Correctness is not Faithfulness in Retrieval Augmented Generation Attributions

Jonas Wallat*
L3S Research Center
Hannover, Germany
jonas.wallat@l3s.de

Maria Heuss*
University of Amsterdam
Amsterdam, The Netherlands
m.c.heuss@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Avishek Anand
Delft University of Technology
Delft, The Netherlands
avishek.anand@tudelft.nl

Abstract

Large language models (LLMs) have transformed information retrieval through chat interfaces, but their hallucination tendencies pose significant risks. While Retrieval Augmented Generation (RAG) with citations has emerged as a solution by allowing users to verify responses through source attribution, current evaluation approaches focus primarily on **citation correctness** – whether cited documents support the corresponding statements. This is insufficient and we introduce **citation faithfulness** – whether the model's reliance on cited documents is genuine rather than post-rationalized to fit pre-existing knowledge. Our contributions are threefold: (i) we introduce coherent notions of attribution and introduce the concept of citation faithfulness; (ii) we propose desiderata for citations beyond correctness and accuracy needed for trustworthy systems; and (iii) we emphasize evaluating citation faithfulness by studying post-rationalization. Through experimentation, we reveal prevalent post-rationalization issues, finding that up to 57% of citations lack faithfulness. This undermines reliable attribution and may result in misplaced trust, highlighting a critical gap in current LLM-based IR systems. We demonstrate why both citation correctness and faithfulness must be considered when deploying LLMs in IR applications, contributing to a broader discussion of building more reliable and transparent information access systems.

CCS Concepts

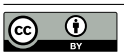
• **Information systems** → **Language models**; • **Computing methodologies** → **Natural language processing**.

Keywords

Large language models; Retrieval-augmented generation; Attributions; Interpretability; Faithfulness; Self-explanations

ACM Reference Format:
Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not Faithfulness in Retrieval Augmented Generation Attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3731120.3744592

*Equal contribution.



ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744592

1 Introduction

Recent years have shown great improvements in large language models (LLMs) and a steep increase in the adoption of chat systems for different tasks, such as information access. They can improve information accessibility through their interactive nature, the possibility to interact with information in a foreign language or the use of simple language. The adoption of these systems spans lots of different societal applications, ranging from healthcare [77] and legal systems [61] to education [34]. Trustworthiness of AI systems is key to their responsible deployment and usage in high-stakes scenarios, particularly in such high stakes domains [29, 32].

A critical challenge in these systems are hallucinations, where LLMs generate plausible but incorrect or fabricated information, potentially undermining their reliability and disproportionately affecting vulnerable populations who may rely on these systems for critical information access [3].

One promising approach to address hallucinations is enabling text generation that is explicitly grounded in retrieved source documents and accompanied by citations [7, 54], which is often operationalized through **retrieval augmented generation (RAG)**. RAG employs a two-step process: first, retrieve relevant documents and then use them to generate answers. While citations cannot eliminate hallucinations, they enhance verifiability by explaining the origin of information [42]. This grounded text generation approach [22] has been successfully applied to various NLP tasks, including summarization and question answering. Recent implementations of RAG mechanisms [38] ensure that content remains coherent, contextually relevant, and anchored in verifiable sources [7].

We investigate the faithfulness of citations in RAG, examining whether cited documents genuinely contribute to the answer generation process or are merely superficially referenced. We conceptualize citations as a form of LLM (self-) explanation that should give insight into the source of generated information, analogous to how chain-of-thought explanations reveal a model's reasoning process. This analogy raises important concerns, as recent research [13] has demonstrated that even reasoning models, which should benefit from coherent chains of thought, frequently exhibit unfaithful behavior by omitting crucial information from their reasoning chains that was evidently used in generating answers.

Current evaluation practices for attributed text focus primarily on two aspects: the **correctness of the answer** and the **correctness of citations**, which is based on the agreement between attributed statements and the information found in referenced source documents. Citation correctness, sometimes called answer faithfulness [25], measures the extent to which cited documents support a generated statement.

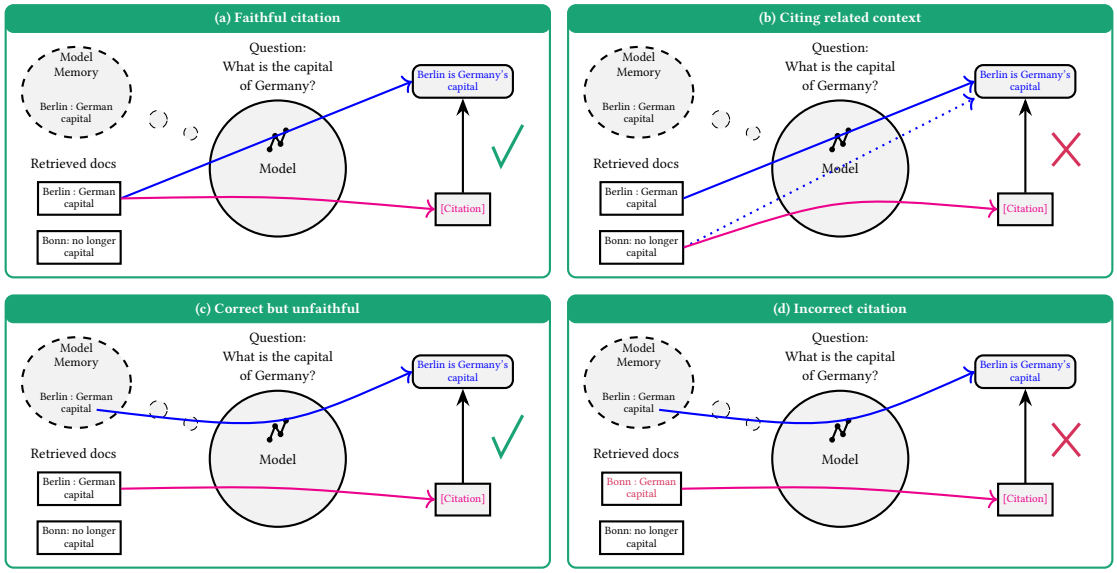


Figure 1: Different answer scenarios for the query "What is the capital of Germany?" (a) The ideal case, i.e., a correct citation that is faithful to the answer's generation process. (b) A citation referring to the context that was used during the answer generation but does not contain the statement itself. (c) A correct but unfaithful citation, where the model post-rationalizes a citation to fit its prior. (d) An incorrect citation.

