A Comparative Analysis of Faithfulness Metrics and Humans in Citation Evaluation

Weijia Zhang
University of Amsterdam
Amsterdam, Netherlands

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, Netherlands

Jiahuan Pei
Centrum Wiskunde & Informatica
Amsterdam, Netherlands

Yifei Yuan
University of Copenhagen
Copenhagen, Denmark

Jia-Hong Huang
University of Amsterdam
Amsterdam, Netherlands

Evangelos Kanoulas
University of Amsterdam
Amsterdam, Netherlands

arXiv:2408.12398v1 [cs.IR] 22 Aug 2024

ABSTRACT

Large language models (LLMs) often generate content with unsupported or unverifiable content, known as "hallucinations." To address this, retrieval-augmented LLMs are employed to include citations in their content, 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 tackle this challenge by leveraging faithfulness metrics to estimate citation support automatically. However, they limit this citation support estimation to a binary classification scenario, neglecting 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 indicate no single metric consistently excels across all evaluations, highlighting the complexity of accurately evaluating fine-grained support levels. Particularly, we find that the best-performing metrics struggle to distinguish partial support from full or no support. Based on these findings, we provide practical recommendations for developing more effective metrics.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; • **Information systems** → Evaluation of retrieval results.

KEYWORDS

Faithfulness metrics; Citation evaluation; Large language models

1 INTRODUCTION

Large language models (LLMs) suffer from generating content known as "hallucinations" [51], which refers to content that either contradicts established world knowledge or cannot be verified by any reliable source of information. Mainstream studies [1, 7, 27] aims to mitigate this issue by leveraging retrieval-augmented LLMs to generate responses with in-line citations, which contain supporting evidence to verify the statements in the responses. One primary challenge in this field is to assess how well a citation supports its associated statement, since manually evaluating citations is labor-intensive and time-consuming. To this end, automated citation evaluation has been explored to minimize reliance on human assessments [1, 27]. Given the early stage of this research, faithfulness evaluation metrics [10, 23, 32] have been employed as proxies to automatically estimate the support levels of the citations [6, 7]. This is motivated by the observation that these metrics measure the extent to how faithful the model-generated text is to the sourced text, aligning closely with the objectives of automated citation evaluation.

Prior studies [6, 7] have primarily limited the application of faithfulness metrics in automated citation evaluation to a binary classification scenario. In this scenario, faithfulness metrics are solely tasked with determining whether a citation supports the associated statement. This binary approach fails to capture the fine-grained citation support encountered in real-world applications. For instance, consider a "*partial support*" scenario illustrated in Figure 1. Given a

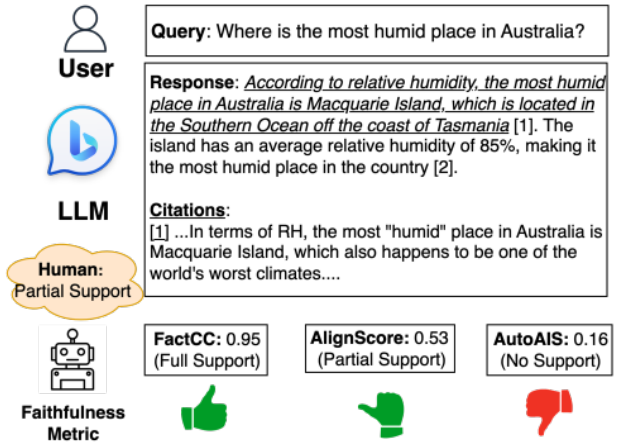

Figure 1: An example of *partial support* in citation evaluation. A retrieval-augmented LLM generates a response that includes citations based on a given user query. The human assessor annotates that the first citation partially supports the associated statement. Inconsistent metric scores are observed when assessing the statement using three distinct faithfulness metrics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
LLM4Eval@SIGIR2024, July 18, 2024, Washington D.C., USA
© 2024 Copyright held by the owner/author(s).

user query "*Where is the most humid place in Australia?*", a retrieval-augmented LLM generates a response along with multiple citations. A human assessor categorizes the first citation as "partial support" since it only supports the initial segment of the statement: "*the most humid place in Australia is Macquarie Island*". However, it does not provide evidence for the latter part of the statement: "*which is located in the Southern Ocean off the coast of Tasmania*". The complexity of this partial support scenario leads to noticeable inconsistencies across three distinct faithfulness metrics. However, the effectiveness of faithfulness metrics in accurately distinguishing citations in such fine-grained citation support scenarios remains largely under-explored.

