Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Ensuring the verifiability of model answers is a fundamental challenge for retrieval-augmented generation (RAG) in the question answering (QA) domain. Recently, self-citation prompting was proposed to make large language models (LLMs) generate citations to supporting documents along with their answers. However, self-citing LLMs often struggle to match the required format, refer to non-existent sources, and fail to faithfully reflect LLMs' context usage throughout the generation. In this work, we present MIRAGE – Model Internals-based RAG Explanations – a plug-and-play approach using model internals for faithful answer attribution in RAG applications. MIRAGE detects context-sensitive answer tokens and pairs them with retrieved documents contributing to their prediction via saliency methods. We evaluate our proposed approach on a multilingual extractive QA dataset, finding high agreement with human answer attribution. On open-ended QA, MIRAGE achieves citation quality and efficiency comparable to self-citation while also allowing for a finer-grained control of attribution parameters. Our qualitative evaluation highlights the faithfulness of MIRAGE's attributions and underscores the promising application of model internals for RAG answer attribution.[1]

1 Introduction

Retrieval-augmented generation (RAG) with large language models (LLMs) has become the de-facto standard methodology for Question Answering (QA) in both academic (Lewis et al., 2020b; Izacard et al., 2022) and industrial settings (Dao and Le, 2023; Ma et al., 2024). This approach was shown to be effective at mitigating hallucinations and producing factually accurate answers (Petroni et al., 2020; Lewis et al., 2020a; Borgeaud et al., 2022; Ren et al., 2023). However, verifying whether the model answer is faithfully supported by the retrieved sources is often non-trivial due to the large context size and the variety of potentially correct answers (Krishna et al., 2021; Xu et al., 2023). In light of this issue, several *answer attribution*[2] approaches were recently proposed to ensure the trustworthiness of RAG outputs (Rashkin et al., 2021; Bohnet et al., 2022; Muller et al., 2023). Initial ef-

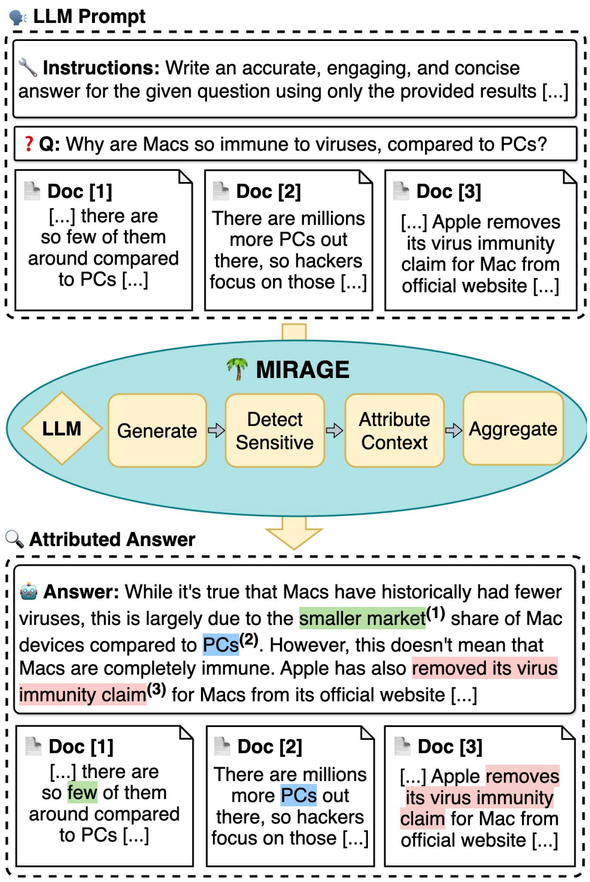

Figure 1: MIRAGE is a model internals-based answer attribution framework for RAG settings. Context-sensitive answer spans (in color) are detected and matched with contextual cues in retrieved sources to evaluate the trustworthiness of models' answers.

[1]Code and data released at `https://anonymized`.

[2]We use the term *answer attribution* (AA) when referring to the task of citing relevant sources to distinguish it from the *feature attribution* methods used in MIRAGE.

forts in this area employed models trained on Natural Language Inference (NLI) to automate the identification of supporting documents (Bohnet et al., 2022; Yue et al., 2023). Being based on an external validator, this approach does not faithfully explain the answer generation process but simply identifies plausible supporting sources post-hoc. Following recent progress in the instruction-following abilities of LLMs, *self-citation* (i.e. prompting LLMs to generate inline citations alongside their answers) has been proposed to mitigate the training and inference costs of external validator modules (Gao et al., 2023a). However, self-citation is hindered by the imperfect instruction-following capacity of modern LLMs (Mu et al., 2023; Liu et al., 2023). Moreover, the black-box nature of these models can make it difficult to evaluate self-citation faithfulness. We argue that this is a pivotal issue since the primary goal of answer attribution should be to ensure that the LLM is not 'right for the wrong reasons' (McCoy et al., 2019).

In light of this, we introduce MIRAGE, an extension of the context-reliance evaluation PECoRe framework (Sarti et al., 2024) that uses model internals for efficient and faithful answer attributions. This approach first identifies context-sensitive tokens in a generated sentence by measuring the shift in LM predictive distribution caused by the added input context. Then, it attributes this shift to specific influential tokens in the context using gradient-based saliency or other feature attribution techniques (Madsen et al., 2022). We adapt this approach to the RAG setup by matching context-dependent generated sentences to retrieved documents that contribute to their prediction and converting the resulting pairs to citations using the standard answer attribution (AA) format. We begin our assessment of MIRAGE on the short-form XOR-AttriQA dataset (Muller et al., 2023), showing high agreement between MIRAGE results and human annotations across several languages. We then test our method on the open-ended ELI5 dataset (Fan et al., 2019), achieving AA quality comparable to or better than self-citation, while ensuring a higher degree of control over attribution parameters. In summary, we make the following contributions:

- We introduce MIRAGE, a model internals-based answer attribution framework optimized for RAG applications.
- We show that MIRAGE outperforms NLI and self-citation methods while being more efficient and controllable.
- We analyze challenging attribution settings, highlighting MIRAGE's faithfulness to LLMs' reasoning process.

2 Background and Related Work

In RAG settings, a set of documents relevant to a user query is retrieved from an external dataset and infilled into an LLM prompt to improve the generation process (Petroni et al., 2020; Lewis et al., 2020a). *Answer attribution* (Rashkin et al., 2021; Bohnet et al., 2022; Muller et al., 2023) aims to identify which retrieved documents support the generated answer (*answer faithfulness*, Gao et al., 2023b), e.g., by exploiting the similarity between model outputs and references.[3] Simplifying access to relevant sources via answer attribution is a fundamental step towards ensuring RAG trustworthiness in customer-facing scenarios (Liu et al., 2023).