We argue that ensuring mere correctness is insufficient for reliable information retrieval systems. This is particularly evident in domains such as legal IR [46] and medical question answering [37], where documents are complex and responses are vulnerable to model biases. In these contexts, simple fact-checking or correctness evaluation may prove inadequate, requiring instead a nuanced understanding of document content. To address this challenge, we propose providing guarantees that cited documents were actually used during answer generation, enabling users to assess response trustworthiness based on source credibility rather than blind faith in model outputs. However, simply providing citations is insufficient, as both unwarranted trust and excessive skepticism toward model outputs can have significant consequences.

This concern is worsened by research showing that explanations can paradoxically increase user trust even when misleading [59], particularly in tasks where output verification is challenging. The core issue lies in distinguishing between genuine document usage and **post-rationalization**, where models cite sources to fit preconceived notions derived from parametric memory rather than authentic retrieval-based reasoning. To address this challenge, we introduce the concept of **citation faithfulness** – defined through a causal relationship where cited documents directly influence the generation of corresponding claims, rather than being superficially appended to justify predetermined answers. Figure 1 illustrates the differences between faithful and unfaithful behavior, as well as correct and incorrect citations.

When building trustworthy IR systems that offer self-explanations – in this case, citations – we should strive to convey the system's decisions accurately. Only if the produced citations are faithful to the underlying processes can we enable justified trust (as opposed to misplaced trust if faithfulness breaks down).

Our contributions are threefold: First, we offer coherent notions of attribution and citation in the context of grounded generation and introduce the concept of citation faithfulness. Second, we propose desiderata for citations that go beyond correctness and accuracy and are needed for trustworthy and usable systems. Third, we emphasize the need to evaluate the faithfulness of citations by studying post-rationalization. Our experiments reveal the existence of unfaithful behavior, with up to 57% of citations being post-rationalized.

Our work on disentangling citation correctness and faithfulness in grounded text generation using LLMs aims to create more reliable IR systems by ensuring accurate and contextually faithful citations. By focusing on post-rationalization, we enhance accountability, helping IR systems avoid propagating biases or misinformation, thus promoting ethical standards in information dissemination and ensuring these systems effectively serve all users, regardless of their technical expertise or background.

2 Related Work

We summarize relevant background and position our work w.r.t. risks of LLMs, the evaluation of attributed generation, faithfulness in interpretability, and faithfulness of self-explanations. The area of knowledge conflicts [76] examines information flow and whether answers originate from parametric memory or the contextual [51, 71]. Its goal of understanding models is similar, but has a different focus (full answers vs. citations) and is therefore out of scope.

2.1 Risks of LLMs

Recent work in the field of responsible AI has identified numerous risks associated with the deployment of LLMs in real-world applications. These risks span multiple domains, from security vulnerabilities and susceptibility to adversarial attacks [74], environmental concerns [55], and challenges related to the trustworthiness

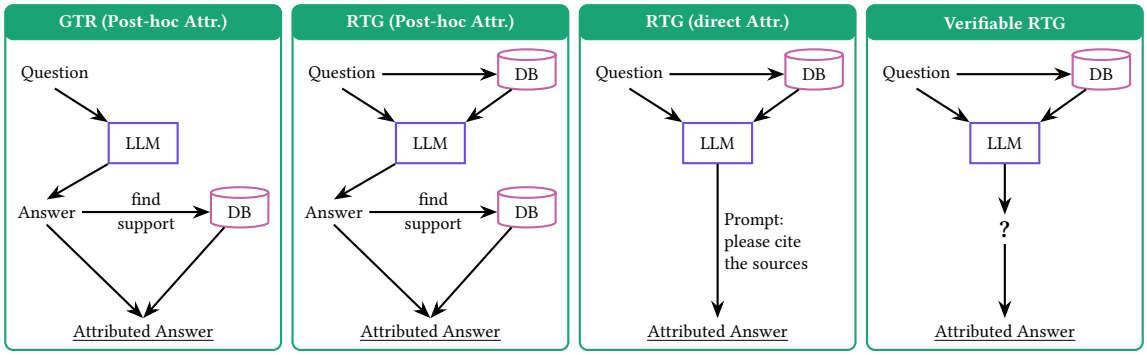


Figure 2: Different methods of attribution generation, using information from the database (DB) at different stages of the generation pipieline. The likelihood for un-faithful behavior and post-rationalization decreases from left to right.

of these systems and their alignment with social norms, values, and regulations [43]. Our work focuses on the risk of unreliable or incorrect information being presented as authoritative and trustworthy. LLMs are known to produce hallucinated information that may be inconsistent with real-world facts or entirely unverifiable [40]. These fabricated "facts" can become sources of misinformation, since the presence or absence of citations can influence users' trust in the presented content [56]. The risk of misinformation becomes particularly concerning when considering demographic variations in susceptibility. Research has shown that certain population groups, including younger individuals, those with lower levels of education, and racial minorities, are especially vulnerable to health misinformation [49]. This vulnerability is particularly troubling given the increasing use of LLMs in high-stakes domains such as healthcare, where misinformation has long been a significant concern among public health practitioners and researchers [15, 65]. The expansion of LLM applications into sensitive domains such as emotional support, financial advice, medical advice, and legal assistance [28] raises additional concerns. For instance, the use of LLMs for self-diagnosis purposes has been identified as a potential new vector for health misinformation [5].

These applications highlight the critical need for robust safeguards and regulatory frameworks. Current regulatory efforts, such as the EU AI Act, attempt to address these risks, though some argue existing frameworks are inadequate for the challenges posed by generative language models [27]. Although legislative frameworks may need refinement, the documented risks associated with LLM-powered information systems underscore the technical community's responsibility to anticipate failures and develop responsible solutions. This work aims to contribute to that effort by examining post-rationalization and unfaithful citations in LLM-powered advice systems which might become sources of misinformation.

2.2 LLMs and Attributions

Supplying LLM-generated answers with attributions aims to improve the quality of the generated answers [24], reduce hallucination [67], and improve users' trust [48] in the generated outputs. Methods for generating attributed answers range from prompting [24], adding post-hoc attributions [24, 66], and training paradigms [4, 11, 48, 66, 78] to generation-planning for more fine-grained citations [64]. Figure 2 provides an overview of common methods. The simplest method is generate-then-retrieve (GTR), a paradigm in which a model produces an answer (without attributions), and supporting evidence is added in a subsequent step [7, 24]. Retrieve-then-generate (RTG) operates similarly, but the model produces the (unattributed) answer after seeing both the question and the retrieved documents. As with GTR, RTG produces attributions in a second retrieval step, independent of the initially retrieved documents [78]. Thus, both GTR and RTG have post-hoc attributions, which are unfaithful to the model by design, i.e., the citation does not reflect the model's decision-making during the answer generation process. It is, however, possible to directly generate attributed answers by prompting the RTG model to do so [7, 24]. The resulting attributed answer *may* be faithful to the model's decision process, but we lack guarantees. As we show below, there is a significant chance of unfaithful behavior. The ultimate goal of attributed answer methodologies is to verify that certain information in the answer *originates* from the source document.