To address the issue above, we propose a comparative evaluation framework designed to assess the effectiveness of faithfulness metrics against human judgments in fine-grained levels of support scenarios. In our framework, we define "*support levels*" as the extent to which a citation supports the corresponding statement [29, 43]. More specifically, in contrast to previous studies that predominantly focus on binary classification scenarios, our framework aims to evaluate the effectiveness of faithfulness metrics in a three-category support level scenario: *full support*, *partial support*, and *no support*. These categories indicate whether a citation provides *full*, *partial*, or *no* support to the associated statement. To comprehensively assess the metric effectiveness, we measure the alignment between metric scores and human judgments by employing three distinct types of evaluation protocols: 1) *Correlation analysis:* we employ a standard correlation analysis to determine the extent to which metric scores correlate with human judgments. This analysis highlights the general trend in the relationship between these two variables, offering insights into their alignment; 2) *Classification evaluation:* we conduct a classification evaluation to assess the metrics' capability to distinguish citations based on their support levels. This evaluation specifically measures the accuracy of the metrics in distinguishing between partial, full, and no support scenarios, providing a clear indication of their effectiveness in three-way classification scenarios; and 3) *Retrieval evaluation:* we undertake a retrieval evaluation to assess the effectiveness of metrics in ranking citations according to their support levels. This is motivated by the observation that the previous two evaluation protocols assume citations are present within statements. However, this assumption is not always valid in practical applications [29]. In these cases, faithfulness metrics are adapted to retrieve potential citations from a pool of candidates [6, 7]. The retrieval evaluation thus plays a pivotal role in determining the practical utility of metric adaptations.

In our experiments, we assess seven widely used faithfulness metrics, categorizing them into *similarity-based* and *entailment-based* metrics. Our experimental findings are as follows: 1) no single faithfulness metric consistently outperforms others across three evaluation protocols. This suggests that these protocols are complementary and should be integrated to provide a comprehensive evaluation of metric performance; 2) the best-performing metrics like the entailment-based AUTOAIS show promising results in distinguishing between full-support and no-support scenarios. Nonetheless, they struggle to identify cases of partial support, highlighting the inherent complexities of automated citation evaluation; and 3) in terms of retrieval evaluation, similarity-based metrics, such as BERTSCORE, consistently surpass best-performing entailment-based metrics. This indicates that entailment-based metrics exhibit higher sensitivity to noisy data, which is introduced by a considerable number of irrelevant documents in such scenarios.

Our primary contributions can be summarized as follows:

- **Exploration of fine-grained levels of support in citation evaluation:** to the best of our knowledge, we are the first to systematically investigate the effect of three-category support levels on faithfulness metrics in the task of automated citation evaluation.
- **Introduction of a comparative evaluation framework:** we propose a comparative evaluation framework designed to assess the alignment between metric scores and human judgments. This framework includes correlation analysis, classification, and retrieval evaluation to comprehensively evaluate the metric performance.
- **Comprehensive experimental findings:** 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 these findings, we offer practical recommendations for the development of more effective metrics.

2 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.

2.1 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. According to human evaluation in the dataset, we categorize the citations into three distinct levels of support: full, partial, and no support. We adopt the definition of these levels of support from Liu et al. [29]:

- 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 $F(s_i, c_i) \rightarrow \mathbb{R}^+$. 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.

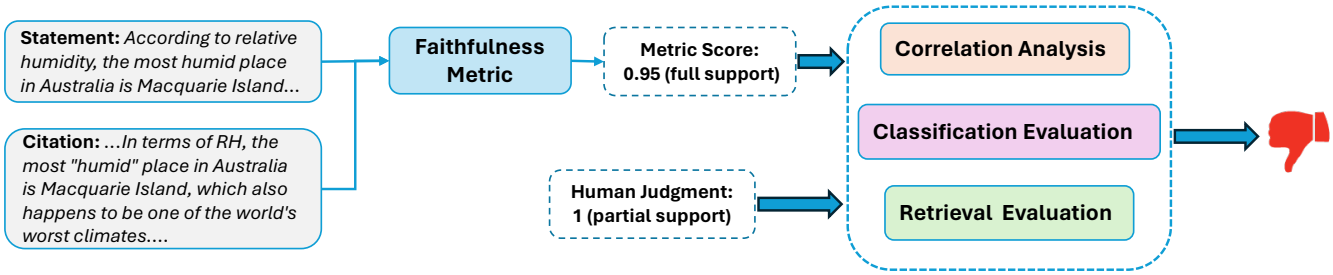

Figure 2: Our proposed comparative evaluation framework. A faithfulness metric assigns scores to given statements and their corresponding citations. Subsequently, our framework comprehensively assesses the alignment between these metric scores and human judgments by employing correlation analysis, classification,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