2.1 Answer Attribution Methods

Entailment-based Answer Attribution Bohnet et al. (2022) and Muller et al. (2023) approximate human annotation by leveraging the prediction of a pre-trained NLI system given a retrieved document as premise and a generated sentence as hypothesis. AAs produced by NLI systems such as TRUE (Honovich et al., 2022) were shown to correlate strongly with human annotations, prompting their adoption in AA studies (Muller et al., 2023; Gao et al., 2023a). Despite their effectiveness, entailment-based methods can be computationally expensive when several answer sentence-document pairs are present. Moreover, this approach assumes that the NLI model can robustly detect entailment between answers and supporting documents across several domains and languages. In practice, however, NLI systems were shown to be brittle in challenging scenarios, exploiting shallow heuristics (McCoy et al., 2019; Nie et al., 2020; Sinha et al., 2021; Luo et al., 2022), and require dedicated efforts for less-resourced settings (Conneau et al., 2018). For example, NLI may fail to correctly attribute answers in multi-hop QA settings when considering individual documents as premises (Yang et al., 2018; Welbl et al., 2018).

[3] Popular frameworks such as `LangChain` (Chase, 2022) and `LlamaIndex` (Liu, 2022) support similarity-based citations using vector databases.

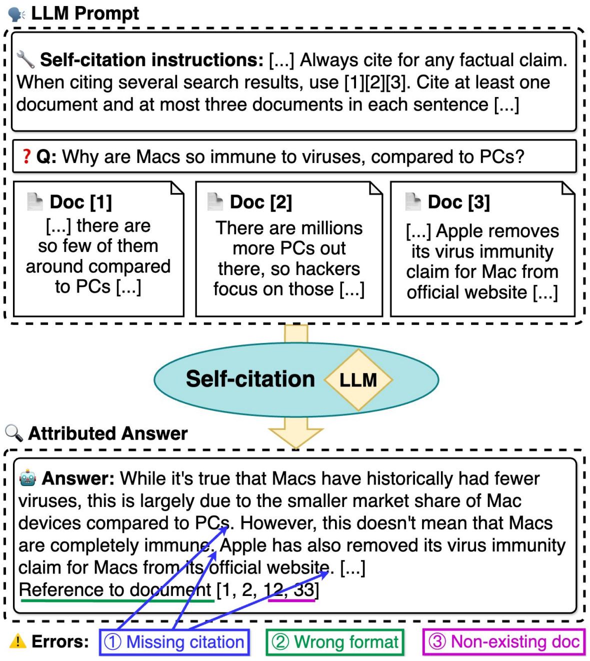

Figure 2: Instruction-following errors in a *self-citation* example, using the setup of Gao et al. (2023a).

Self-citation Gao et al. (2023a) is a recent AA approach exploiting the ability of recent LLMs to follow instructions in natural language (Raffel et al., 2020; Chung et al., 2022; Sanh et al., 2022; Brown et al., 2020), thereby avoiding the need for an external validator. Nakano et al. (2021) and Menick et al. (2022) propose citation fine-tuning for LLMs, while Gao et al. (2023a) instruct general-purpose LLMs to produce inline citations in a few-shot setting. Answers produced via self-citation prompting are generally found to be of higher quality and more related to information contained in provided sources, but can still contain unsupported statements and inaccurate citations (Liu et al., 2023). In our preliminary analysis, we find that self-citation often misses relevant citations, uses a wrong format, or refers to non-existing documents (Figure 2). Table 1 shows LLaMA 2 7B Chat (Touvron et al., 2023) and Zephyr β 7B (Tunstall et al., 2023) results on the ELI5 dataset (Fan et al., 2019) using Gao et al. (2023a) self-citation setup. Both tested models fail to produce AAs matching the prompt instructions for the majority of generated sentences, with almost all answers having at least one unattributed sentence.

2.2 Attribution Faithfulness

Answer Attribution can be Unfaithful The aforementioned approaches do not account for attributions' *faithfulness*, i.e. whether the selected documents influence the LLM during the generation. Indeed, the presence of an entailment relation or high semantic similarity does not imply that a retrieved document was functional in generating the selected answer. For example, an LLM may rely on memorized knowledge while ignoring the provided relevant context. Even in the case of self-citation, recent work showed that, while the justifications of self-explaining LLMs appear plausible, they generally do not align with their internal reasoning process (Atanasova et al., 2023; Madsen et al., 2024; Agarwal et al., 2024), with little to no predictive efficacy (Huang et al., 2023). Concurrent to our work, Phukan et al. (2024) propose an internals-based method for granular AA of LLM generations. While the two-step approach they proposed is similar to MIRAGE, their usage of embedding similarity as an attribution indicator has inherent faithfulness limitations since it does not capture the functional aspect of context usage during prediction.

Model	Missing citation (%) Answer	Missing citation (%) Sentence
Zephyr 7B β	54.5	95.7
LLaMA 2 7B Chat	62.4	99.3

Table 1: % of unattributed sentences and answers with ≥ 1 unattributed sentences on ELI5.

Feature Attribution in Interpretability The task of faithfully identifying salient context information has been studied extensively in the NLP interpretability field (Ferrando et al., 2024). In particular, *post-hoc feature attribution* approaches (Madsen et al., 2022) exploit information sourced from model internals, e.g., attention weights or gradients of next-word probabilities, to identify input tokens playing an important role towards the model's prediction. While feature attribution studies in NLP typically focused on classification tasks (Atanasova et al., 2020; Wallace et al., 2020; Chrysostomou and Aletras, 2022), recent work applies these methods to evaluate context usage in language generation (Yin and Neubig, 2022; Ferrando et al., 2023; Sarti et al., 2023, 2024). Importantly, feature attribution techniques are designed to maximize the faithfulness of selected context tokens by accessing models' intermediate computations, as opposed to the AA methods of Section 2.1. While the faithfulness of such approaches can still vary depending on models and tasks, the development of robust and faithful methods is an active area of research (Jacovi and Goldberg, 2020; Chan et al., 2022; Bastings et al., 2022; Lyu et al., 2024).

3 Method

Identifying which generated spans were most influenced by preceding information is a key challenge for LM attribution. The Model Internals-based RAG Explanations (MIRAGE) method we propose is an extension of the Plausibility Evaluation for Context Reliance (PECoRE) framework (Sarti et al., 2024) for context-aware machine translation. This section provides an overview of PECoRE's two-step procedure and clarifies how MIRAGE adapts it for RAG answer attribution.