2.3 Evaluation of Attributed Generation

Attributed generation is a complex process that requires evaluation across multiple dimensions. One dimension is the *usability* of the generated response, which includes factors like fluency and perceived utility [42]. Traditionally, these factors have been assessed through user studies and automatic evaluation methods [24]. Other important dimensions include *answer relevance*, which measures how well the response addresses the question, and context relevance, which looks at the compactness of the retrieved context [20]. Datasets like HAGRID [33] are useful for evaluation, with human evaluations of the informativeness and attributability of the responses, which can be used to measure overlap with gold citations [18]. Weller et al. [73] use the QUIP-Score, a method based on n-gram comparisons, to measure grounding and quoting from model pre-training data.

Next to the generated answer, the citation to the referenced document needs to be evaluated, too. To this end, prior work often uses natural language inference (NLI) classifiers [7, 23]. These help evaluate citation precision, which measures the average correctness of citations, and comprehensiveness/citation recall, which quantifies the proportion of accurately cited statements in all statements [18, 39]. The correctness of citations is a major focus in prior work [1, 18, 25, 39, 47, 54, 57, 58, 79]. We differentiate between

citation correctness and the related but distinct aspect of **citation faithfulness**. Citation faithfulness requires a causal relationship between the cited document and the generated statement, an area that has so far received little attention.

2.4 Faithfulness in Interpretability

In retrieval-augmented generation (RAG) attributions, (citation) faithfulness has not been studied much. In contrast, the evaluation of faithfulness of explanations has been studied extensively. Here, faithfulness refers to how accurately an explanation reflects the model's decision-making process, clearly differentiating it from explanation plausibility [31]. It lacks a universally accepted formal definition and is often defined in an ad-hoc manner [44]. Faithfulness establishes a causal relationship. Various methods have been proposed for evaluating faithfulness: (i) axiomatic evaluation, (ii) predictive power evaluation, (iii) robustness evaluation, (iv) perturbation-based evaluation, (v) white-box evaluation, and (vi) human perception evaluation [44]. Twelve desirable properties of explanations have been identified by Nauta et al. [50], including correctness (of explanations), which is equated with faithfulness. Overall, the concepts of faithfulness and correctness appear entangled in the explainability literature. We take a step towards disentangling those two aspects for attributed text. Inspired by Lyu et al. [44], we consider the causal relationship between the attributed text and generated answer to be a fundamental condition of faithful attribution.

2.5 Faithfulness of LLM Self-Explanations

Self-explanations are explanations that an LLM is prompted to generate along with the answer to a posed question. Self-explanations have been divided into (i) chain-of-thought (CoT) reasoning, which involves generating a sequence of intermediate steps that lead to the response [72], (ii) token importance, which highlights tokens that significantly influence the response generation [41, 75], and (iii) counterfactual explanations, which provide insights into how different inputs might lead to a different response [2]. Faithfulness of self-explanations has recently received attention [2, 36, 45, 68], with work on evaluating faithfulness [36, 68] and its importance in contrast to plausibility [2]. There is high variation in how much LLMs use CoT on different tasks, some relying upon it heavily, others merely generating it in a post-hoc manner [36]. Recent work on evaluating the faithfulness of reasoning models reveals that even models explicitly trained for reasoning tasks exhibit an astonishing level of unfaithful CoT reasoning [13, 16]. We view attributed generation that generates citations along with the text, rather than post-hoc, as a special class of self-explanation. We use a similar evaluation strategy as was previously used for the evaluation of faithfulness for CoT explanations [68] to show that similar faithfulness concerns arise for attributed generation as for CoT reasoning. We identify the problem of post-rationalization, which is closely related to post-hoc reasoning [36].

3 Attributions

RAG systems provide a way of grounding LLM-generated answers in documents that are retrieved from a corpus. By ensuring high quality of information in the corpus, this can improve the quality of the generated answers. RAG operates in two stages, where the first stage retrieves documents that match the information need/query of the user, and the second stage uses the retrieved documents to generate an answer. In the context of attributed text generation in RAG, an answer may be accompanied by references to documents, emphasizing that certain information originates from the referenced document. Merriam-Webster defines the verb *to attribute* as explaining by indicating a *cause*, emphasizing the causal nature.[1]

3.1 Notation

Let $A = \{a_i\}_i$ be a set of retrieved documents and let s be a text snippet, i.e., a factual statement that needs to be grounded in the retrieved documents A. A citation $cit : s \mapsto a_j \in A$, or simply (s, a_j), connects a statement to a document that supports the stated statement. We use the term *attribution* to refer to the referenced document a_j or the process of referencing source documents.

Example 1: Attributed Answer

Question: Whats the biggest penguin in the world?
Answer: The Emperor Penguin [0] is the tallest [0] or biggest penguin in the world.

In Example 1, "tallest" would be a factual statement s attributed to document $a = 0$ through the citation ("tallest", 0). We note that many attributed statements are underspecified. Therefore, we distinguish between the statement ("tallest") and the underlying *claim* ("Emperor penguin: tallest: in the world"). Ideally, a citation should map claims to documents, but it is currently operationalized as statement to document, which can cause problems of misalignment between those two.

When attribution generation is integrated with answer generation, citations can be considered a form of self-explanation, others being chain-of-though explanations [72], explain-then-predict and predict-then-explain frameworks [10], and counterfactuals [14].

3.2 Desiderata for Good Attributions

Here we define several dimensions that can make attributions good or bad; Table 1 provides an overview.

Table 1: Desiderata for good attributions.

Desideratum	Description
Correctness	The attribution accurately represents the content of the cited document.
Faithfulness	The attribution accurately represents how the model derived its answer.
Appropriateness	The attribution is relevant and meaningful, not noisy or irrelevant.
Comprehensiveness	The attributions cover all the key points in the answer.

Correctness. Most importantly, good citations should be correct, meaning that the cited documents should support the generated

[1] https://www.merriam-webster.com/dictionary/attribute#h2

statement. Ensuring correctness in attribution is crucial for maintaining the integrity and reliability of the information being presented. However, there are several ways in which the outputs of an LLM can be right or wrong.

