Towards Fine-Grained Citation Evaluation in Generated Text: A Comparative Analysis of Faithfulness Metrics

Weijia Zhang[1] **Mohammad Aliannejadi**[1] **Yifei Yuan**[2] **Jiahuan Pei**[3]
Jia-Hong Huang[1] **Evangelos Kanoulas**[1]
[1]University of Amsterdam [2]University of Copenhagen [3]Centrum Wiskunde & Informatica
w.zhang2@uva.nl

Abstract

Large language models (LLMs) often produce unsupported or unverifiable content, known as "hallucinations." To mitigate this, retrieval-augmented LLMs incorporate citations, grounding the content in verifiable sources. Despite such developments, manually assessing how well a citation supports the associated statement remains a major challenge. Previous studies use faithfulness metrics to estimate citation support automatically but are limited to binary classification, overlooking fine-grained citation support in practical scenarios. To investigate the effectiveness of faithfulness metrics in fine-grained scenarios, we propose a comparative evaluation framework that assesses the metric effectiveness in distinguishing citations between three-category support levels: *full*, *partial*, and *no* support. Our framework employs correlation analysis, classification evaluation, and retrieval evaluation to measure the alignment between metric scores and human judgments comprehensively. Our results show no single metric consistently excels across all evaluations, revealing the complexity of assessing fine-grained support. Based on the findings, we provide practical recommendations for developing more effective metrics.

1 Introduction

Large language models (LLMs) often generate content known as "hallucinations" (Li et al., 2022; Ji et al., 2022; Zhang et al., 2023b), which contradicts established knowledge or lacks verification from reliable sources. Mainstream studies (Bohnet et al., 2022; Gao et al., 2023a) aim to mitigate this by using retrieval-augmented LLMs to generate responses with in-line citations that provide supporting evidence. One primary challenge is to assess how well a citation supports its statement, as manual evaluation is labor-intensive and time-consuming. Automated citation evaluation has been explored to reduce reliance on human assessments (Gao et al., 2023b; Li et al., 2024b). To this end, faithfulness evaluation metrics are employed as proxies to automatically estimate the citation support (Xia et al., 2024; Li et al., 2024a). These metrics measure the faithfulness between model-generated and sourced text, which aligns closely with the objectives of automated citation evaluation.

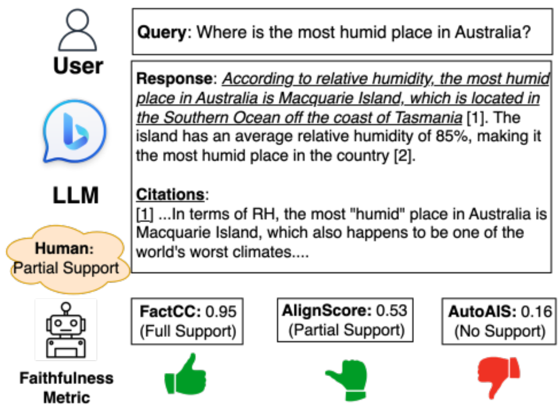

Figure 1: An example of *partial support* in citation evaluation. Inconsistent metric scores are observed when assessing the statement with three faithfulness metrics.

Prior studies in faithfulness metrics have primarily limited this task to a binary classification problem (Tahaei et al., 2024; Huang et al., 2024d), where faithfulness metrics are leveraged to determine whether a citation supports the associated statement. However, this binary approach fails to capture the fine-grained citation support encountered in real-world applications. For instance, in Figure 1, a retrieval-augmented LLM generates a response with multiple citations given a query. A human assessor labels the first citation as *"partial support"* since it only supports "the most humid place in Australia is Macquarie Island" but not "which is located in the Southern Ocean off the coast of Tasmania." This partial support scenario causes noticeable inconsistencies across three dif-

Proceedings of the 17th International Natural Language Generation Conference, pages 427–439
September 23–27, 2024. ©2024 Association for Computational Linguistics

ferent faithfulness metrics. Therefore, there is a significant research need to evaluate the effectiveness of faithfulness metrics in accurately distinguishing citations in such fine-grained support scenarios.

To address this issue, we propose a comparative evaluation framework for assessing the metric effectiveness in fine-grained support scenarios. In our framework, we define "*support levels*" as the extent to which a citation supports the associated statement (Liu et al., 2023; Yue et al., 2023). Specifically, we consider a three-category support level scenario: *full*, *partial*, and *no* support. These categories indicate whether a citation fully, partially or does not support the associated statement, respectively. To comprehensively assess the metric effectiveness, we measure the alignment between metric scores and human judgments with three types of evaluation protocols: 1) *Correlation analysis:* we employ it to measure how well metric scores correlate with human judgments. 2) *Classification evaluation:* we conduct a classification evaluation to assess the metrics' capability to distinguish citations based on their support levels. 3) *Retrieval evaluation:* we undertake a retrieval evaluation to assess the metric effectiveness in ranking citations according to their support levels. This is motivated by the observation that the previous two evaluation protocols assume citations are within statements, which is not always valid in practice (Asai et al., 2024). In such cases, faithfulness metrics are adapted to perform post-hoc retrieval, aiming to retrieve potential citations from a candidate pool (Kang et al., 2023; Gou et al., 2024). Thus, retrieval evaluation is crucial for determining the practical utility of these metric adaptations.