3.1 Step 1: Context-sensitive Token Identification (CTI)

For every token in an answer sentence $\mathbf{y} = \langle y_1, \dots, y_n \rangle$ generated by a LM prompted with a query $\mathbf{q}$ and a context $\mathbf{c} = \langle c_1, \dots, c_{|\mathbf{c}|} \rangle$, a contrastive metric m such as KL divergence (Kullback and Leibler, 1951) is used to quantify the shift in the LM predictive distribution at the i-th generation step when the context is present or absent (P^i_{ctx} or $P^i_{\text{no-ctx}}$). Resulting scores $\mathbf{m} = \langle m_1, \dots, m_n \rangle$ reflect the context sensitivity of every generated token and can be converted into binary labels using a selector function s_{CTI}:

$$\begin{aligned} \text{CTI}(\mathbf{q}, \mathbf{c}, \mathbf{y}) = \{\, y_i \,|\, s_{\text{CTI}}(m_i) = 1\, \forall y_i \in \mathbf{y} \} \\ \text{with } m_i = \text{KL}(P^i_{\text{ctx}} \,\|\, P^i_{\text{no-ctx}}) \end{aligned} \tag{1}$$

3.2 Step 2: Contextual Cues Imputation (CCI)

For every context-sensitive token y_i identified by CTI, a contrastive alternative $y_i^{\setminus \mathbf{c}}$ is produced by excluding $\mathbf{c}$ from the prompt, but using the original generated prefix $\mathbf{y}_{<i}$. Then, *contrastive feature attribution* (Yin and Neubig, 2022) is used to obtain attribution scores $\mathbf{a}^i = \langle a^i_1, \dots, a^i_{|\mathbf{c}|} \rangle$ for every context token $c_j \in \mathbf{c}$:

$$a^i_j = \left\{\, \nabla_j \big(p(y_i) - p(y_i^*) \big),\ \forall c_j \in \mathbf{c} \,\right\} \tag{2}$$

where ∇_j is the L2 norm of the gradient vector over the input embedding of context token c_j, and both probabilities are computed from the same contextual inputs $(\mathbf{q}, \mathbf{c}, \mathbf{y}_{<i})$. Intuitively, this procedure identifies which tokens in $\mathbf{c}$ influence the prediction of y_i while accounting for the non-contextual option $y_i^{\setminus \mathbf{c}}$. Resulting scores are once again binarized with a selector s_{CCI}:

$$\text{CCI}(y_i) = \{\, c_j \,|\, s_{\text{CCI}}(a^i_j) = 1,\ \forall c_j \in \mathbf{c} \} \tag{3}$$

This results in pairs of context-sensitive generated tokens and the respective input-context tokens influencing their prediction:

$$\mathcal{P} = \big\{ \langle\, y_i, c_j \,\rangle,\ \forall y_i \in \text{CTI}, \forall c_j \in \text{CCI}(y_i) \big\} \tag{4}$$

3.3 From Granular Attributions to Document-level Citations

CTI Filtering First, we set $s_{\text{CTI}}(m_i) = m_i \geq m^*$, where m^* is a threshold value for selecting context-sensitive generated tokens. We experiment with two variants of m^*: a **calibrated threshold** m^*_{CAL} obtained by maximizing agreement between the contrastive metric and human annotations on a calibration set with human AA annotations, and an **example-level threshold** m^*_{EX} using only within-example scores to avoid the need of calibration data. In our experiments, we follow the approach by Sarti et al. (2024) and set $m^*_{\text{EX}} = \overline{\mathbf{m}} + \sigma_{\mathbf{m}}$, where $\overline{\mathbf{m}}$ and $\sigma_{\mathbf{m}}$ are respectively the average and standard deviation of $\mathbf{m}$ scores for the given example.

CCI Filtering To extract granular document citations (e.g., colored spans with document indices in Figure 1), we set $s_{\text{CCI}} = a^i_j \geq a^{i*}$, where a^{i*} is either the Top-K or Top-% highest attribution value in $\mathbf{a}^i$, to filter attributed context tokens $c_j \in \text{CCI}(y_i)$. Then, we use the identifier $\text{docid}(c_j)$ of the documents they belong to as citation for context-sensitive token y_i. Since token-level citations may be hard to interpret, we collate consecutive tokens citing the same documents into a single span and map highlights from subword to word-level for visualization purposes.

Sentence-level Aggregation AA is commonly performed at the sentence level to follow standard citation practices and facilitate user assessment. To enable a direct comparison with other sentence-level methods, we aggregate token-level citations as the union over all cited documents $\text{docid}(\cdot)$ across context-sensitive tokens in $\mathbf{y}$:

$$\begin{aligned} \text{MIRAGE}(\mathbf{y}) = \bigcup_{y_i \in \text{CTI}(\mathbf{y})} \text{docid}(c_j)\ \forall c_j \in \text{CCI}(y_i) \\ \text{with } s_{\text{CTI}} = m_i \geq m^*, s_{\text{CCI}} = a^i_j \geq a^{i*} \end{aligned} \tag{5}$$

In the following sections, we use MIRAGE $_{\text{CAL}}$ and MIRAGE $_{\text{EX}}$ to refer to sentence-level answer attribution using m^*_{CAL} and m^*_{EX} thresholds, respectively.

4 Agreement with Human Answer Attribution Annotations

We begin our evaluation by comparing MIRAGE predictions with human-produced answer attributions. We employ the XOR-AttriQA dataset (Muller et al., 2023), which, to our knowledge, is the only open dataset with human annotations over RAG outputs produced by a publicly accessible LM.[4] We limit our assessment to open-weights LLMs to ensure that MIRAGE answer attribution can faithfully reflect the model's inner processing towards the natural production of the annotated answer used for evaluation. Moreover, while cross-linguality is not the focus of our work, XOR-AttriQA allows us to assess the robustness of MIRAGE across several languages and its agreement with human annotations compared to an entailment-based system.

4.1 Experimental Setup

XOR-AttriQA consists of 500/4720 validation/test tuples, each containing a concise factual query $\mathbf{q}$, a set of retrieved documents that we use as context $\mathbf{c} = \langle \text{doc}_1, \dots, \text{doc}_k \rangle$, and a single-sentence answer $\mathbf{y}$ produced by an mT5-base model (Xue et al., 2021) fine-tuned on cross-lingual QA in a RAG setup (CORA; Asai et al., 2021).[5] Queries and documents span five languages (Bengali, Finnish, Japanese, Russian, and Telugu), with no constraint on documents to match the language of the query.[6] Although the RAG generator employs a set of retrieved documents during generation, human annotators were asked to label tuples $(\mathbf{q}, \text{doc}_i, \mathbf{y})$ to indicate whether the information in doc_i supports the generation of $\mathbf{y}$. Importantly, MIRAGE requires extracting model internals in the naturalistic setting that leads to the generation of the desired answer, i.e., the one assessed by human annotators. Hence, we perform a selection procedure to identify XOR-AttriQA examples where the answer produced by filling in the concatenated documents $\mathbf{c}$ in the LM prompt matches the one provided. The resulting subset, which we dub XOR-AttriQA$_{\text{match}}$, contains 142/1144 calibration/test examples and is used for our evaluation.[7]

[4] E.g., the human-annotated answers in Bohnet et al. (2022) were generated by the proprietary PALM 540B (Chowdhery et al., 2023), whose internals are inaccessible.