2.2 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 analysis**, **classification performance**, and **retrieval effectiveness**.

2.2.1 Correlation Analysis. The correlation analysis aims to measure the general trend in the relationship between metric scores and human judgments. Previous research [23, 33] 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). Following them, we adapt correlation analysis to the task of automated citation evaluation. Specifically, Given statements and their associated citations, we assess how well predicted metric scores correlate with human-annotated support levels. To facilitate correlation analysis, we arbitrarily assign support levels $\{\text{FS}, \text{PS}, \text{NS}\}$ to values $\{0, 1, 2\}$. We then utilize standard correlation metrics such as the Pearson correlation coefficient to assess metric effectiveness.

2.2.2 Classification Evaluation. In addition to correlation analysis, we perform classification evaluation to determine the effectiveness of faithfulness metrics 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 [10]. 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 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 such as ROC-AUC. The overall metric performance is then computed by averaging the results across all binary tasks.

2.2.3 Retrieval Evaluation. The objective of retrieval evaluation is to measure the effectiveness of metrics 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 [11, 12, 29]. In post-hoc retrieval, candidate documents are retrieved to form a pool of potential citations using information retrieval techniques [22]. Faithfulness metrics are then employed to rank citations based on their predicted metric scores, aiming to identify the citation with the highest support level [1, 6, 7]. 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 arbitrarily assign support levels $\{\text{FS}, \text{PS}, \text{NS}\}$ to relevance labels $\{2, 1, 0\}$. The metric effectiveness is assessed using standard information retrieval evaluation metrics, such as nDCG. This evaluation also provides a deeper understanding of metric performance in post-hoc citation retrieval scenarios.

3 FAITHFULNESS METRICS

In our experiments, we evaluate seven widely used faithfulness evaluation metrics, dividing them into similarity-based and entailment-based categories. Similarity-based metrics assess the level of support of a citation by measuring the degree of similarity between the citation and the associated statement. In contrast, entailment-based metrics leverage natural language inference (NLI) models [40] to estimate the support level based on the likelihood that the citation entails the statement.

3.1 Similarity-Based Metrics

BERTSCORE [45] adopts BERT [2] 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 [33], We use recommended `deberta-xlarge-mnli` [8] as the backbone of BERTSCORE.

BARTSCORE [42] adopts BART [25] 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

Table 1: Data statistics of the VeJudge dataset. The dataset comprises 12,681 statement-citation pairs. Each pair has been annotated by human assessors based on three categories: full, partial, and no support.

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

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 [9] as the backbone of BARTSCORE.

3.2 Entailment-Based Metrics

FACTCC [23] 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 which contains simulated examples with different factual errors [23]. This metric is also widely used for faithfulness evaluation in abstractive text summarization.

SUMMAC [24] is a RoBERTa-based model [30] that is fine-tuned on NLI datasets. In this metric, a source text and its generated text are split into sentences. Entailment scores for all source/generated sentence pairs are then computed. Finally, the metric aggregates the scores of all pairs to obtain the final faithfulness score. The metric has two variants: (i) $\text{SUMMAC}_{\text{ZS}}$ is a zero-shot version that is only pre-trained on NLI datasets; and (ii) $\text{SUMMAC}_{\text{Conv}}$ adds extra convolutional layers and is further fine-tuned on synthetic training data proposed in Kryscinski et al. [23]. We include both variants in our experiments.

AUTOAIS [6, 10] is a T5-11B [34] model trained on a collection of NLI datasets, which is commonly used in recent automated citation evaluation [1, 6]. As the original output of AUTOAIS is a numeric, either "1" (faithful) or "0" (unfaithful), we use the generated token probability of "1" as the predicted metric score.