In our experiments, we assess various widely used faithfulness metrics, categorizing them into *similarity-based*, *entailment-based*, and *LLM-based* metrics. We find that: 1) No single faithfulness metric consistently outperforms others across three evaluation protocols, suggesting that these protocols are complementary and should be integrated for a comprehensive evaluation of metric performance; 2) The best-performing metrics show promise in distinguishing some support scenarios but struggle with others. This highlights the inherent complexities of automated citation evaluation. 3) Similarity-based metrics surpass best-performing entailment-based metrics in retrieval evaluation. This indicates that entailment-based metrics exhibit higher sensitivity to noisy data, which is introduced by irrelevant documents in such scenarios.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to systematically investigate the effect of fine-grained support levels on faithfulness metrics in the task of automated citation evaluation.
- We propose a comparative evaluation framework to assess the alignment between metric scores and human judgments. This framework includes three evaluation protocols to comprehensively evaluate the metric performance.
- Our experimental results demonstrate the best-performing faithfulness metrics still struggle to identify partially supporting citations, underscoring the inherent challenges of automated citation evaluation. Based on our findings, we offer practical recommendations for the development of more effective metrics.

2 Related Work

Faithfulness Evaluation Metrics Faithfulness evaluation metrics are crucial for assessing the factual consistency of text generated by models relative to the source text. It receives great interest within the field of natural language generation (NLG) (Huang et al., 2019, 2021b; Zhang et al., 2021, 2023a; Huang et al., 2024b,c; Zhu et al., 2024), particularly in abstractive summarization (Maynez et al., 2020; Kryscinski et al., 2020; Huang and Worring, 2020; Huang et al., 2021a; Zhang et al., 2024). In general, faithfulness metrics are categorized into three types: entailment-based, similarity-based, and QA-based metrics. Entailment-based metrics employ natural language inference (NLI) models to determine if the source text entails the generated text (Falke et al., 2019; Laban et al., 2022; Honovich et al., 2022; Zha et al., 2023). Similarity-based metrics, such as BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), quantify text similarity and have demonstrated robust performance in faithfulness evaluation (Pagnoni et al., 2021; Honovich et al., 2022). QA-based metrics utilize a combination of question generation and question answering to estimate faithfulness levels (Durmus et al., 2020; Wang et al., 2020; Scialom et al., 2021; Fabbri et al., 2022). In this work, we exclude QA-based metrics from our work, following recent works suggesting the challenging limitations in these metrics (Kamoi et al., 2023). We focus on the extrinsic evaluation of faithfulness metrics against human judgments in scenarios requiring fine-grained citation support.

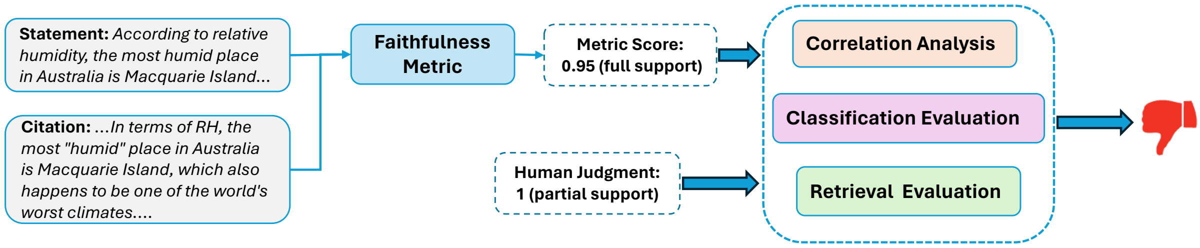

Figure 2: The overview of the proposed comparative evaluation framework. A faithfulness metric assigns scores to given statements and their corresponding citations. Subsequently, our framework comprehensively assesses the alignment between these metric scores and human judgments by employing correlation analysis, classification, and retrieval evaluation.

Citation Evaluation Citation evaluation seeks to enhance the trustworthiness of retrieval-augmented LLMs by verifying the support provided by citations to the generated statements (Rashkin et al., 2023; Yue et al., 2023; Huang and Chang, 2023; Huang et al., 2024a). Given the labor-intensive nature of manual citation evaluation, there has been a shift towards automated approaches to reduce dependence on human evaluation. Since the goals of automated citation evaluation align closely with faithfulness evaluation in NLG, faithfulness metrics are employed to verify whether a citation supports the corresponding statement (Li et al., 2024c; Sun et al., 2023; Ye et al., 2024; Li et al., 2024d; Shen et al., 2024; Huang et al., 2024d). Despite their widespread usage, the effectiveness of these metrics in more practical fine-grained citation support scenarios, such as those involving partial support by citations, has not been adequately addressed. Questions remain about the metrics' capability to differentiate citations in these fine-grained scenarios. This work addresses these gaps by examining the effectiveness of faithfulness metrics across three distinct levels of citation support: full, partial, and no support.