[5] https://hf.co/gsarti/cora_mgen

[6] In practice, Muller et al., 2023 report that most retrieved documents are in the same language as the query or in English.

[7] See Appendix A for more details on this selection. Appendix B presents experiments on the full XOR-AttriQA.

4.2 Entailment-based Baselines

Muller et al. (2023) use an mT5 XXL model fine-tuned on NLI for performing answer attribution on XOR-AttriQA. Since neither the tuned model nor the tuning data are released, we opt to use TRUE[8] (Honovich et al., 2022), a fine-tuned T5 11B model (Raffel et al., 2020), which was shown to highly overlap with human annotation on English answer attribution tasks (Muller et al., 2023; Gao et al., 2023a). We evaluate TRUE agreement with human annotation in two setups. In NLI$_{\text{ORIG}}$, we evaluate the model directly on all examples, including non-English data. While this leads the English-centric TRUE model out-of-distribution, it accounts for real-world scenarios with noisy data, and can be used to assess the robustness of the method in less-resourced settings. Instead, in NLI$_{\text{MT}}$, all queries and documents are machine translated to English using the Google Translate API.[9] While this simplifies the task by ensuring all TRUE inputs are in English, it can lead to information loss caused by imprecise translation.

4.3 Results and Analysis

MIRAGE agrees with human answer attribution Table 2 presents our results. MIRAGE is found to largely agree with human annotations on XOR-AttriQA$_{\text{match}}$, with scores on par or slightly better than those of the ad-hoc NLI$_{\text{MT}}$ system augmented with automatic translation. Although calibration appears to generally improve MIRAGE's agreement with human annotators, we note that the uncalibrated MIRAGE$_{\text{EX}}$ achieves strong performances despite having no access to external modules or tuning data. These findings confirm that the inner workings of LMs can be used to perform answer attribution, resulting in performances on par with supervised answer attribution approaches even in the absence of annotations for calibration.

MIRAGE is robust across languages and filtering procedures Table 2 shows that NLI$_{\text{ORIG}}$ answer attribution performances are largely language-dependent due to the unbalanced multilingual abilities of the TRUE NLI model. This highlights the brittleness of entailment-based approaches in OOD settings, as discussed in Section 2.1. Instead, MIRAGE variants perform similarly across all languages by exploiting the internals of the multilingual RAG model. MIRAGE's performance across

[8] https://hf.co/google/t5_xxl_true_nli_mixture

[9] https://cloud.google.com/translate

Method	Extra Requirements	CCI Filter	BN	FI	JA	RU	TE	Avg. / Std
NLI $_{\text{ORIG}}$ (Honovich et al.) NLI $_{\text{MT}}$ (Honovich et al.)	11B NLI model 11B NLI model + MT engine	–	33.8 82.6	83.7 83.7	86.5 90.5	**85.8** 81.7	50.0 82.5	68.0 / 21.9 84.2 / 3.2
MIRAGE $_{\text{CAL}}$ (Ours)	142 annotated AA examples	Top 3 Top 5%	81.7 **84.4**	**84.2** 83.0	87.8 **91.4**	83.3 **85.8**	87.0 **88.9**	84.8 / 2.3 **86.7** / 3.1
MIRAGE $_{\text{EX}}$ (Ours)	–	Top 3 Top 5%	80.2 81.7	78.5 80.1	83.8 89.2	77.2 84.4	75.2 81.8	79.0 / 2.9 83.4 / 3.2

Table 2: Agreement % of MIRAGE and entailment-based baselines with human AA on XOR-AttriQA$_{\text{match}}$ using CORA for RAG. **Extra Requirements**: data/models needed for AA in addition to the RAG model and the current example. **Filter**: s_{CCI} filtering for saliency scores. **Best overall** and best uncalibrated scores are highlighted.

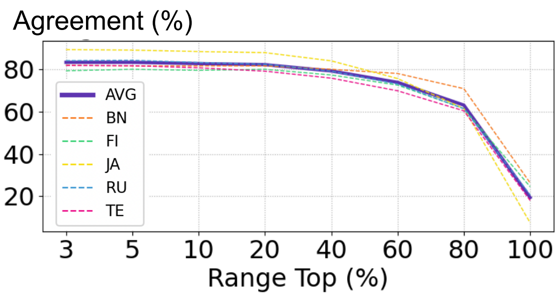

Figure 3: Robustness of MIRAGE $_{\text{EX}}$ agreement with human annotations across Top-% CCI filtering thresholds.

languages is comparable to that of NLI $_{\text{MT}}$, which requires an extra translation step to operate on English inputs.

We further validate the robustness of the CCI filtering process by testing percentile values between Top 3-100% for the MIRAGE $_{\text{EX}}$ setting. Figure 3 shows that Top % values between 3 and 20% lead to a comparably high agreement with human annotation, suggesting this filtering threshold can be selected without ad-hoc parameter tuning.

5 Answer Attribution for Long-form QA

XOR-AttriQA can only provide limited insights for real-world answer attribution evaluation since its examples are sourced from Wikipedia articles, and its answers are very concise. In this section, we extend our evaluation to ELI5 (Fan et al., 2019), a challenging long-form QA dataset that was recently employed to evaluate LLM self-citation capabilities (Gao et al., 2023a). Different from XOR-AttriQA, ELI5 answers are expected to contain multiple sentences of variable length, making it especially fitting to assess MIRAGE context-sensitive token identification capabilities before document attribution. Alongside our quantitative assessment of MIRAGE in relation to self-citation baselines, we conduct a qualitative evaluation of the disagreement between the two methods.

5.1 Experimental Setup

Dataset The ELI5 dataset contains open-ended why/how/what queries $\mathbf{q}$ from the "Explain Like I'm Five" subreddit[10] eliciting long-form multi-sentence answers. For our evaluation, we use the RAG-adapted ELI5 version by Gao et al. (2023a), containing top-5 matching documents $\mathbf{c} = \langle \text{doc}_1, \ldots, \text{doc}_5 \rangle$ retrieved from a filtered version of the Common Crawl (Sphere; Piktus et al., 2021) for every query. The answer attribution task is performed by generating a multi-sentence answer $\mathbf{ans} = \langle \mathbf{y}_1, \ldots, \mathbf{y}_m \rangle$ with an LLM using $(\mathbf{q}, \mathbf{c})$ as inputs, and identifying documents in $\mathbf{c}$ supporting the generation of answer sentence $\mathbf{y}_i$, $\forall \mathbf{y}_i \in \mathbf{ans}$.

