI: 10.18653/v1/2024.acl-long.585. URL: aclanthology.org/2024.acl-long.585.

Research problem. There is a need for benchmarks to evaluate hallucination in RAG, and the current benchmarks are either synthesized, small, or not RAG-specific [3].

Methodology. The authors created a manually annotated, case and word level corpus of naturally generated RAG responses, each of which was annotated with four types of hallucination. Each response was annotated by two annotators, and a third for large disagreements. At response level, 91.8% of the participants agreed, and at span level, 78.8% of the participants agreed [3]. They then benchmarked detection methods and optimised a small LLM as a detector [3].

Dataset. 2,965 instances across question answering (MS MARCO, three passages per question), data-to-text writing (Yelp Open Dataset) and news summarization (CNN/Daily Mail and recent news), yielding 17,790 responses [3].

Results. Response-level overall F1 was 63.4% for prompt-based GPT-4-turbo, 58.8% for SelfCheckGPT with GPT-3.5-turbo and 78.7% for the fine-tuned Llama-2-13B. At span level, GPT-4-turbo reached 28.3% and the fine-tuned model 52.7% [3]. The authors conclude that detection is still a challenge, particularly at span level, and that detection of obvious hallucinations is more successful than detection of subtle hallucinations. Using the fine-tuned detector to select among two candidate responses lowered the hallucination rate; for example, choosing a response with no detected spans reduced it by 63.2% for the smaller-model group and 51.0% for the GPT group relative to random selection [3].

Limitations. The authors recognize that there are other practical situations that are not addressed [3]. The level of agreement at the span is 78.8%, which is subjective, and the corpus includes 3 tasks and models from 2023 (our interpretation).

Contribution and relevance to RAG. It is the central RAG-specific resource in this review, and two other reviewed papers build on it [4], [5]. Relevance to RAG is direct.

D. ReDeEP: Detecting Hallucination in RAG via Mechanistic Interpretability [4]

Authors: Z. Sun, X. Zang, K. Zheng, J. Xu, X. Zhang, W. Yu, Y. Song, H. Li. Year: 2025. Venue: ICLR 2025. URL: openreview.net/forum?id=ztzZDzgfrh (no DOI identified; preprint arXiv:2410.11414).

Research problem. Despite providing accurate and relevant context, RAG models can still generate outputs that contradict the context. Existing detectors focus on one of the two knowledge sources (external context or parametric knowledge) or mix them without decoupling [4].

Methodology. An empirical study on Llama2-7B-chat with RAGTruth shows that hallucinations are associated with Knowledge FFNs adding too much parametric knowledge and Copying Heads failing to retain external knowledge. Two scores are defined: an External Context Score from attention heads and a Parametric Knowledge Score from feed-forward layers via the logit lens. ReDeEP regresses these scores to predict hallucination, at token level or chunk level. A similar mitigation technique, AARF, reweights the two contributions [4].

Results. The authors report that ReDeEP improves detection across two datasets, three backbones and several metrics. For example, on RAGTruth with LLaMA2-7B, ReDeEP (chunk) reached AUC 0.7458 and F1 0.7190, and with LLaMA2-13B AUC 0.8244 and F1 0.7587. The table also shows that ReDeEP is not best in every cell; for LLaMA2-13B on RAGTruth, LMvLM has higher recall

(0.8357 versus 0.7198 for chunk-level ReDeEP) [4]. Of 1,024 attention heads in Llama2-7B, 1,006 showed higher external-context scores on truthful than on hallucinated responses [4]. The authors' efficiency analysis indicates that chunk-level ReDeEP is faster than token-level and that methods without external models are faster than those calling them [4].

Limitations. The work targets cases where retrieved context is accurate and relevant [4]. It requires access to internal states and was evaluated on LLaMA-family models only (our interpretation of the reported setup).

Contribution and relevance to RAG. It links RAG hallucination to internal mechanisms, distinguishes it from knowledge conflict, and is designed specifically for RAG [4].