3 Evaluation Framework

In this section, we introduce the proposed comparative evaluation framework. We begin by formalizing the task of automated citation evaluation. Subsequently, we detail three distinct evaluation protocols within this framework, ensuring a comprehensive assessment in alignment between faithfulness metrics and human judgments. Our framework is demonstrated in Figure 2.

3.1 Task Formulation

The objective of automated citation evaluation is to automatically quantify the support level of a citation based on the citation and its associated statement. In this work, we assume access to a dataset for automated citation evaluation, comprising pairs of statements and their corresponding citations, denoted as (s_i, c_i). Each s_i is a statement from the set S of all statements produced by an LLM and each c_i is a citation from a set C of citations returned by the LLM. We categorize the citations into three distinct support levels: full, partial, and no support. We adopt the definition of these support levels from Liu et al. (2023):

- Full Support (FS): the citation fully supports every detail in the statement.
- Partial Support (PS): the citation supports certain aspects of the statement, while other details remain unsupported or are contradicted.
- No Support (NS): none of the content in the statement is supported by the citation. For instance, the citation is entirely irrelevant or contradicts the statement.

To this end, without loss of generality, we define a faithfulness metric as a scoring function, denoted as $F(s_i, c_i) \rightarrow R^+$. For any given statement s_i and its associated citation c_i, this scoring function provides a numeric score that indicates the extent of support provided by the citation to the statement.

3.2 Evaluation Protocols

The objective of evaluation protocols is to comprehensively assess the extent to which metric scores align with human judgments. In this work, we assess this alignment across three distinct dimensions: **correlation**, **classification performance**, and **retrieval effectiveness**.

3.2.1 Correlation Analysis

The correlation analysis measures the general trend in the relationship between metric scores and human judgments. Previous research (Kryscinski et al., 2020; Pagnoni et al., 2021) has employed correlation analysis to meta-evaluate faithfulness metrics in abstractive text summarization. They involve measuring the extent to which metric scores align with binary levels of faithfulness, which are annotated by human assessors as either faithful (1) or unfaithful (0). Inspired by them, we adapt correlation analysis to the task of automated citation evaluation. Specifically, given the statements and their associated citations, we assess how well predicted metric scores correlate with human-annotated support levels. To facilitate correlation analysis, we assign support levels $\{\mathrm{FS}, \mathrm{PS}, \mathrm{NS}\}$ to values $\{0, 1, 2\}$. We then utilize standard correlation metrics to assess metric performance. The details are shown in Section 5.2.

3.2.2 Classification Evaluation

In addition to correlation analysis, we perform classification evaluation to determine the metric effectiveness in discriminating citations based on their support level. Specifically, the metrics need to categorize a citation into one of three support levels: FS, PS, NS. Notably, existing faithfulness metrics do not apply to this three-way classification scenario, as they are unable to accurately determine the extent to which a statement is partially supported by its corresponding citation (Laban et al., 2022). To address this issue, we adopt a one-vs-one strategy, by effectively decomposing the three-way classification into three binary classification task settings: (i) Full Support vs. No Support (FS-vs-NS), (ii) Full Support vs. Partial Support (FS-vs-PS), and (iii) Partial Support vs. No Support (PS-vs-NS). For each binary classification task setting, we construct a specialized dataset comprising only instances with the corresponding binary support levels derived from the original dataset. We assess the performance of metrics on these tailored binary datasets using standard binary classification evaluation metrics. The overall metric performance is then computed by averaging the results across all binary tasks.

3.2.3 Retrieval Evaluation

The objective of retrieval evaluation is to measure the metric effectiveness in ranking citations according to their support levels. This evaluation is motivated by the observation that previous correlation and classification evaluations presuppose the presence of citations within generated statements. However, real-world scenarios frequently present instances where citations are absent or irrelevant, highlighting the need for post-hoc retrieval to enhance citation quality (Liu et al., 2023; Huang et al., 2024a). In post-hoc retrieval, candidate documents are retrieved to form a pool of potential citations using information retrieval techniques (Karpukhin et al., 2020). Faithfulness metrics are then employed to rank citations based on their predicted metric scores, aiming to identify the citation with the highest support level. Ideally, a faithfulness metric should rank fully supporting citations at the top, followed by partially supporting citations, and finally non-supporting citations. Similar to correlation analysis, we assign support levels $\{\mathrm{FS}, \mathrm{PS}, \mathrm{NS}\}$ to relevance labels $\{2, 1, 0\}$. The metric effectiveness is assessed using standard information retrieval evaluation metrics. This evaluation also provides a deeper understanding of metric performance in post-hoc citation retrieval scenarios.