Models and Answer Attribution Procedure We select LLaMA 2 7B Chat (Touvron et al., 2023) and Zephyr β 7B (Tunstall et al., 2023) for our experiments since they are high-quality open-source LLMs of manageable size. To enable a fair comparison between the tested attribution methods, we first generate answers with inline citations using the self-citation prompt by Gao et al. (2023b).[11] Then, we remove citation tags and use MIRAGE to attribute the resulting answers to retrieved documents. This process ensures that citation quality is compared over the same set of answers, controlling for the variability that could be produced by a different prompt.[12] For more robust results, we perform generation three times using different sampling seeds, and report the averaged scores. Since human-annotated data is not available, we only assess the calibration-free MIRAGE $_{\text{EX}}$.

Entailment-based Evaluation Differently from the XOR-AttriQA dataset used in Section 4, ELI5 does not contain human annotations of AA. For this reason, and to ensure consistency with Gao et al.

[10] https://reddit.com/r/explainlikeimfive

[11] The full prompt is provided in Appendix D (Table 9).

[12] For completeness, we also report MIRAGE results without self-citation prompting in Appendix D.

Model	Answer Attrib.	Citation ↑ Prec.	Rec.	F1
Zephyr β	Self-citation	41.4	24.3	30.6
	MIRAGE EX Top 3	38.3	46.2	41.9
	MIRAGE EX Top 5%	**44.7**	**46.5**	**45.6**
LLaMA 2	Self-citation	**37.9**	19.8	26.0
	MIRAGE EX Top 3	21.8	**29.6**	25.1
	MIRAGE EX Top 5%	26.2	29.1	**27.6**

Table 3: Answer attribution quality estimated by TRUE for self-citation and MIRAGE on ELI5.

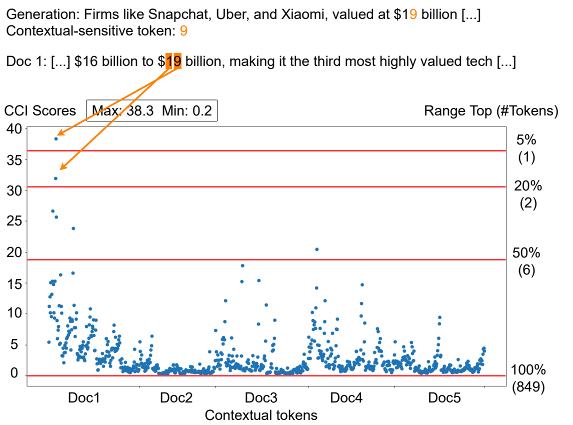

Figure 4: Attribution scores over retrieved documents' tokens for the prediction of context-sensitive token '9'.

(2023a)'s self-citation assessment, we adopt the TRUE model as a high-quality approximation of expected annotation behavior. Despite the potential OOD issues of entailment-based AA highlighted in Section 4, we expect TRUE to perform well on ELI5 since it closely matches the general/scientific knowledge queries in TRUE's fine-tuning corpora and contains only English sentences. To overcome the multi-hop issue when using single documents for entailment-based answer attribution, we follow the ALCE evaluation (Gao et al., 2023a)[13] to measure citation quality as NLI precision and recall (summarized by F1 scores) over the concatenation of retrieved documents.

5.2 Results

Results in Table 3 show that MIRAGE provides a significant boost in answer attribution precision and recall for the Zephyr β model, while it greatly improves citation recall at the expense of precision for LLaMA 2, resulting in an overall higher F1 score for the MIRAGE EX Top 5% setting. These results confirm that MIRAGE can produce effective answer attributions in longer and more complex settings while employing no external resources like the self-citation approach.

[13]ALCE is an evaluation framework for RAG, evaluating LLM responses in terms of citation quality, correctness, and fluency. More details can be found in Appendix C

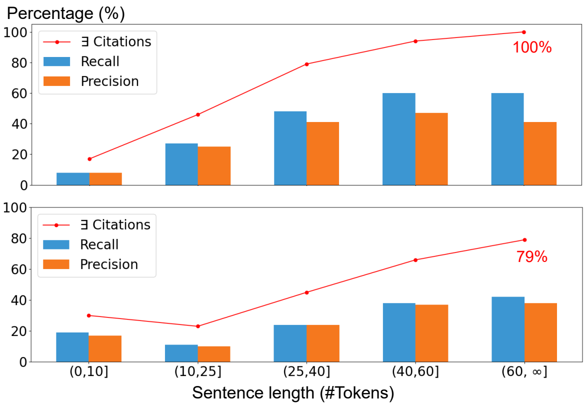

Figure 5: MIRAGE EX (top) and self-citation (bottom) average performance on ELI5 answer sentences binned by length. red: % of sentences with ≥ 1 citation.

From the comparison between Top 3 and Top 5% CCI filtering strategies, we note that the latter generally results in better performance. This intuitively supports the idea that an adaptive selection strategy is more fitting to accommodate the large variability of attribution scores across different examples. Figure 4 visualizes the distributions of attribution scores a^i_j for an answer produced by Zephyr β, showing that most context tokens in retrieved documents receive low attribution scores, with only a handful of them contributing to the prediction of the context-sensitive token '9' in the generation. This example also provides an intuitive explanation of the robustness of Top-% selection thresholds discussed in Section 4.3. Ultimately, the Top 5% threshold is sufficient to select the document containing the direct mention of the generated token.

Since the m^*_{EX} threshold used to select context-sensitive tokens by MIRAGE EX depends on the mean and standard deviation of generated answer's scores, we expect that the length of the generated answer might play a role in citation quality. As shown in Figure 5, MIRAGE citation quality is indeed lower for shorter answer sentences. However, a similar trend is observed for self-citation, which is outperformed by MIRAGE for all but the shortest length bin ($\leq$ 10 tokens). The proportion of non-attributed sentences (red line) suggests that the lower quality could be a byproduct of the ALCE evaluation protocol, where non-attributed sentences receive 0 precision/recall. Future availability of human-annotated RAG datasets may shed more light on this effect.

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

Document [1] [...] Q. What does it mean for books to be Sensitized or Desensitized? A security strip is embedded into each library book. When a book is checked out, it must be "desensitized" so that it will not set off the alarm when it leaves or enters the library. When the book is returned, it is “sensitized” so that the alarm will be set off should someone attempt to take the book from the library without going through the proper borrowing procedure.
Document [2] [...]