E. Rethinking Evaluation for LLM Hallucination Detection [5]

Authors: W. Chen, V. Padmanabhan, T. Giyahchi, E. Wong, L. Akoglu. Year: 2026. Venue: ACL 2026 (Volume 1: Long Papers), pp. 14912–14931. DOI: 10.18653/v1/2026.acl-long.680.

Research problem. Many benchmarks exist, but none exhibits all properties the authors consider necessary for effective evaluation. The largest gaps are the scarcity of long-context RAG-based benchmarks and the lack of realistic label noise for stress-testing detectors [5].

Methodology. The authors establish a desiderata for hallucination detection benchmarks, review existing benchmarks against it, and build TRIVIA+, a RAG-based benchmark with rigorous human annotation. They also design and share three sets of noisy labels with sample-dependent noise schemes in the published version [5]. A preprint version mentions four sets, so readers should rely on the published record. They then run popular state-of-the-art detectors on RAG-based benchmarks [5].

Results. The authors report that TRIVIA+ has the longest context in the literature, that ample room remains for detectors to approach the performance ceiling on RAG-based benchmarks, that a basic LLM-as-a-judge baseline performs competitively, and that label noise hinders detection performance [5].

Limitations. It contributes one new benchmark, and its noise is constructed by design rather than observed (our interpretation of the abstract).

Contribution and relevance to RAG. It is the only reviewed paper that treats long-context and noisy-label evaluation explicitly, and it focuses on RAG-based tasks [5].

F. Detecting Hallucinations Using Semantic Entropy [6]

Authors: S. Farquhar, J. Kossen, L. Kuhn, Y. Gal. Year: 2024. Venue: Nature, vol. 630, no. 8017, pp. 625–630. DOI: 10.1038/s41586-024-07421-0.

Research problem. There is a need for a general approach to identify hallucinations in LLMs even for novel queries for which humans might not have the correct answer [6].

Methodology. Entropy-based uncertainty estimators are calculated at the meaning level, not at the level of the individual word sequences, since one meaning can be conveyed in many ways. The target are arbitrary and incorrect generations (confabulations) [6] of hallucinations.

Results. The authors report that the method generalizes robustly to tasks not seen before [6].

Limitations. By design it covers confabulations only [6]. It needs multiple sampled answers (our inference) and does not use retrieved context.

Contribution and relevance to RAG. It gives a principled uncertainty-based detector that is not tied to surface wording. It is not RAG-specific; a probe-based approximation of semantic entropy (SEP) appears among the baselines in [4].

G. Lookback Lens: Detecting Contextual Hallucinations Using Attention Maps [7]

Authors: Y.-S. Chuang, L. Qiu, C.-Y. Hsieh, R. Krishna, Y. Kim, J. R. Glass. Year: 2024. Venue: EMNLP 2024, pp. 1419–1436. DOI: 10.18653/v1/2024.emnlp-main.84.

Research problem. When summarizing an article or answering from a passage, LLMs can produce details that are unsubstantiated with respect to the input context [7].

Methodology. The authors hypothesize that contextual hallucinations relate to how much the model attends to the context versus its own generated tokens. Per attention head, they compute the ratio of attention weights on the context to those on newly generated tokens and train a linear classifier on these lookback-ratio features [7].

Results. The linear classifier was as effective as a richer detector using full hidden states or a text-based entailment model, transferred across tasks and models without retraining from 7B to 13B, and, used in classifier-guided decoding, reduced hallucination by 9.6% on XSum [7].

Limitations. It needs attention maps, so it is unavailable for closed models (our inference). Transfer is

shown from 7B to 13B; other model families are not covered in the abstract.

Contribution and relevance to RAG. It targets context-grounded generation directly, making it closely related to RAG even though it is framed as contextual hallucination [7].

H. RAGAs: Automated Evaluation of Retrieval Augmented Generation [8]

Authors: S. Es, J. James, L. Espinosa Anke, S. Schockaert. Year: 2024. Venue: EACL 2024, System Demonstrations, pp. 150–158. DOI: 10.18653/v1/2024.eacl-demo.16.