4 Faithfulness Metrics

In our experiments, we evaluate diverse faithfulness evaluation metrics, dividing them into similarity-based, entailment-based, and LLM-based metrics. Similarity-based metrics assess the support levels mainly based on the degree of similarity between the citation and the associated statement. Entailment-based metrics leverage pre-trained NLI models to estimate the support levels. LLM-based metrics directly prompt LLMs to measure the support levels.

4.1 Similarity-Based Metrics

BERTScore (Zhang et al., 2020) adopts BERT (Devlin et al., 2019) to measure semantic similarity between a pair of text by aggregating cosine similarity among token-level BERT representation without further fine-tuning. We report the precision version of BERTScore since it correlates more with human judgments in faithfulness evaluation (Pagnoni et al., 2021), We use recommended `deberta-xlarge-mnli` (He et al., 2021) as the backbone model.

BARTScore (Yuan et al., 2021) adopts BART (Lewis et al., 2020) to measure the similarity between two texts based on conditional log-likelihood of generating target text from source text. In our experiments, we leverage the faithfulness version of BARTScore, in which we treat the citation and the statement as the source and target text, respectively. We use the BART model fine-tuned on the CNN/DailyMail dataset (Hermann et al., 2015) as the backbone model.

Human Judgment	# Statement-Citation Pair
Full Support	6,616
Partial Support	1,445
No Support	4,620
Total	12,681

Table 1: Data statistics of the GenSearch dataset. Each pair has been annotated by human assessors based on three categories: full, partial, and no support.

4.2 Entailment-Based Metrics

FactCC (Kryscinski et al., 2020) is a BERT-based model to verify whether a generated text is faithful to a source text, which is fine-tuned on synthetic training data containing simulated examples with different factual errors (Kryscinski et al., 2020).

SummaC (Laban et al., 2022) is a RoBERTa-based model (Liu et al., 2019) fine-tuned on NLI datasets. This metric splits source and generated texts into sentences, computes entailment scores for each pair, and aggregates these scores to obtain the final faithfulness score. It has two variants: (i) SummaC$_{\text{ZS}}$ is a zero-shot version that is only pre-trained on NLI datasets; (ii) SummaC$_{\text{Conv}}$ adds extra convolutional layers and is further fine-tuned on synthetic training data proposed in Kryscinski et al. (2020).

AutoAIS (Honovich et al., 2022) is a T5-11B (Raffel et al., 2020) model trained on a collection of NLI datasets, which is commonly used in recent automated citation evaluation. As the original output of AutoAIS is a numeric, either "1" (faithful) or "0" (unfaithful), we utilize the generated token probability of "1" as the predicted metric score.

AlignScore (Zha et al., 2023) further fine-tunes a RoBERTa-based model (Liu et al., 2019) with a unified alignment loss function. To this end, a unified dataset containing a variety of related natural language processing datasets has been collected. In this work, we adapt the `large` version as it demonstrates the best performance.

4.3 LLM-Based Metrics

In addition to established faithfulness metrics, we utilize LLMs as faithfulness evaluators for comparison. Specifically, we introduce two prompting methods as follows: (i) **Discrete scoring** prompts the LLM to assign discrete scores from the set $0, 1, 2$ for a given statement and its citation, where 0, 1, and 2 indicate no support, partial support, and full support, respectively; (ii) **Continuous scoring** prompts the LLM to assign continuous scores in the range $[0, 1]$ for a given statement and its citation. Here, 1 indicates full support, 0 indicates no support, and values between 0 and 1 indicate partial support.

Metric	Pearson	Spearman	Kendall
LLM-based			
GPT-3.5-CON	0.023	0.057	0.035
GPT-3.5-DIS	0.101	0.181	0.128
Entailment-based			
FactCC	0.121	0.199	0.140
SummaC$_{\text{ZS}}$	0.364	0.180	0.137
SummaC$_{\text{Conv}}$	0.565	0.444	0.342
AlignScore	0.585	<u>0.488</u>	<u>0.393</u>
AutoAIS	**0.638**	**0.639**	**0.547**
Similarity-based			
BERTScore	0.542	0.227	0.170
BARTScore	<u>0.598</u>	0.235	0.176

Table 2: Correlation coefficients between human-annotated support levels and metric scores on the GenSearch dataset. The best and second-best correlations are marked in **bold** and <u>underline</u>, respectively.

In the experiments, we employ the latest version of GPT-3.5 (`gpt-3.5-turbo-0125`) as the base model. Moreover, we utilize the chain of thought (CoT) method (Wei et al., 2022; Kojima et al., 2022) to enhance the reasoning capabilities of the LLM. We use GPT-3.5-DIS and GPT-3.5-CON to denote GPT-3.5 using discrete and continuous scoring methods, respectively. The detailed prompts are shown in Appendix A.