Query: How does a small paper library bar code set off the security alarm when you haven’t checked a book out?

ANSWER ATTRIBUTION RESULTS

Self-citation: [...] When a book is checked out, it is "desensitized" to prevent the alarm from going off. [∅] When the book is returned, it is "sensitized" so the alarm will sound if the item is taken without authorization. [∅] [...]

MIRAGE: [...] When a book is checked (1) out, it (1) is "desensitized" (1) to prevent the alarm from going off. [1] When the book (1) is returned, it is "sensitized" (1) so the alarm will sound if the item is taken without authorization. [1] [...]

NLI (TRUE model): [1] entails both answer sentences.

Table 4: Example of self-citation failure using Zephyr β on ELI5. NLI and MIRAGE produce the correct citation, while self-citation does not cite any document ([∅]).

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

Document [2] [...]
Document [3] [...] What will happen if you accidentally set off your security system? The siren will sound and it will be loud, but you should be able to stop the siren by entering your code into your keypad. [...]
Document [4] [...]

Query: How does a small paper library bar code set off the security alarm when you haven’t checked a book out?

ANSWER ATTRIBUTION RESULTS

Self-citation: [...] False alarms can be prevented by entering the code on the keypad, as documented in [3]. [...]

MIRAGE: [...] False alarms can be prevented by entering the code on the keypad [∅] [...]

NLI (TRUE model): [3] entails the answer sentence.

Table 5: Example showcasing the brittleness of entailment-based AA. MIRAGE correctly finds that the answer cannot be attributed ([∅]), while NLI and self-citation attribute the lexically similar Document [3].

5.3 Qualitative Analysis of Disagreements

To better understand MIRAGE’s performance, we examine some ELI5 examples where MIRAGE disagrees with self-citation on Zephyr β’s generations. Table 4 and 5 illustrate two cases in which the entailment-based TRUE model results agree with either MIRAGE or self-citation. In Table 4, the answer provided by the model is directly supported by Document [1], as also identified by TRUE. However, self-citation fails to cite the related document at the end of the two sentences. By contrast, MIRAGE attributes several spans to Document [1], resulting in the correct answer attribution for both sentences.

While TRUE achieves high consistency with human judgment (e.g., for the example in Table 4), NLI-based AA can still prove unreliable in cases of high lexical overlap between the answer and supporting documents. Table 5 illustrates one such case, where both self-citation and TRUE attribute the answer to Document [3], whereas MIRAGE does not label any context document as salient for the answer. Here, the answer wrongly states that the bar code can used to ***prevent*** the alarm, while Document [3] mentions that the code can be used to ***cancel*** the alarm after an accidental activation. Thus, despite the high lexical and semantic relatedness, the answer is not supported by Document [3]. The failure of TRUE in this setting highlights the sensitivity of entailment-based systems to surface-level similarity, making them brittle in cases where the model’s context usage is not straightforward. Using another sampling seed for the same query produces the answer *“[...] the individual can **cancel** the alarm by providing their password at the keypad”*, which MIRAGE correctly attributes to Document [3].[14]

6 Conclusion

In this study, we introduced MIRAGE, a novel approach to enhance the faithfulness of answer attribution in RAG systems. By leveraging model internals, MIRAGE effectively addresses the limitations of previous methods based on prompting or external NLI validators. Our experiments demonstrate that MIRAGE produces outputs that strongly agree with human annotations while being more efficient and controllable than its counterparts. Our qualitative analysis shows that MIRAGE can produce faithful attributions that reflect actual context usage during generation, reducing the risk of false positives motivated by surface-level similarity.

In conclusion, MIRAGE represents a promising first step in exploiting interpretability insights to develop faithful answer attribution methods, paving the way for the usage of LLM-powered question-answering systems in mission-critical applications.

[14] This and other examples are provided in Appendix E.

7 Limitations

LLMs Optimized for Self-citation In this study, we focus our analysis on models that are not explicitly trained to perform self-citation and can provide citations only when prompted to do so. While recent systems include self-citation in their optimization scheme for RAG applications[15], we believe incorporating model internals in the attribution process will remain a valuable and inexpensive method to ensure faithful answer attributions.

Brittleness of NLI-based evaluation Following Gao et al. (2023a), the evaluation of Section 5 employs the NLI-based system TRUE due to the lack of AA-annotated answers produced by open-source LLMs. However, using the predictions of NLI models as AA references is far from ideal in light of their brittleness in challenging scenarios and their tendency to exploit shallow heuristics. While the ELI5 dataset is reasonably in-domain for the TRUE model, this factor might still undermine the reliability of some of our quantitative evaluation results. Future work should produce a wider variety of annotated datasets for reproducible answer attribution using open-source LLMs, enabling us to extend our analysis to a broader set of languages and model sizes and ultimately enhance the robustness of our findings.

Applicability to Other Domains and Model Sizes Our evaluation is conducted on relatively homogeneous QA datasets and does not include language models with >7B parameters. This limits the generalizability of our findings to other domains and larger models. Future work should extend our analysis to a broader range of domains and model sizes to further validate the robustness and applicability of MIRAGE. This said, we expect MIRAGE to be less vulnerable to language and quality shifts compared to existing AA methods that depend on external validators or on the model's instruction-following abilities.

MIRAGE's Parametrization and Choice of Attribution Method While Section 4.1 highlights the robustness of MIRAGE to various CCI filtering thresholds, the method still requires non-trivial parametrization. In particular, we emphasize that the choice of the attribution method employed to generate attribution scores in the CCI step can significantly impact the faithfulness of the resulting answer attributions. Although we employed a relatively simple gradient-based approach in this study, we note that our proposed framework is method-agnostic and can incorporate more sophisticated feature attribution techniques. Finally, we remark that MIRAGE can produce redundant citations for repeated information across multiple documents, which might result in misleading answer attributions (see e.g. Appendix E).

[15]For example, the Command-R models: https://huggingface.co/CohereForAI/c4ai-command-r-plus

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34:7547–7560.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

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*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *ArXiv*, abs/2212.08037.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

Harrison Chase. 2022. LangChain.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

George Chrysostomou and Nikolaos Aletras. 2022. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Xuan-Quy Dao and Ngoc-Bich Le. 2023. Chatgpt is good but bing chat is better for vietnamese students. *arXiv preprint arXiv:2307.08272*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *Preprint*, arXiv:2405.00208.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen

Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

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*, pages 4198–4205, Online. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Solomon Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

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. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jerry Liu. 2022. LlamaIndex.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. Simple but challenging: Natural language inference models fail on simple sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3449–3462, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Lijia Ma, Xingchen Xu, and Yong Tan. 2024. Crafting knowledge: Exploring the creative mechanisms of chat-based search engines. *arXiv preprint arXiv:2402.19421*.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? *ArXiv*, abs/2401.07927.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8).