Wrong answers. A direct way in which an LLM-generated answer can be wrong is if the statement itself is wrong, not matching the ground truth answer. This is the property that is evaluated most frequently in the open-domain QA and attribution literature [e.g., 7, 18, 38]. Wrong answers can result from hallucinations or correct attributions from a document containing false information. Therefore, an answer can be wrong despite having proper citations.

Hallucinated attributions. Attributions that do not exist, i.e., when a model hallucinates a reference to a non-existing document, are relatively easy to spot. LLMs without a retrieval component, such as the early versions of ChatGPT, especially, commonly generate broken links or hallucinate titles and authors of the source document from which certain information should come.

Wrong citations. Attributions can be incorrect, for example, by misrepresenting the content of the attributed documents or by attributing claims from document a to document b. In these cases, the citation (s, b) is incorrect. Compared to answer correctness, less work focuses on the correctness of attributions. Attributions are usually evaluated by testing if the attributed document implies the statement. To do so, recent work employs NLI models [7, 18, 23].

Appropriateness & Comprehensiveness – *What* do we cite? Besides unfaithful behavior and incorrect attributions, bad citations may (appear to) be inappropriate or non-comprehensive and, therefore, dilute our understanding or evaluation of the answer. Appropriateness of attributions means that the attribution should be relevant, understandable, and meaningful; comprehensiveness refers to covering all the key points in the answer. The question of how much we need to cite and whether attributions cover the important claims are less prominent in current evaluations frameworks, but these aspects may heavily skew the results of other evaluation metrics like correctness.

Example 2: Inappropriate Citations

Question: how long was gabby in a coma in the choice
Answer: In the novel [0, 4] the choice [0, 3, 4], Gabby is in a coma for three months.

Inappropriate citations. In Example 2, neither citation offers much value given the question. Attributing the title "the choice" provided in the question to documents 0, 3, and 4 offers no additional insights. On the contrary, when evaluating the quality of the provided citations, common approaches average over all existing citations. A large number of such *low-value* citations, which re-state information from the question, may heavily skew the evaluation metrics.

Short statements – What is the actual claim? Capturing a comprehensive, standalone statement in an LLM-generated response that maintains specificity even when separated from the rest of the text can be complex. The statement is often reduced to a single word or concept, subtly referring to other parts of the generated response. In our example, it remains ambiguous what the highlighted word "novel" pertains to (i.e., the actual claim). This lack of clarity makes interpreting and evaluating such references more challenging.

For which statements do we need a citation? An answer may contain several citations, but one may be missing for the factual answer to the question. In the above example, the focus of the question is the time that Gabby spent in a coma ("three months"). This is the most critical statement in the answer and should be attributed to a source document. The above answer is not comprehensive since a central requested fact is not attributed to any source.

Faithfulness – Right for the *wrong* reason? Can an attribution be correct and still be bad? Like model explanations, attributions can be right for the wrong reason. To judge whether an attribution is right for the wrong reason, it is key to understand the internal model processes and understand whether a document a was considered during answer generation. If a is cited for another reason, then the attribution is not faithful to the underlying model behavior. Importantly, unfaithful attributions might still be factually correct and, therefore, difficult to spot – yet they foster misguided trust.

Post-rationalization. We hypothesize that post-rationalized attributions are a special case of unfaithful behavior. In this setting, an LLM's parametric memory produces an answer to the question, and the model looks for support in the documents in some shallow way (e.g., by token-matching). The resulting citation is not faithful since the attribution superficially maps to a document, while using the model's internal knowledge. Let us consider Example 3.

Example 3: Faithfulness, Post-rationalization, Correctness

Question: What is the capital of Germany?
Answer: The capital of Germany is Berlin [1, 2]

Document 1: The capital of Germany is Berlin [...]
Document 2: Berlin has the best night-life [...]

Faithful (right for the right reason): Citing document 1 because the LLM used document 1's information to generate the answer.
Post-rationalized but **correct** (right for the wrong reason): Citing document 1 because the model knows the answer and finds a document that agrees with its priors.
Post-rationalized and **wrong**: Citing document 2 because the model knows the answer, and the answer token is mentioned in document 2.

Since the outputs in the faithful and unfaithful cases are identical (citing document 1), unfaithful behavior is hard to identify. We propose that a comprehensive evaluation of faithfulness must consider both the *attributions* themselves and the *process* through which they are derived. Given that citation faithfulness and correctness have often been conflated in previous research, we provide a detailed discussion and definition of citation faithfulness in Section 4.

4 Citation Faithfulness

The Cambridge Dictionary defines *faithful* as "true or not changing any of the details, facts, style, etc. of the original."[2] In the explainability literature, a "faithful explanation should accurately reflect

[2] https://dictionary.cambridge.org/us/dictionary/english/faithful

the reasoning process behind the model's prediction" [31]. Lyu et al. [44] further clarify that faithfulness establishes causality, distinguishing between "what is known by the model" and "what is actually used in making predictions."

Prior work on attributed answer generation defines *answer faithfulness* as the extent to which the cited document supports the generated statement [79]. Answer faithfulness considers the answer itself rather than the citation. In the context of the citation, this property is often called the correctness of the citation. In this work, we define **citation faithfulness** and disentangle the concepts of *answer faithfulness/correctness* and *citation faithfulness*. Prior work on attributed answers often has defined faithfulness loosely, for example, as "whether the selected documents influence the LLM during the generation" [53]. We take inspiration from the rich literature on the faithfulness of explanations and define the faithfulness of citations through a casual dependency of the generated answer and referenced document.

Definition 1: Citation Faithfulness

Let s be a generated statement underlying claim c. Let $A = \{a_i\}_i$ be a set of documents that the model has retrieved as context. We call (s, a_j) a faithful citation if:

- $a_j \in A$,
- The underlying claim c is supported by a_j (correctness), and
- c is causally impacted by a_j.

The second condition, often referred to as the "correctness" of the citation, has been a focal point in previous studies evaluating RAG attribution. Correctness tests whether a statement or claim is supported by the attributed document (measured by NLI models). However, while correctness is a necessary condition for faithfulness, it is insufficient. For a citation to be deemed faithful, the model must also rely causally on the cited document to generate the answer so that information flows from the document to the generated claim. Evaluation of this causal dependence of the model output on the cited statement has been largely overlooked, which is why we advocate for increased attention to the topic in future research.