Research problem. Evaluating RAG pipelines is hard because retrieval relevance, faithful use of passages, and generation quality all matter [8].

Methodology. A suite of reference-free metrics assesses these dimensions without ground-truth human annotations [8]. In the account given by [4], the RAGAs faithfulness metric breaks a response into shorter assertions, verifies each against the context with an LLM, and reports the ratio of supported statements.

Results. The authors claim that such a system accelerates the evaluation process [8]. As a third-party data point, [4] reports RAGAs as a baseline on RAGTruth with LLaMA2-7B at AUC 0.7290 and PCC 0.3865, using a GPT-4o-mini backend in that study [4].

Limitations. It depends on an LLM judge, with the cost and bias concerns discussed in [10] (our interpretation), and being a demo paper it reports limited validation in the abstract.

Contribution and relevance to RAG. It is designed for RAG and gives practitioners a ready-made faithfulness check alongside retrieval-quality metrics [8].

I. AlignScore: Factual Consistency with a Unified Alignment Function [9]

Authors: Y. Zha, Y. Yang, R. Li, Z. Hu. Year: 2023. Venue: ACL 2023 (Volume 1: Long Papers), pp. 11328–11348. DOI: 10.18653/v1/2023.acl-long.634.

Research problem. Earlier factual-consistency metrics depend on specific functions such as NLI or question answering trained on limited data and cannot assess diverse inconsistencies across inputs, outputs and tasks [9].

Methodology. AlignScore is built on a general function of information alignment between two text pieces, trained within a unified framework using 4.7M examples from seven tasks (NLI, QA, paraphrasing, fact verification,

information retrieval, semantic similarity and summarization) [9].

Results. The authors report substantial improvement over a wide range of earlier metrics and state that AlignScore matches or outperforms metrics based on ChatGPT and GPT-4 [9].

Limitations. It is a separate model to deploy, and for RAG the choice of comparison unit is left to the user (our interpretation).

Contribution and relevance to RAG. It is the entailment-style representative in this review. Any application to RAG checks a response against a context, so it is applicable but not RAG-specific [9].

J. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena [10]

Authors: L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica. Year: 2023. Venue: NeurIPS 2023, Datasets and Benchmarks Track. URL: neurips.cc/virtual/2023/poster/73434 (arXiv:2306.05685).

Research problem. Evaluating LLM-based chat assistants is difficult because their capabilities are broad and existing benchmarks do not measure human preferences [10].

Methodology. The authors study strong LLMs as judges on open-ended questions, examine their usage and limitations, including position, verbosity and self-enhancement biases and limited reasoning ability, and propose ways to mitigate some of them [10].

Results. A summary of the paper reports that GPT-4's judgments agreed with human preferences more than 80% of the time [10]; readers should consult the original for details.

Limitations. The above mentioned biases [10]. This paper is about the quality of chat, not hallucination.

Contribution and relevance to RAG. It provides the basic and the cautionary information for judge-based detectors. All of the above methods, such as prompt-based detection in [3] and [4] and the competitive LLM-judge baseline in [5] share this idea. It is indirectly relevant.

VI. COMPARATIVE ANALYSIS

Tables II and III make a comparison of the ten papers. The type of detection is the main category of Section IV. The comparison is qualitative and there is no ranking of

the papers. Numbers reported are from various datasets, models, splits and metrics and cannot be compared between papers. Even the two papers that share RAGTruth

[3], [4] differ in models, label treatment and evaluation protocol; for example, [4] annotated responses of LLaMA3-8B itself because RAGTruth had none for that model [4].