5 Experiments

In this section, we describe the dataset used in the experiments. Subsequently, we discuss the evaluation metrics incorporated within our proposed framework, which assess the performance of faithfulness metrics in alignment with human judgments.

5.1 Datasets

In our experiments, we utilize the GenSearch dataset (Liu et al., 2023) as our evaluation benchmark, which consists of data from generative search engines (GSE) like BingChat.[1] These GSEs represent commercial applications of retrieval-augmented LLMs. As depicted in Figure 1, each example includes a user query and a corresponding response generated by the GSE. The user queries

[1] `https://www.bing.com/chat`

Category	Metric	FS-vs-NS	FS-vs-PS	PS-vs-NS	Overall
LLM-based	GPT-3.5-CON	54.80	54.13	51.60	53.51
	GPT-3.5-DIS	57.84	52.79	55.48	55.37
Entailment-based	FactCC	68.45	62.58	56.39	62.47
	SummaC$_{\text{ZS}}$	78.60	72.96	58.67	70.08
	SummaC$_{\text{Conv}}$	85.01	78.74	61.84	75.20
	AlignScore	90.79	<u>81.41</u>	69.78	80.66
	AutoAIS	**92.61**	**82.31**	<u>73.90</u>	**82.94**
Similarity-based	BARTScore	87.43	75.42	71.34	78.07
	BERTScore	<u>91.55</u>	75.94	**78.72**	<u>82.07</u>

Table 3: Classification performance of faithfulness metrics regarding ROC-AUC score (%) on the GenSearch dataset. The overall performance is the macro-averaged performance of three binary classification settings. The best and second-best scores are marked in **bold** and <u>underline</u>, respectively.

are sourced from various QA datasets (Fan et al., 2019; Kwiatkowski et al., 2019). Each response consists of multiple statements, each containing in-line citations linking to web documents. Notably, these statements are supported by one or more citations. For this benchmark, human assessors are enrolled to annotate each statement-citation pair based on the degree to which the citation supports the associated statement.

Data Statistics The GenSearch dataset comprises a total of $12,681$ statement-citation pairs. For each pair, human assessors categorize the citation into one of three categories of support levels: full, partial, or no support. The details of data statistics are shown in Table 1. Notably, for citations classified under the full or partial support categories, human assessors additionally extract explicit evidence sentences from the citation that support the associated statement.

Data Processing While the GenSearch dataset aligns well with our research objectives, we encounter a significant challenge: the extensive length of most citations within the dataset. These citations often comprise a web document with thousands of words, far exceeding the maximum input capacity of most faithfulness metrics, which is limited to 512 tokens. This limitation necessitates input truncation, potentially compromising the reliability of faithfulness metrics. To mitigate this issue, we adopt a strategy similar to previous studies (Zha et al., 2023). Specifically, we segment each cited document into shorter text chunks, with a maximum length of 150 words per chunk. These text chunks, along with their corresponding statements, serve as the inputs for faithfulness metrics to predicted metric scores. Furthermore, to determine human judgments for the text chunks, we employ the Jaccard similarity index to identify text chunks containing human-annotated evidence sentences, classifying them as either fully or partially supporting text chunks.

5.2 Evaluation Metrics

We report Pearson, Spearman, and Kendall coefficients for correlation analysis, as recommended by previous research (Pagnoni et al., 2021). In terms of classification evaluation, following previous studies (Honovich et al., 2022; Ma et al., 2023), we report the macro-averaged Receiver Operating Characteristic-Area Under Curve (ROC-AUC) score, as it obviates the need for manual threshold setting for each binary classification task. For retrieval evaluation, we report standard normalized discounted cumulative gain (NDCG@n) scores where $n \in \{5, 10, 20\}$.

6 Results and Analyses

In this section, we discuss the performance of faithfulness metrics across three distinct evaluation protocols. Subsequently, we conduct a qualitative analysis through case studies.

6.1 Correlation Results

The correlation results are demonstrated in Table 2. The following observations can be made: 1) The best-performing metrics reveal moderate correlations when analyzed using the Pearson coefficient. For instance, AutoAIS achieves the highest Pearson coefficient, recording a value of 0.638, largely surpassing the second-best BARTScore, which posts a coefficient of 0.598. 2) There is notable variation in correlation trends among high-performing metrics. BARTScore shows the second-best Pearson