We recognize that our definition of faithfulness is somewhat abstract. As Lyu et al. [44] observe, formulating a concrete definition with a single, comprehensive test to evaluate explanation faithfulness remains an open challenge – one that extends beyond our field to explanation methods in general. Thus, a set of more tangible necessary conditions with corresponding tests should be established in practice. These can assist in approximating the level of faithfulness of specific explanations. Consider the following examples of more concrete necessary conditions for faithful attribution. For a citation (s, a) to be considered faithful, the following should hold:

(1) If the relevant information in the cited document a is altered, the model should either provide a different generated statement s or modify the decision-making process. This could involve using different evidence a' or the model's memory to generate the answer.
(2) Adding irrelevant documents to the context should not affect the attribution, provided that the answer remains unchanged.

In Section 5, we design and implement an experiment to test the second necessary condition, providing empirical evidence for post-rationalization. While the first condition might offer broader insights into model faithfulness, testing it directly would require a deeper understanding of the model's internal decision-making process. Current analytical techniques are insufficient for this level of investigation. Therefore, we leave this analysis to future work.

5 Post-Rationalization – A Study of Unfaithful Behavior

We study attributions of a prominent RAG model and produce evidence of unfaithful behavior. As Jacovi and Goldberg [31] argue, faithfulness, as opposed to plausibility, should not be measured through human evaluation. Therefore rather than doing a human study to evaluate the quality of the citation self-explanations, we deploy a test based on input-output relationships. We investigate a particular case of unfaithful behavior, post-rationalization, i.e., the process in which a model generates a prior answer from model memory without regard to the documents and then searches retrieved documents to find supporting evidence.

5.1 Setup

Cohere's Command-R+ model is a "RAG-optimized" LLM specifically trained to produce grounded answers.[3] It has 104B parameters and a context length of 128k tokens, which we use in 4-bit quantization to run on a single NVIDIA A100 GPU. We evaluate Command-R+'s attributions on the NaturalQuestions QA dataset, containing 1,444 real user questions answered by Wikipedia pages [35]. We use the temporally-aligned KILT [52] Wikipedia dump[4] as a retrieval base. Following [17], we split passages into chunks of 100 tokens and prepend the title of the page to the chunks. We index the resulting chunks and, for each query, retrieve the top 30 documents using BM25. We rerank the 30 retrieved documents using ColBERT v2 [60] and feed the top 5 documents together with the question into Command-R+.

We use the grounded generation prompt template provided by Cohere.[5] The grounded generation pipeline with Command-R+ follows four steps: (i) predict the relevance of the retrieved documents; (ii) predict which documents should be cited; (iii) produce an answer without citations, and (iv) one with citations. This setup makes Command-R+ a retrieve-then-generate (RTG) model with direct attributions via prompting (see Figure 2). We selected an instance from this class of models since its chances of faithful behavior are higher than in the case of post-hoc attributions.

5.2 Citing Behavior

As an initial step, we study the answers and attributions performed by Command-R+. Figure 3 provides an overview. The model produces relatively short answers with on average 2.4 sentences and roughly five citations per answer. The cited spans (statements) have a median length 4 tokens and are, thus, relatively short. Further looking at individual documents (Figure 4), we see that Command-R+ cites on average 3 documents for a given statement, with almost

[3] https://cohere.com/blog/command-r-plus-microsoft-azure
[4] Available here: https://huggingface.co/datasets/facebook/kilt_tasks.
[5] https://huggingface.co/CohereForAI/c4ai-command-r-plus

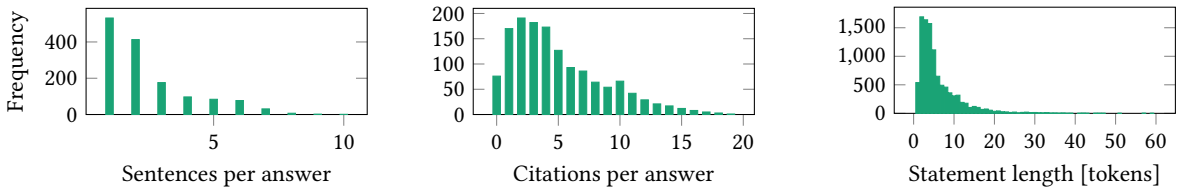


Figure 3: Statistical analysis of citations performed by Command-R+ on NaturalQuestions.

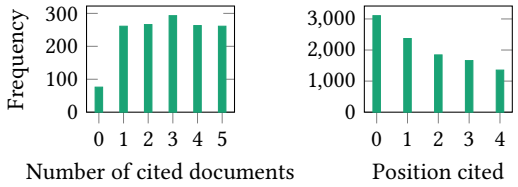


Figure 4: Number of cited documents and position of the cited documents in the input.

equal frequency of 1–5 documents being cited. We also find 76 instances where the model did refrain from answering and, therefore, cited nothing. With regard to the position of the cited documents, we observe a tendency to citing the first documents. The first document is cited more than twice as much as the fifth document. However, since we order the input documents by reranking scores, it is to be expected that the earlier documents are more relevant.

To better understand Command-R+'s grounded generation process, we also investigate whether the model cites the documents it predicted to be relevant and to be cited (step 1 and 2, c.f. Section 5.1). We present the results in Table 2. While it is expected that the model does not cite all documents it predicted to be relevant, it is somewhat surprising that it only cited 46% of the documents it predicted to be cited. In the remaining 54%, the model cited either nothing, fewer documents, or some documents it did not predict to be cited (1%). We hypothesize that the model was specifically trained to cite only the documents selected in the earlier processes. Furthermore, we did not find the model hallucinating attributions (e.g., citing document IDs other than the five retrieved documents). Nevertheless, the large mismatch between the documents predicted to be cited and the actual citations lets us question the faithfulness of the model's attribution behavior.

Table 2: Investigation into the citing behavior of Command-R+. We explore whether all documents that the model predicted as relevant (step 1 in Command-R+'s grounded generation) and predicted to be cited (step 2) were cited in the grounded answer (step 4).

Split	Pred. Rel.	Pred. Cited
Cited all selected documents	636	820
Cited less than all selected documents	708	522
Cited not selected documents	8	12
Cited nothing	76	76

5.3 Unfaithful Attributions

We devise the following experiments to better understand the extent to which Command-R+ post-rationalizes citations. One possible way of post-rationalization could be finding documents to cite by token matching, so we (i) generate attributed answers for QA pairs, and (ii) select statements from these answers and append them to other documents. Since statements are, usually, around 4 tokens, they mostly contain short concepts such as "Emperor penguin" or "The Choice," which should not be cited when appearing without factual context. We append these adversarial statements into three kinds: random documents retrieved using BM25 for arbitrary questions, documents predicted to be relevant in step 1 of Section 5.1 but never cited, and documents cited for other statements in the attributed answer step 4 of Section 5.1. The created dataset with adversarial documents consists of 1,344 QA pairs (random), 702 (relevant but not cited), and 829 (cited for other reasons). In step (iii) we again generate attributed answers, but this time with our adversarial documents. In the case that the adversarial document was created from a random document, we append it to the list of documents in the context. If the original document was part of the context, we substitute it with the adversarial one. Lastly, (iv) we observe whether the model now cites our adversarial documents for the statements selected in step (ii). We operate under the assumption that citing documents that just randomly contain the statement ("Emperor penguin") indicates *post-rationalization*. This process is also depicted in Figure 5.