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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*.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Dan Hendrycks, and David Wagner. 2023. Can llms follow simple rules? *arXiv preprint arXiv:2311.04235*.

Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. Peering into the mind of language models: An approach for attribution in contextual question answering. *Preprint*, arXiv:2405.17980.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, et al. 2021. The web is your oyster-knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and D. Reitter. 2021. Measuring attribution in natural language generation models. *Computational Linguistics*, 49:777–840.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria. OpenReview.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

A Construction of XOR-AttriQA$_{\text{match}}$

XOR-AttriQA$_{\text{match}}$ is a subset of the original XOR-AttriQA containing only examples for which our LLM generation matches exactly the answer annotated in the dataset. Replicating the original answer generation process is challenging since the original ordering of the documents doc_i in $\mathbf{c}$ unavailable.[16] To maximize the chances of replication, we attempt to restore the original document sequence by randomly shuffling the order of doc_is until LLM can naturally predict the answer $\mathbf{y}$ (otherwise, at most 200 iterations), as shown in Algorithm 1. The statistics of the original XOR-AttriQA and XOR-AttriQA$_{\text{match}}$ are shown in Table 6.

Algorithm 1 Restore original document sequence

Input: $\{Doc_1, ..., Doc_n\}$, $query$, $answer$, $\mathbb{M}$

```
iter ← 0,  found = False
while iter < 200 do
    pred ← M({Doc_1, ..., Doc_n}, query)
    if pred == answer then
        found = True  break
    else
        Shuffle({Doc_1, ..., Doc_n})
    end if
    iter += 1
end while
if found then
    return {Doc_1, ..., Doc_n}
end if
```

Dataset	BN	FI	JA	RU	TE	Total
Orig.	1407	659	1066	954	634	4720
Match	274	214	232	254	170	1144

Table 6: Statistic for tests sets of the original (Orig.) XOR-AttriQA and XOR-AttriQA$_{\text{match}}$.

B Answer Attribution on the Full XOR-AttriQA

Differently from the concatenation setup in Section 4.1, we also test MIRAGE on the full XOR-AttriQA dataset by constraining CORA generation to match the annotated answer $\mathbf{y}$. We adopt a procedure similar to Muller et al. (2023) by considering a single document-answer pair $(\text{doc}_i, \mathbf{y})$ at a time, and using MIRAGE's CTI step to detect whether $\mathbf{y}$ is sensitive to the context doc_i. Results in Table 7 show that MIRAGE achieves performances in line with other AA methods despite these approaches employing ad-hoc validators trained with as many as 540B parameters.

C ALCE Evaluation Benchmark

Gao et al. (2023a) propose ALCE, an evaluation framework for RAG QA tasks. ALCE assesses the LLMs' response from three diverse aspects: citation quality, correctness, and fluency. **Cita-**

[16] Muller et al. 2023 only provide the split documents without the original ordering.

Method	Extra Requirements	BN	FI	JA	RU	TE	Avg. / Std
mT5 XXL$_{\text{NLI}}$ (Honovich et al.)	11B NLI model (250 FT ex.)	81.9	80.9	94.5	87.1	88.7	86.6 / 4.9
	11B NLI model (100k FT ex.)	89.4	88.3	91.5	91.0	92.4	90.5 / 1.5
	11B NLI model (1M FT ex.)	91.1	90.4	93.0	92.9	93.8	92.2 / 1.3
PALM2$_{\text{LORA}}$ (Anil et al.)	540B LLM (250 FT ex.)	91.5	88.3	94.7	93.7	93.7	92.4 / 2.3
PALM2 (Anil et al.)	540B LLM (250 FT ex.)	**92.3**	**92.6**	**96.4**	**94.5**	**94.8**	**94.1** / 1.5
PALM2 (Anil et al.)	540B LLM (4-shot prompting)	91.5	87.4	92.0	90.5	90.6	90.4 / 1.6
PALM2$_{\text{CoT}}$ (Anil et al.)	540B LLM (4-shot prompting)	83.7	78.8	71.7	81.9	84.7	80.2 / 4.7
MIRAGE$_{\text{CAL}}$ (Ours)	500 AA calibration ex.	82.2	82.5	92.0	87.7	90.2	86.9 / 4.0
MIRAGE$_{\text{EX}}$ (Ours)	–	79.0	74.1	90.8	82.6	86.9	82.7 / 5.8

Table 7: Agreement % of MIRAGE and entailment-based baselines with human AA on the full XOR-AttriQA using CORA for RAG (annotated answers not matching the LM's natural generation are force-decoded). **Extra Requirements**: data/models needed for AA in addition to the RAG model itself. **Best overall** and best validator-free scores are highlighted. PALM and mT5 results are taken from Muller et al. (2023).

Model	Prompt	Answer Attribution	Filter	Citation↑ Prec.	Rec.	F1	Correctness↑	Fluency↑
Zephyr	Self-citation	Self-citation	-	41.4	24.3	30.6	9.9	28.6
		MIRAGE$_{\text{EX}}$	Top 3	38.3	46.2	41.9		
			Top 5%	**44.7**	**46.5**	**45.6**		
	Standard	MIRAGE$_{\text{EX}}$	Top 3	29.8	34.5	32.0	**11.3**	**34.3**
			Top 5%	34.1	34.2	34.1		
LLaMA	Self-citation	Self-citation	-	**37.9**	19.8	26.0	11.8	24.8
		MIRAGE$_{\text{EX}}$	Top 3	21.8	**29.6**	25.1		
			Top 5%	26.2	29.1	**27.6**		
	Standard	MIRAGE$_{\text{EX}}$	Top 3	17.8	20.9	19.2	**13.0**	**26.4**
			Top 5%	21.1	20.1	20.6		

Table 8: Citation quality (F1, Recall, Precision), correctness, and fluency of self-citation and MIRAGE on ELI5 when using self-citation instruction and standard prompts of Table 9.

tion quality evaluates the answer attribution performance with recall and precision scores. The recall score calculates if the concatenation of the cited documents entails the generated sentence. The precision measures if each document is cited precisely by verifying if the concatenated text still entails the generation whenever one of the documents is removed. We further calculate F1 scores to summarize the overall performance. **Correctness** checks whether the generated answer entails the golden reference answer according to the NLI model TRUE. Gold-reference answers are provided in the original dataset, and some were summarized by Gao et al. (2023b) by using GPT-4 in case they were too long. **Fluency** reflects the coherence and fluency of the generated response according to MAUVE (Pillutla et al., 2021), a popular NLG metric. We report the average score for all instances for each evaluation metric.