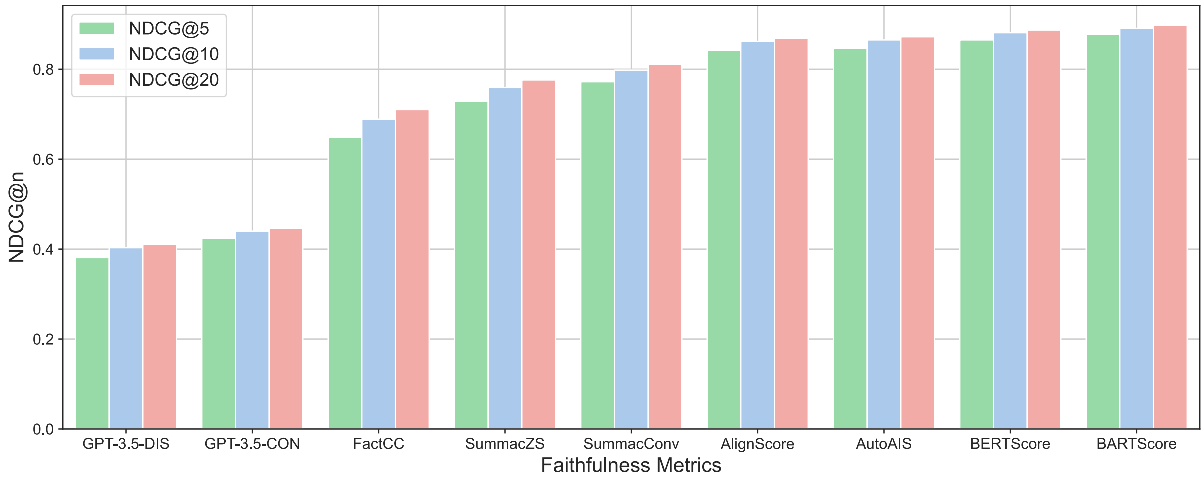

Figure 3: Retrieval performance of faithfulness metrics regarding NDCG@n scores on the GenSearch dataset. Note that we assign relevance labels 2, 1, and 0 to full, partial, and no support, respectively (shown in the color).

correlation but much lower Spearman and Kendall correlations. This divergence likely arises from the Pearson coefficient's assumption of linear relationships between two variables, which is often invalid in automated citation evaluation. 3) Similarity-based metrics generally show lower Spearman and Kendall correlations compared to Pearson. For instance, BERTScore has a substantial Pearson correlation of 0.542 but lower Spearman and Kendall correlations of 0.227 and 0.170. This indicates that similarity-based metrics do not align well with human judgments, highlighting their limitations in fine-grained support scenarios. 4) LLM-based metrics show little correlation with human judgments among all correlation coefficients, with the correlation of the GPT-3.5-CON metric being almost zero. This finding suggests a negligible relationship between LLM-based metric scores and human judgments. Furthermore, the GPT-3.5-DIS metric significantly outperforms GPT-3.5-CON, highlighting that more fine-grained support levels present greater challenges in correlation analysis.

6.2 Classification Results

Table 3 presents the results of the classification evaluation. The observations can be summarized as follows: 1) Among all three binary classification task settings, most faithfulness metrics demonstrate superior performance in the FS-vs-NS setting. Notably, entailment-based AutoAIS achieves the highest ROC-AUC score of 92.61, which shows significant discriminability between full support and no support instances. This can be attributed to its much more extensive parameters compared to other entailment-based metrics. 2) We observe the performance decline across the other two settings (i.e. FS-vs-PS and PS-vs-NS). For instance, when comparing the FS-vs-NS and PS-vs-NS settings, the ROC-AUC score of AutoAIS diminishes from 92.61 to 73.90. This decline indicates that even the best-performing metric struggles with granular sensitivity to varying levels of support. 3) While entailment-based AutoAIS generally surpasses other metrics, it is outperformed by similarity-based BERTScore in the PS-vs-NS setting. Interestingly, while most metrics perform worst in this setting, BERTScore shows its least effectiveness in FS-vs-PS. This highlights the unique prediction behaviors of different metrics across binary classification settings. 4) The performance of LLM-based metrics significantly lags behind other metrics. For instance, GPT-3.5-DIS achieves only a ROC-AUC score of 57.84 in the FS-vs-NS setting, markedly lower than the best-performing AutoAIS, which achieves a ROC-AUC score of 92.61. Furthermore, the overall performance of LLM-based metrics approaches random guessing. This underscores the inefficacy of LLM-based metrics in distinguishing fine-grained support levels.