The results are presented in Figure 6. First and foremost, we note that recovering the old statement in the newly generated answer worked in 63–70% of the cases, while the generated answer changed at least to some extent in the remainder of the cases. Since a change in answer statement might reflect a change in the used attention mechanisms and makes it impossible to compare citations for previously generated statements with newly generated ones, we discard those cases. This is necessary to understand if the adversarial document has been cited for the same statement. By injecting the statement into random documents and passing them to the model, the model cited these documents in 12% (116/936) cases. Interestingly, the number of adversarial documents cited is much higher when forging relevant but uncited documents (57%) and documents cited for different reasons (55%). Based on our results, we conjecture post-rationalization to be a *common phenomenon*. We additionally present an example of Command-R+'s post-rationalization behavior, citing a random adversarial document below:

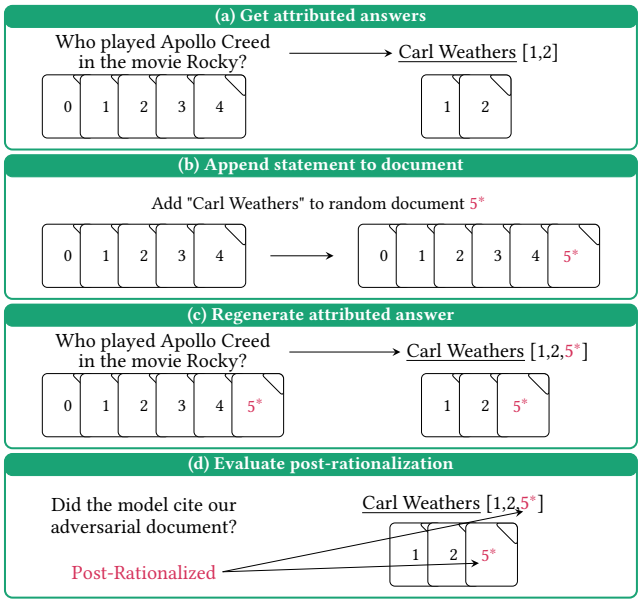


Figure 5: Experimental setup of the post-rationalization experiment. We inject attributed statements into random documents and regenerate the answer to see if the model cites unrelated documents when injected with statements.

Example 4: QA pairs with adversarial random documents added

Question: who played apollo creed in the movie rocky
Answer: Carl Weathers [1,2,5*] played Apollo Creed in the Rocky films.
Adversarial Document 5*: 1974 State of the Union Address
The 1974 State of the Union Address was given to the 93rd United States Congress, on Wednesday, January 30, 1974, by Richard Nixon, the 37th President of the United States. He said, "We meet here tonight at a time of great challenge and great opportunities for America. We meet at a time when we face great problems at home and abroad that will test the strength of our fiber as a nation. But we also meet at a time when that Carl Weathers

Are the adversarial documents actually adversarial? Our first experiment is based on the assumption that adversarial documents, generated by appending statements to unrelated documents, do not contain the actual claims. If this assumption does not hold, the model might be able to use the information within the document to generate the answer, hence the citation referencing this document might in fact be faithful. Therefore, to verify our estimation of post-rationalization, we examine whether adversarial documents alone might be sufficient to generate the investigated claims. Recall from Section 3.1 that we differentiate between the text snippet that the citation is referencing called statement s ("Carl Weathers" in Figure 5) and the underlying claim c ("Carl Weathers played Apollo Creed in Rocky"). Since statements are typically quite short (median of 4 tokens, cf. Figure 4, right), we do not expect that adding

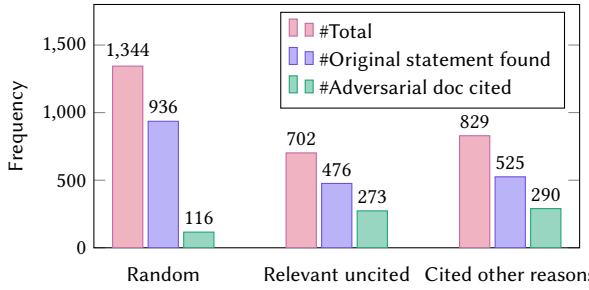


Figure 6: Results of the post-rationalization tests. We measure the cases in which the model cited our adversarial document (which had the previously cited statement appended). Since we also change the input, the model is not guaranteed to produce the same statements again. Thus, we also include the number of cases where we could match the old statement.

statements alone provides sufficient context to generate an answer (or claim) using only information from the adversarial documents. To validate this assumption, we conducted an additional experiment focusing on instances where the model cites adversarial documents.

We ran inference using three different context configurations (instead of the full list of retrieved documents (e.g., [0,1,2,3,4]):

(a) The complete set of originally cited documents for the corresponding statement (e.g., [1,2] in Figure 5).
(b) One randomly sampled document that was originally cited for the corresponding statement (e.g., [2] in in Figure 5).
(c) The adversarial document alone (e.g., [5] in Figure 5).

We hypothesize that the model should recover original statements more frequently from previously cited documents than from adversarial documents, assuming that at least some of the original citations were faithful to their sources.

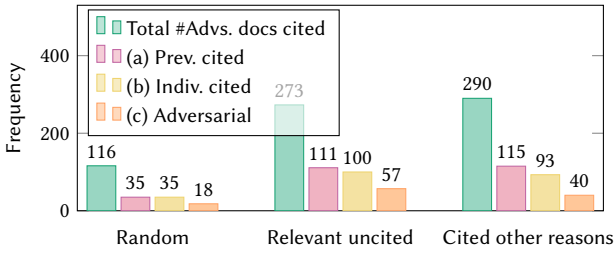


Figure 7: Frequency of statement recovered when providing different types of context documents to the model. The bars show the total number of adversarial instances investigated (green), and the number of recovered statements when all previously cited documents (pink), only a single cited document (yellow), or only the adversarial document (orange) is used as context.