TABLE II COMPARISON OF THE REVIEWED PAPERS (PART A)

Paper	Year	Detection type	Black-box / white-box	RAG-specific?	Dataset	Main method
SelfCheckGPT [1]	2023	Consistency-based	Black-box	No	WikiBio passages generated by GPT-3, manually annotated	Sample several responses; score sentence consistency against samples
FActScore [2]	2023	Claim-level; retrieval-grounded	Black-box (retrieval + LM)	No (retrieval used for verification)	People biographies	Split into atomic facts; percentage supported by a knowledge source
RAGTruth [3]	2024	Benchmark with word-level labels	Model-agnostic	Yes	MS MARCO QA, Yelp data-to-text, CNN/DM and recent-news summaries	Manually annotated corpus; benchmarks prompt, SelfCheckGPT, LMvLM and a fine-tuned detector
ReDeEP [4]	2025	Mechanistic interpretability; model-internal	White-box	Yes	RAGTruth; Dolly (AC)	Regress External Context Score and Parametric Knowledge Score
Chen et al. [5]	2026	Benchmark; robust evaluation (long context, label noise)	Model-agnostic	Yes	TRIVIA+ and existing RAG-based benchmarks	Benchmark desiderata; new benchmark with noisy-label sets
Semantic entropy [6]	2024	Consistency-based (uncertainty)	Sampling-based	No	Several datasets and tasks (not named in abstract)	Entropy over meanings of sampled answers
Lookback Lens [7]	2024	Model-internal (attention)	White-box	Closely related (contextual hallucination)	Summarization and QA; XSum for mitigation	Linear classifier on lookback-ratio features
RAGAs [8]	2024	LLM-as-a-judge; retrieval-grounded	Black-box	Yes	Not stated in abstract	Reference-free metrics for retrieval, faithfulness and generation
AlignScore [9]	2023	NLI/entailment (alignment)	Trained checker	Applicable, not RAG-specific	22 evaluation datasets (19 unseen in training)	Unified alignment function trained on 4.7M examples
Zheng et al. [10]	2023	LLM-as-a-judge	Black-box	No	MT-Bench; Chatbot Arena	Strong LLM as judge; bias analysis

TABLE III COMPARISON OF THE REVIEWED PAPERS (PART B)

Paper	Evaluation metrics	Main contribution	Advantages	Limitations
[1]	AUC-PR (sentence); correlation (passage)	Zero-resource black-box detection	No external database or output probabilities needed	Consistency is not correctness (our interpretation); needs samples; weak on RAGTruth per [3]
[2]	FActScore; estimator error rate	Atomic-fact precision relative to a knowledge source	Fine-grained; handles partial support	Human evaluation costly; depends on source; evaluated on biographies

Paper	Evaluation metrics	Main contribution	Advantages	Limitations
[3]	Precision, recall, F1 (response and span)	First large-scale word-level RAG hallucination corpus	Natural responses; span labels; four types	Span detection still hard; three tasks; span agreement 78.8%
[4]	AUC, PCC, accuracy, recall, F1	Decouples context use from parametric knowledge	Explains as well as detects; no external judge	Needs internal access; LLaMA family only; assumes accurate context
[5]	Detector performance with clean and noisy labels (details in full paper)	Desiderata; long-context TRIVIA+; noisy labels	Stress-tests detectors	One new benchmark; constructed noise
[6]	Not detailed in abstract	Meaning-level uncertainty	No task-specific data; generalizes to unseen tasks	Confabulations only; sampling cost
[7]	Detection metrics in full paper; 9.6% reduction on XSum via decoding	Attention-ratio features	Simple; transfer across tasks and models	Needs attention maps; transfer shown from 7B to 13B
[8]	Suite of reference-free metrics	Reference-free RAG evaluation	No ground-truth annotations required	Relies on an LLM judge; demo paper
[9]	Not detailed in abstract	Unified alignment metric	355M-parameter model competitive with ChatGPT/GPT-4-based metrics	Separate trained model; comparison unit must be chosen for RAG
[10]	Agreement with human preferences	Systematic study of LLM-as-a-judge	Scalable and explainable	Position, verbosity and self-enhancement bias; limited reasoning

A. Written Comparison

Detection accuracy. Accuracy is only interpretable within a study. RAGTruth shows that fine-tuning a small model with in-domain labels outperforms prompt-based GPT-4-turbo at response and span level [3]. ReDeEP reports gains over many baselines but not uniformly in every metric [4]. Chen et al. conclude that room for improvement remains on RAG benchmarks [5]. None of the reviewed papers report a detector that is reliable in all settings.