6.3 Retrieval Results

Figure 3 presents the results of the retrieval evaluation. The key findings are as follows: 1) Similarity-based metrics, BARTScore and BERTScore, outperform other entailment-based metrics in all NDCG@n scores. For instance, entailment-based AutoAIS exhibits weaker NDCG@5 scores than BARTScore. This is likely because entailment-based metrics are more sensitive to noisy information than similarity-based metrics, as many

Error Reason	Example
The citation does not explicitly mention coreference.	**Statement:** Others believe that performance-enhancing drugs should be allowed in sports. **Citation:** However, if children are allowed to train as professional athletes, then they should be allowed to take the same drugs , provided that they are no more dangerous than their training is ... **Human Judgment:** full support **Metric Score:** 0.055 (no support)
The complex statement includes independent claims.	**Statement:** Love leads to growth while being in love is about ownership ... **Citation:** " Growing to love the real person and accepting who they are, with both strengths and weaknesses, can make a wonderful difference in your relationship," McCoy says ... **Human Judgment:** partial support **Metric Score:** 0.0004 (no support)
The citation is semantically similar but non-supporting.	**Statement:** Carpal tunnel syndrome can be treated with various methods, including wrist splinting, anti-inflammatory medication , and surgery. **Citation:** If diagnosed and treated early, the symptoms of carpal tunnel syndrome can often be relieved without surgery . If your diagnosis is uncertain or if your symptoms are mild, your doctor will recommend nonsurgical treatment first ... **Human Judgment:** no support **Metric Score:** 0.52 (partial support)

Table 4: Case study of the faithfulness metric AutoAIS. Green phrases indicate supported content in the statement and corresponding supporting evidence. Red phrases indicate unsupported content in the statement and corresponding misleading information in the citation.

irrelevant documents exist in retrieval scenarios. It suggests the need for the robustness improvements of metrics in post-hoc retrieval scenarios. 2) The best-performing BERTScore achieves more than twice the NDCG@n scores compared to LLM-based metrics. This result suggests that LLM-based metrics are ineffective in ranking documents with higher support levels. A plausible explanation is that LLM-based metrics lack fine-grained sensitivity to variations in support levels. Interestingly, our observations reveal that GPT-3.5-CON surpasses GPT-3.5-DIS, highlighting the advantage of fine-grained scoring methods in retrieval evaluation. 3) NDCG@n scores effectively capture the performance variations as the number of text chunks increases. For instance, as the chunk count increases, BARTScore shows a marginal performance improvement, while FactCC exhibits a more pronounced enhancement.

6.4 Case Study

Table 4 presents three cases of AutoAIS. In the first example, where human judgment indicates full support. AutoAIS incorrectly assigns a very low score. This may be due to the lack of explicit mention of drug coreference in the cited text chunk. This indicates coreference resolution remains a significant challenge in automated citation evaluation. In the second example, where human judgment indicates partial support. The complex statement implicitly contains two independent claims that require verification. However, the provided citation fails to offer sufficient evidence, resulting in an almost zero metric score. In the third example, where human judgment indicates no support. The given citation is semantically similar to the statement, leading to a metric score of partial support. Despite this semantic similarity, specific treatments mentioned in the statement, such as wrist splinting, are not explicitly referenced in the citation.

7 Discussions

Overall, our results across three evaluation protocols indicate that the evaluation protocols are complementary and should be integrated for a comprehensive assessment of metrics. Based on the evaluation results, we further propose the following practical recommendations to develop more effective metrics for automated citation evaluation: 1) **Development of training resources:** motivated by the observation that the best-performing metrics still struggle with identifying partial support, we recommend the development of training resources that include fine-grained support level annotations. These resources could significantly enhance the metrics' fine-grained sensitivity to varying support levels; 2) **Introduction of contrastive learning:** to improve the robustness of metrics in post-hoc retrieval scenarios, we recommend fine-tuning metrics using contrastive learning frameworks. This method has demonstrated effectiveness across various information retrieval tasks (Izacard et al., 2022). 3) **Development of more explainable metrics:**

traditional faithfulness metrics often only provide final scores without sufficient explainability (Xu et al., 2023). This limitation hinders a deeper understanding of the models' behavior. Therefore, it is crucial to develop more explainable faithfulness metrics, potentially using large language models (LLMs).

8 Conclusion

We propose a comparative evaluation framework to explore the efficacy of faithfulness metrics beyond the binary scenario by examining three levels of citation support. Our framework employs correlation analysis, classification evaluation, and retrieval evaluation to measure the alignment between metric scores and human judgments. Experimental results reveal that no single metric consistently excels across all evaluation protocols, indicating the complexity of automated citation evaluation and the limitations of existing faithfulness metrics. We provide practical suggestions based on the findings.

Limitations

In this work, we consider a citation that explicitly contains human-annotated evidence as the fully supporting citation for each statement. However, for some complex statements, their evidence is distributed among multiple citations. For instance, about 2% statements on the GenSearch dataset require multiple citations to be fully supported. Also, we focus on statement-level citation evaluation. Since answer-level citation evaluation is much more complicated and requires proper aggregation methods, we leave this exploration as future work. We do not evaluate QA-based faithfulness metrics as a recent study shows that such metrics have some fundamental issues, such as failing to localize errors (Kamoi et al., 2023). However different findings could be explored with QA-based metrics.