Results are presented in Figure 7. We found that adversarial documents alone are only sufficient to recover the statement in 14–21% of cases, varying by the document type used for adversarial generation. In contrast, contexts containing cited documents yielded much higher recovery rates of 30–43%. These findings were unexpected, as we anticipated higher recovery rates from originally cited documents and near-zero recovery from adversarial documents.

The relatively low recovery rates from cited documents might stem from the inherent instability of language generation, as can also be seen in Figure 6, where simply adding irrelevant adversarial documents to the context reduces answer consistency, as well as the potential existence of unfaithful citations in the original answers.

On the other hand, several factors may explain the non-zero recovery from adversarial contexts: (i) the model might generate answers from its parametric memory and then match tokens to create citations, though this process is not directly observable; (ii) reasoning models change answers based on subtle hints in prompts without faithfully reflecting these changes in their reasoning [13] — similarly, the appended statements in our adversarial documents might serve as subtle hints pointing to plausible answers; and (iii) a few adversarial documents may contain the target claims, although preliminary qualitative analysis suggests otherwise.

Considering the big differences in recovery rate between the adversarial only setup (c) and the two baselines (a), (b), we conclude that at least most of the adversarial documents do **not** contain the claim that is necessary to generate the correct answer and can hence be considered adversarial.

6 Discussion

Citing Parametric Memory. Our results are the first step toward understanding unfaithful behavior in RAG systems due to post-rationalization. We focus on attributions, where a *faithful* attribution should signify the origin of the corresponding information. In contrast to past work, which values high citation recall, we argue that statements that were not generated from context but rather from model memory should not be accompanied by a citation. If the parametric model memory is used to generate an answer, a faithful model should either omit the citation or acknowledge their use of parametric memory rather than attempting to provide potentially misleading citations. This could for example be done by adding "model memory" as an explicit source, increasing transparency about the true origin of information.

The Importance of Faithfulness Evaluation. Our work underscores the importance of establishing control settings that yield conclusive evidence regarding faithfulness in model-generated content. Several issues necessitate a principled approach to measuring faithfulness in future research. The challenges we encountered are reminiscent of those seen in explainability research within IR and other fields, where ensuring validity in attribution metrics remains difficult [6, 9, 45]. A lack of ground truth and the inherently interpretative nature of attribution for RAG systems present a challenge for constructing evaluation criteria for accurately identifying unfaithful outputs. We suggest using evaluation strategies from explainability in IR, such as deliberate data contamination techniques [30, 63], model probing to gain first insights into specific model capabilities [21, 62, 69, 70], or reverse engineering parts of decision process [12]. However, validating LLM-based attributions introduces new challenges that call for the development of novel evaluation paradigms. We have proposed a preliminary test designed to assess faithfulness. This test, however, implicitly assumes that the model internals, or in other words, where the model looks and based on what it generates the answer, do not change through the insertion of additional irrelevant documents. To verify this assumption, we need to investigate the model's internal states during answer generation, e.g., based on recent advances in understanding internal model processes [8, 19, 26].

7 Conclusion

We have demonstrated that citation faithfulness is a crucial yet often overlooked aspect of reliable information retrieval systems. We have defined desiderata of faithful attribution and defined and disentangled the notions of citation correctness and citation faithfulness. We provide empirical evidence of unfaithful citation behavior through post-rationalization in Command-R+, a state-of-the-art LLM trained for the RAG task, by measuring the impact of short text insertions into irrelevant documents on the generated citations. Up to 57% of such insertions result in post-rationalization, highlighting a significant gap between correctness measured through token matching and true faithfulness in citation behavior. This highlights the importance of evaluating faithfulness, along with correctness, especially in high-stakes decision-making and decision-support.

Our study has several limitations that warrant consideration. First, the relatively small scale of our empirical analysis may limit the generalizability of our findings. Second, our research builds on the assumption that citations in AI-generated responses enhance user trust. While there is initial evidence that misleading explanations can increase user trust [59], the impact of misleading citations on user trust still requires further empirical validation. Third, a user study would be necessary to empirically measure the actual impact of misleading citations on information consumption and trust in reliable sources, particularly among vulnerable populations. Lastly, the conducted experiments raise several questions beyond the scope of this work, such as the instances where the adversarial document alone suffices to generate the model's answer, as observed in our experiments (Section 5).

These limitations point to several directions for future research. First, larger-scale studies are needed to validate our findings on post-rationalization and unfaithful attribution across a more diverse range of language models and datasets. Second, systematic human studies should investigate how different user groups, particularly vulnerable populations, interpret and interact with AI-generated citations. Third, researchers should develop robust evaluation frameworks for algorithmic accountability that specifically address attribution faithfulness in RAG systems. Finally, there is a need to explore alternative citation mechanisms that clearly distinguish between information drawn from model memory versus document-sourced statements.

Data and code. To facilitate reproducibility, code and parameters are available at https://github.com/jwallat/RAG-attributions.

Acknowledgments

This research was supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.-LTP.20.006, the European Union's Horizon Europe program under grant agreement No 101070212, the German Research Foundation (DFG), under Project IREM with grant No. AN 996/1-1, and by the Lower Saxony Ministry of Science and Culture (MWK), in the zukunft.niedersachsen program of the Volkswagen Foundation (HybrInt). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Transactions of the Association for Computational Linguistics* 12 (2024), 775–793.

[2] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (Un) reliability of Explanations from Large Language Models. *arXiv preprint arXiv:2402.04614* (2024).

[3] Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. *arXiv preprint arXiv:2311.01463* (2023).

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

[5] Francois Barnard, Marlize Van Sittert, and Sirisha Rambhatla. 2023. Self-diagnosis and Large Language Models: A New Front for Medical Misinformation. *arXiv preprint arXiv:2307.04910* (2023).

[6] Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 976–991.

[7] Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. *arXiv preprint arXiv:2212.08037* (2022).

[8] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations*.

[9] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods. *arXiv preprint arXiv:1910.02065* (2019).

[10] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems* 31 (2018).

[11] Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. PURR: Efficiently Editing Language Model Hallucinations by Denoising Language Model Corruptions. *arXiv preprint arXiv:2305.14908* (2023).

[12] Catherine Chen, Jack Merullo, and Carsten Eickhoff. 2024. Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*. ACM, 1401–1410.

[13] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner Fabien Roger Vlad Mikulik, Sam Bowman, Jan Leike Jared Kaplan, et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410* (2025).

[14] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2024. Do Models Explain Themselves? Counterfactual Simulatability of Natural Language Explanations. In *Forty-first International Conference on Machine Learning*.