D ELI5 Evaluation with Standard Prompt

In the main experiments, we use self-citation prompts by Gao et al. (2023a) for MIRAGE answer attribution to control for the effect of different prompts on model responses, enabling a direct comparison with self-citation. In Table 8, we provide additional results where a standard prompt without citation instructions is used ("Standard" prompt in Table 9). We observe the overall citation quality of MIRAGE drops when a standard prompt is used instead of self-citation instructions. We conjecture this might be due to answers that are, in general, less attributable to the provided context due to a lack of explicit instructions to do so. We also observe higher correctness and fluency in the standard prompt setting, suggesting a trade-off between answer and citation quality.

SELF-CITATION (Gao et al., 2023a)
Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.
STANDARD
Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant). Use an unbiased and journalistic tone.

Table 9: Self-citation prompt, taken from Gao et al., 2023a, and standard prompt with no citation instruction.

E More Examples of Disagreement

Table 10 to 12 show three cases where MIRAGE answer attributions disagree with self-citation attributions of the same generation[17]. We adopt the Top-5% threshold for CCI Filtering. In Table 10, the generated answer becomes the consistent description '*cancel* the alarm' as mentioned in Document [3]. In this case, MIRAGE attributes this sentence to the corresponding Document [3] while NLI maintains its attribution of Document [3] due to lexical overlap, as suggested in Section 5.3.

On several occasions, we observe that MIRAGE attributes all occurrences of lexically similar tokens in the context when the LLM is generating the same word. For example, in Table 11 the named entity "Science ABC" is mentioned in both Document [1] and [4], and MIRAGE finds both occurrences as salient towards the prediction of the same entity in the output. Similarly, in Table 12, the generated word 'Document' is attributed to the previous mentions of the same word in the context. In both cases, when moving from token-level to sentence-level AA, this dependence would result in wrong AA according to NLI, since the documents are not entailing the answer, but rather making a specific token more likely. These cases reflect the possible discrepancy between AA intended as logical entailment and actual context usage during generation. Future work could explore more elaborate ways to aggregate granular information at sentence level while preserving faithfulness to context usage.

[17]Note that we remove citation tags in self-citation generated answers and use MIRAGE to attribute the resulting answers, as introduced in Section 5

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY
Document [2] [...] **Document [3]** [...] What will happen if you accidentally set off your security system? The siren will sound and it will be loud, but you should be able to stop the siren by entering your code into your keypad. [...] **Document [4]** [...]
Query: How does a small paper library bar code set off the security alarm when you haven't checked a book out.
ANSWER ATTRIBUTION RESULTS
Self-citation: [...] The response center may contact the individual to confirm if it's a false alarm, and the individual can cancel the alarm by providing their password at the keypad. [∅] [...]
MIRAGE: ... The response center (3) may (3) contact the individual to confirm if it (3)'s a false alarm, and the individual (3) can cancel (3) the alarm by providing (3) their password at the keypad (3).[3] ...
NLI (TRUE model): [3] entails the answer sentence.

Table 10: Example described in Section 5.3: MIRAGE attributes the generation to Document [3] when *cancel* is used instead of *prevent* (Table 5).

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

Document [1](Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity? » Science ABC): [...] They know that some people simply won't show up for their appointments [...] To account for this discrepancy, where the flight would have to fly with some very expensive seats empty, they decide to opt for a more profitable strategy. Assuming that out of the maximum capacity of 200 seats only 185 show up for the flight, this leaves 15 seats available. The airline could try to sell these seats quickly to avoid losses, but this rarely happens. What the airline does instead is try to sell 220 seats in advance.
Document [2] [...] **Document [3]** [...]
Document [4](Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity? » Science ABC): your own interests, as the airlines have clearly looked after theirs. First of all, if the airline is offering you a seat on a different flight, check if that seat is definitely available. If it is also full and it looks like you might be bumped yet again, you might as well be stranded! If the airline is offering compensation, such as free meals, hotel costs, and transportation between the airport and the hotel, look for the most profitable outcome. The airline might offer you vouchers for future flights, but since airline employees are given guidelines for negotiation, you might
Document [5] [...]

Query: Why are airlines able to sell more seats than are available on a flight?

ANSWER ATTRIBUTION RESULTS

Self-citation: [...] According to a report from Science ABC [1], airlines assume that out of the maximum capacity of 200 seats, only 185 will actually board the aircraft. To avoid empty seats and potential losses, airlines opt to sell up to 220 seats in advance [1]. [...]

MIRAGE: [...] According to a report from Science ABC (1)(4), airlines assume that out (1) of the maximum capacity (1) of 200 seats (1), only 185 (1) will actually board the aircraft. [1][4] [...]

NLI (TRUE model): Only [1] entails the answer sentence.

Table 11: Example of counterintuitive MIRAGE attribution: the generation is grounded to Document [4] because the named entity 'Science ABC' is made more probable by the occurrence of the same named entity 'Science ABC'. However, Document [4] does not logically entails the answer.

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

Document [1](Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity? » Science ABC): board the aircraft. They know that some people simply won't show up for their appointments (just like at hospitals, hotels, restaurants etc.). To account for this discrepancy, where the flight would have to fly with some very expensive seats empty, they decide to opt for a more profitable strategy. Assuming that out of the maximum capacity of 200 seats only 185 show up for the flight, this leaves 15 seats available. The airline could try to sell these seats quickly to avoid losses, but this rarely happens. What the airline does instead is try to sell 220 seats in advance.
Document [2] [...]
Document [3] [...]
Document [4](Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity? » Science ABC): your own interests, as the airlines have clearly looked after theirs. First of all, if the airline is offering you a seat on a different flight, check if that seat is definitely available. If it is also full and it looks like you might be bumped yet again, you might as well be stranded! If the airline is offering compensation, such as free meals, hotel costs, and transportation between the airport and the hotel, look for the most profitable outcome. The airline might offer you vouchers for future flights, but since airline employees are given guidelines for negotiation, you might
Document [5] [...]

Query: Why are airlines able to sell more seats than are available on a flight?

ANSWER ATTRIBUTION RESULTS

Self-citation: [...] According to a source in Document [1], airlines assume that not all passengers will show up for their flights, leaving some seats empty. [...]

MIRAGE: [...] According to a source in Document (4), airlines assume that not (1) all passengers will show up for their flights, leaving some seats empty. [1][4] [...]

NLI (TRUE model): Only [1] entails the answer sentence.

Table 12: Example of counterintuitive MIRAGE attribution: Document [4] is attributed by MIRAGE due to the repetition of the keyword 'Document'.