Ethical Considerations

We realized there are some risks in exploring citation evaluation for LLM-generated text. Since we have used publicly available datasets and open-source implementation of faithfulness metrics, we carefully avoid potential ethical problems caused by datasets or open-source codes. As we address the issue of the effectiveness of faithfulness metrics for LLM-generated text, concerning hallucination. We acknowledged the hallucinated text generated by LLMs may contain potentially harmful or misleading information. Our final goal is to mitigate such hallucination issues, which should support the discussion around hallucinations of LLMs and all ethical aspects around them.

Acknowledgments

This research was supported by the China Scholarship Council (CSC) under grant number 202008440470. The views expressed in the content belong solely to the authors and may not reflect the perspectives or endorsements of their respective employers or sponsors.

References

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*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-Based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. 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.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-Enhanced Bert with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

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.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024a. Training language models to generate text with citations via fine-grained rewards. *CoRR*, abs/2402.04315.

Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. 2019. A novel framework for robustness analysis of visual qa models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8449–8456.

Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. 2021a. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 580–589.

Jia-Hong Huang and Marcel Worring. 2020. Query-controllable video summarization. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 242–250.

Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, et al. 2021b. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2442–2452.

Jia-Hong Huang, Chao-Chun Yang, Yixian Shen, Alessio M Pacces, and Evangelos Kanoulas. 2024b. Optimizing numerical estimation and operational efficiency in the legal domain through large language models. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

Jia-Hong Huang, Hongyi Zhu, Yixian Shen, Stevan Rudinac, Alessio M. Pacces, and Evangelos Kanoulas. 2024c. A novel evaluation framework for image2text generation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval, LLM4Eval Workshop*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *CoRR*, abs/2307.02185.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024d. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14095–14113.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Ryo Kamoi, Tanya Goyal, and Greg Durrett. 2023. Shortcomings of question answering based factuality frameworks for error localization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 132–146.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *CoRR*, abs/2311.09114.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Tom Kwiatkowski, Jennimaria Palomaki, and et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024a. Improving attributed text generation of large language models via preference learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5079–5101.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024b. LLatrieval: LLM-verified retrieval for verifiable 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)*, pages 5453–5471.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024c. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 493–516.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024d. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14919–14935.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liang Ma, Shuyang Cao, Robert L Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. BUMP: A benchmark of unfaithful minimal pairs for meta-evaluation of faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12788–12812.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Colin Raffel, Noam Shazeer, Adam Roberts, and et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

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):777–840.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Jiajun Shen, Tong Zhou, Suifeng Zhao, Yubo Chen, and Kang Liu. 2024. Citekit: A modular toolkit for large language model citation generation. *arXiv preprint arXiv:2408.04662*.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2023. Towards verifiable text generation with evolving memory and self-reflection. *CoRR*, abs/2312.09075.

Marzieh Tahaei, Aref Jafari, Ahmad Rashid, David Alfonso-Hermelo, Khalil Bibi, Yimeng Wu, Ali Ghodsi, Boxing Chen, and Mehdi Rezagholizadeh. 2024. Efficient citer: Tuning large language models for enhanced answer quality and verification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4443–4450.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2024. Ground every sentence: Improving retrieval-augmented llms with interleaved reference-claim generation. *arXiv preprint arXiv:2407.01796*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994.

Xi Ye, Ruoxi Sun, Sercan 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)*, pages 6237–6251.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, pages 27263–27277.

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.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Weijia Zhang, Vaishali Pal, Jia-Hong Huang, Evangelos Kanoulas, and Maarten de Rijke. 2024. QFMTS: Generating query-focused summaries over multi-table inputs. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*.

Weijia Zhang, Svitlana Vakulenko, Thilina Rajapakse, and Evangelos Kanoulas. 2021. Scaling up query-focused summarization to meet open-domain question answering. *ArXiv preprint, abs/2112.07536*.

Weijia Zhang, Svitlana Vakulenko, Thilina Rajapakse, Yumo Xu, and Evangelos Kanoulas. 2023a. Tackling query-focused summarization as a knowledge-intensive task: A pilot study. *The First Workshop on Generative Information Retrieval (Gen-IR) at SIGIR*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. 2024. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 978–987.

A Details of Prompts

Details of prompts used in the paper are shown in Table 5.

Prompt Name	Prompt Content
Discrete Scoring	**Instruction:** Your task is to quantify how well a provided citation supports a given statement. You should predict a *discrete* score from the set $\{0, 1, 2\}$, where 0, 1, 2 represent that the statement is not supported, partially supported, and fully supported, respectively. Let's think step by step. **Statement:** {*statement*} **Citation:** {*cited text chunk*} **Prediction:**
Continuous Scoring	**Instruction:** Your task is to quantify how well a provided citation supports a given statement. You should predict a *continuous* score between 0 and 1 (inclusive), where 0 is not supported, 1 is fully supported, and a float value between 0 and 1 is partially supported. Let's think step by step. **Statement:** {*statement*} **Citation:** {*cited text chunk*} **Prediction:**

Table 5: Detailed prompts for discrete and continuous scoring methods.