[15] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing Health-related Misinformation on Social Media. *Jama* 320, 23 (2018), 2417–2418.

[16] James Chua and Owain Evans. 2025. Are DeepSeek R1 And Other Reasoning Models More Faithful?. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

[17] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 719–729.

[18] Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An Evaluation Framework for Attributed Information Retrieval using Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5354–5359.

[19] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* 1, 1 (2021), 12.

[20] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 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*. 150–158.

[21] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2022. Match Your Words! A Study of Lexical Matching in Neural Information Retrieval. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13186)*. Springer, 120–127.

[22] Jianfeng Gao, Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Heung-Yeung Shum. 2020. Robust Conversational AI with Grounded Text Generation. *arXiv preprint arXiv:2009.03457* (2020).

[23] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

[24] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6465–6488.

[25] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2023).

[26] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5484–5495.

[27] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and Other Large Generative AI Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1112–1123.

[28] Claudia E Haupt and Mason Marks. 2023. AI-generated Medical Advice—GPT and Beyond. *Jama* 329, 16 (2023), 1349–1350.

[29] High-Level Expert Group on AI. 2019. *Ethics Guidelines for Trustworthy AI*. Technical Report. EU. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[30] Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju, and Avishek Anand. 2021. Towards Benchmarking the Utility of Explanations for Model Debugging. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*. 68–73.

[31] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4198–4205.

[32] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 624–635.

[33] Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution. *arXiv preprint arXiv:2307.16883* (2023).

[34] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and individual differences* 103 (2023), 102274.

[35] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[36] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702* (2023).

[37] Minsuk Lee, James Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. 2006. Beyond Information Retrieval—Medical Question Answering. In *AMIA annual symposium proceedings*, Vol. 2006. American Medical Informatics Association, 469.

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. 9459–9474.

[39] Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A Survey of Large Language Models Attribution. *arXiv preprint arXiv:2311.03731* (2023).

[40] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The Dawn after the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. *arXiv preprint arXiv:2401.03205* (2024).

[41] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 681–691.

[42] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7001–7025.

[43] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374* (2023).

[44] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics* (2024), 1–67.

[45] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 305–329.

[46] K Tamsin Maxwell and Burkhard Schafer. 2008. Concept and Context in Legal Information Retrieval. In *Legal Knowledge and Information Systems*. IOS Press, 63–72.

[47] James Mayfield, Eugene Yang, Dawn J. Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Selin Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the Evaluation of Machine-Generated Reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*. ACM, 1904–1915.

[48] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching Language Models to Support Answers with Verified Quotes. *arXiv preprint arXiv:2203.11147* (2022).

[49] Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why Do People Believe Health Misinformation and Who is at Risk? A Systematic Review of Individual Differences in Susceptibility to Health Misinformation. *Social Science & Medicine* 314 (2022), 115398.

[50] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *Comput. Surveys* 55, 13s (2023), 1–42.

[51] Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 10056–10070.

[52] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics, 2523–2544.

[53] Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation. *arXiv preprint arXiv:2406.13663* (2024).

[54] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics* 49, 4 (2023), 777–840.

[55] Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology* 57, 9 (2023), 3464–3466.

[56] Jennifer Rowley and Frances Johnson. 2013. Understanding Trust Formation in Digital Information Sources: The Case of Wikipedia. *Journal of Information Science* 39, 4 (2013), 494–508.

[57] Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna Bala. 2024. Evaluation of RAG Metrics for Question Answering in the Telecom Domain. In *ICML 2024 Workshop on Foundation Models in the Wild*.

[58] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 338–354.

[59] Mersedeh Sadeghi, Daniel Pöttgen, Patrick Ebel, and Andreas Vogelsang. 2024. Explaining the Unexplainable: The Impact of Misleading Explanations on Trust in Unreliable Predictions for Hardly Assessable Tasks. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 36–46.

[60] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics, 3715–3734.

[61] Tina Seabrooke, Eike Schneiders, Liz Dowthwaite, Joshua Krook, Natalie Leesakul, Jeremie Clos, Horia Maior, and Joel Fischer. 2024. A Survey of Lay People's Willingness to Generate Legal Advice Using Large Language Models (LLMs). In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*. 1–5.

[62] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J. F. Jones. 2020. The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 2069–2072.

[63] Jaspreet Singh and Avishek Anand. 2020. Model Agnostic Interpretability of Rankers via Intent Modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 618–628.

[64] Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute First, then Generate: Locally-attributable Grounded Text Generation. *arXiv preprint arXiv:2403.17104* (2024).

[65] Brian G Southwell, Jeff Niederdeppe, Joseph N Cappella, Anna Gaysynsky, Dannielle E Kelley, April Oh, Emily B Peterson, and Wen-Ying Sylvia Chou. 2019. Misinformation as a Misunderstood Challenge to Public Health. *American Journal of Preventive Medicine* 57, 2 (2019), 282–285.

[66] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language Models for Dialog Applications. *arXiv preprint arXiv:2201.08239* (2022).

[67] S.M. Towhidul Islam Tonmoy, S.M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv preprint arXiv:2401.01313* (2024).

[68] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems* 36 (2024).

[69] Jonas Wallat, Fabian Beringer, Abhijit Anand, and Avishek Anand. 2023. Probing BERT for Ranking Abilities. In *European Conference on Information Retrieval*. Springer Nature Switzerland Cham, 255–273.

[70] Jonas Wallat, Hauke Hinrichs, and Avishek Anand. 2024. Causal Probing for Dual Encoders. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2292–2303.

[71] Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving Knowledge Conflicts in Large Language Models. *arXiv preprint arXiv:2310.00935* (2023).

[72] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing systems*, Vol. 35. 24824–24837.

[73] Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. "According to...": Prompting Language Models Improves Quoting from Pre-Training Data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2288–2301.

[74] Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. 2024. A New Era in LLM Security: Exploring Security Concerns in Real-world LLM-based Systems. *arXiv preprint arXiv:2402.18649* (2024).

[75] Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

[76] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge Conflicts for LLMs: A Survey. *arXiv preprint arXiv:2403.08319* (2024).

[77] Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large Language Models in Health Care: Development, Applications, and Challenges. *Health Care Science* 2, 4 (2023), 255–263.

[78] Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2024. Effective Large Language Model Adaptation for Improved Grounding and Citation Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*. Association for Computational Linguistics, 6237–6251.

[79] Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. 2024. Towards Fine-Grained Citation Evaluation in Generated Text: A Comparative Analysis of Faithfulness Metrics. In *Proceedings of the 17th International Natural Language Generation Conference*. 427–439.