Factuality and faithfulness. Factuality or confabulation is covered by SelfCheckGPT [1], FActScore [2] and semantic entropy [6]. Faithfulness to a context is covered by RAGTruth, ReDeEP, Lookback Lens, RAGAs and AlignScore [3, 4, 7–9]. FActScore is the bridge because its precision is relative to a chosen source [2].

Context grounding and explainability. Grounded methods use the context explicitly [3], [7]–[9]. Explanations differ in kind: span labels [3], atomic facts [2], attention features [7], and component-level mechanisms [4].

Long context and label noise. Only Chen et al. build long-context RAG data and noisy-label stress tests [5]. According to RAGTruth, the length of the responses and the length of the context (only for summarization) are correlated with hallucinations [3].

Applicability to real-world RAG. RAGTruth, ReDeEP and Chen et al. are built for RAG [3]–[5]; RAGAs offers a practical toolkit [8]; Lookback Lens and AlignScore are adaptable [7], [9]; the remaining papers supply components or cautions [1], [2], [6], [10]. The suitability is based on access, latency and risk profile of the application.

VII. RESULTS AND FINDINGS FROM THE REVIEW

A. Reported by Individual Studies

The following is the original authors' statement: Hallucinations are harder to detect than they are to see, especially at span level, in RAG [3]. (b) A small model that is fine-tuned on high-quality RAG annotations can outperform prompt-based GPT-4-turbo [3]. (c) RAG hallucination is correlated with high parametric knowledge usage and forgetting retrieved knowledge within the model [4]. (d) An attention-ratio classifier can be as rich as richer detectors and transfer across tasks and

models [7]. LLM-as-a-judge is a basic baseline that is competitive on RAG benchmarks and label noise makes detection difficult [5]. (f) There are known biases among LLM judges [10]. One of the shortcomings of previous benchmarks [5] was the lack of long-context grounded benchmarks and noisy labels.

B. Our Synthesis of the Literature

Evolution of techniques. The papers seem to be following a progression: output-only consistency and claim-level checks [1, 2, 6]; then context-grounded metrics and models [8, 9]; then benchmarks specific to RAG [3, 5]; and most recently internal-state and mechanistic signals [4, 7]. The order is not necessarily chronological, and previous families are still employed as reference families [3] and [4].

VIII. RESEARCH GAPS

The gaps below are only based on evidence in the reviewed papers.

No universally reliable detector. RAGTruth did not find any of the evaluated methods to be reliable at detecting hallucinations and span-level results were even less reliable. Similarly, Chen et al. claim that there is plenty of room for improvement [5].

Consistency vs Correctness vs Grounding. Sampling-based methods measure the agreement between generations [1, 6] while RAG needs agreement with context [3]. Only the RAG-specific studies compare these views on the same data, and RAGTruth found SelfCheckGPT unsatisfactory there [3].

Measuring true grounding. RAGTruth needed extra annotations for implicit truths and null values and found subtle hallucinations harder to detect than evident ones [3]. Whether a detector's score reflects real grounding remains hard to verify.

Retrieval quality. ReDeEP assumes accurate, relevant context [4], and RAGAs treats retrieval and generation as separate dimensions [8]. Detection under irrelevant, incomplete or conflicting retrieval is not studied by the ten papers.

Black-box versus white-box. The most explanatory RAG-specific detectors need white-box access [4], [7], while accessible ones show lower reported accuracy in the RAG comparisons [3], [4]. Bridging the two is an open problem.

Long-context RAG and noisy labels. Only [5] addresses both, and RAGTruth's context-length analysis is limited to three tasks [3]. Annotation agreement at span

level (78.8%) [3] and the reported effect of noise [5] show that labels themselves are a source of uncertainty.

Cost, real-time use and explainability. ReDeEP reports efficiency comparisons [4], but no reviewed paper evaluates latency in a production RAG pipeline. Explanations exist [2]–[4], [7], yet whether they help users or developers is not studied.

Multimodal RAG. None of the ten papers, as described in the sections analyzed for this review, addresses multimodal inputs.

IX. CHALLENGES IN REAL-WORLD RAG HALLUCINATION DETECTION

This section is mainly our synthesis, tied to evidence where available.

Retrieval problems. Real systems retrieve irrelevant, contradictory or redundant documents, and their sources change over time. The reviewed detectors mostly assume a fixed, reasonable context [3], [4]. RAGTruth limited each question to three passages to ease annotation [3], which is simpler than many production settings.

X. FUTURE RESEARCH DIRECTIONS

The following directions are suggested by limitations in the reviewed literature. None is a guaranteed solution.

Hybrid detectors. The use of retrieval side signals and model internal signals is a consequence of [4] and [7] that demonstrate the importance of internal signals, and [8] that separates retrieval from generation.

XI. CONCLUSION

Content generated by RAG that is not supported by or contradicts the retrieved context is considered hallucination in RAG, and can happen even if retrieval is correct [3] [4]. It is important to detect it since outputs are not trustworthy if they are not retrieved. The reviewed literature offers consistency-based, claim-level, retrieval-grounded, entailment-based, judge-based, model-internal and benchmark-oriented approaches [1]–[10].

The ten papers demonstrate a transition from output-level checks to grounding-aware and internal signals, and reveal that RAG-specific benchmarks reveal challenges, especially at span level, that are not found in general factuality evaluation [3]–[5]. There are still significant problems: there is no detector that is reliable in all settings, labels are subjective and noisy, long context and bad retrieval are little studied, and there are still different methods for accessible and explanatory methods. These limitations lead to directions like hybrid, retrieval-aware,

black-box compatible and noise-robust detection. This review is not intended to provide a solution to hallucination detection, but rather to outline the known and the unknown.

REFERENCES

- 1) P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in Proc. 2023 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Singapore, Dec. 2023, pp. 9004–9017, doi: 10.18653/v1/2023.emnlp-main.557. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.557>
- 2) S. Min et al., "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proc. 2023 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Singapore, Dec. 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.741>
- 3) C. Niu et al., "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers), Bangkok, Thailand, Aug. 2024, pp. 10862–10878, doi: 10.18653/v1/2024.acl-long.585. [Online]. Available: <https://aclanthology.org/2024.acl-long.585>
- 4) Z. Sun et al., "ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," in Proc. 13th Int. Conf. Learn. Representations (ICLR), Singapore, Apr. 2025. [Online]. Available: <https://openreview.net/forum?id=ztzZDzgfrh> (preprint: arXiv:2410.11414)
- 5) W. Chen, V. Padmanabhan, T. Giyahchi, E. Wong, and L. Akoglu, "Rethinking evaluation for LLM hallucination detection: A desiderata, a new RAG-based benchmark, new insights," in Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers), San Diego, CA, USA, Jul. 2026, pp. 14912–14931, doi: 10.18653/v1/2026.acl-long.680. [Online]. Available: <https://aclanthology.org/2026.acl-long.680>
- 6) S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: 10.1038/s41586-024-07421-0.
- 7) Y.-S. Chuang, L. Qiu, C.-Y. Hsieh, R. Krishna, Y. Kim, and J. R. Glass, "Lookback Lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps," in Proc. 2024 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Miami, FL, USA, Nov. 2024, pp. 1419–1436, doi: 10.18653/v1/2024.emnlp-main.84. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.84>
- 8) S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations, St. Julian's, Malta, Mar. 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16. [Online]. Available: <https://aclanthology.org/2024.eacl-demo.16>
- 9) Y. Zha, Y. Yang, R. Li, and Z. Hu, "AlignScore: Evaluating factual consistency with a unified alignment function," in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers), Toronto, ON, Canada, Jul. 2023, pp. 11328–11348, doi: 10.18653/v1/2023.acl-long.634. [Online]. Available: <https://aclanthology.org/2023.acl-long.634>
- 10) L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track, 2023, arXiv:2306.05685. [Online]. Available: <https://neurips.cc/virtual/2023/poster/73434>