ALIGNSCORE [44] further fine-tunes a RoBERTa-based model [30] with a unified alignment loss function. To this end, a unified dataset containing a variety of related natural language processing datasets, such as NLI, question answering, and fact verification datasets, have been collected. In this work, we adapt the `large` version as it demonstrates the best performance.

4 EXPERIMENTS

In this section, we provide a description of the dataset statistics and the data processing method. Subsequently, we discuss the evaluation metrics incorporated within our proposed framework, which assess the performance of faithfulness metrics in alignment with human judgments.

4.1 Datasets

4.1.1 Data Statistics. In the experiments, we employ the dataset of verifiability judgments [29] as our evaluation benchmark, referred to as VeJudge. This 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. These categories indicate whether a citation provides full, partial, or no support to the associated statement. The data statistics are illustrated in Table 1. Notably, for citations classified under the full or partial support categories, human assessors additionally extract explicit evidence from the citation that substantiates the associated statement.

4.1.2 Data Processing. While the VeJudge 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 words. This limitation necessitates input truncation, potentially compromising the reliability of faithfulness metrics. To mitigate this issue, we adopt a strategy similar to previous studies [44]. 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 construct human judgments on the text chunks, we employ the Jaccard similarity index to identify text chunks containing human-annotated evidence, classifying them as either fully or partially supporting text chunks.

4.2 Meta-Evaluation

For correlation evaluation, we report partial Pearson, Spearman, and Kendall coefficients, as recommended by previous research [33]. In terms of classification evaluation, following previous studies [10, 31], we report the Receiver Operating Characteristic-Area Under Curve (ROC-AUC) score as it obviates the need for manual threshold setting for each binary classification task. Moreover, to capture the comprehensive performance across all binary classification tasks, we compute and report the macro-averaged ROC-AUC score. For retrieval evaluation, we report standard information retrieval metrics: mean reciprocal rank (MRR) and normalized discounted cumulative gain (nDCG@n) scores.

5 RESULTS & ANALYSES

In this section, we discuss the results of the performance of faithfulness metrics across three distinct evaluation protocols. Following this, we integrate the observations derived from these evaluation protocols to discuss our main implications, offering practical recommendations to enhance metric effectiveness in automated citation evaluation.

5.1 Correlation Results

The correlation analysis 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. Specifically, AUTOAIS achieves the highest Pearson coefficient, recording a value of 0.604, marginally surpassing the second-best BARTSCORE, which posts a coefficient of 0.593; 2) there is a noticeable variation in correlation trends among high-performing metrics. Notably, AUTOAIS shows a more substantial Pearson

Table 2: Partial correlation coefficients between human-annotated levels of support and faithfulness metric scores on the VeJudge dataset. The best correlations are marked in bold.

METRIC	Pearson	Spearman	Kendall
FACTCC	0.108	-0.018	-0.008
SUMMAC$_{ZS}$	0.326	0.143	0.106
BERTSCORE	0.512	0.218	0.165
ALIGNSCORE	0.551	0.275	0.199
BARTSCORE	0.593	0.279	0.211
AUTOAIS	**0.604**	0.407	0.297
SUMMAC$_{Conv}$	0.565	**0.444**	**0.342**

Table 3: Classification performance of faithfulness metrics regarding ROC-AUC score (%) on the VeJudge dataset. The overall performance is the macro-averaged performance of three binary classification settings. The best scores are marked in bold.

METRIC	FS-VS-NS	FS-VS-PS	PS-VS-NS	OVERALL
FACTCC	68.77	62.58	56.81	62.72
SUMMAC$_{ZS}$	78.37	72.96	58.12	69.82
SUMMAC$_{Conv}$	85.32	78.74	62.57	75.54
BARTSCORE	87.65	75.42	71.94	78.34
ALIGNSCORE	90.97	81.41	70.53	80.97
BERTSCORE	91.92	75.94	**79.89**	82.58
AUTOAIS	**92.65**	**82.31**	74.21	**83.06**

correlation, whereas SUMMAC$_{Conv}$ outperforms in Spearman and Kendall correlations. This divergence might be attributed to the Pearson coefficient assuming linear relationships between two variables. Such an assumption is often invalid in automated citation evaluation, rendering the Pearson coefficient less suitable for capturing the true relationships between metric scores and human judgments; and 3) Generally, most metrics display relatively low Spearman and Kendall correlations compared to their Pearson correlations. For instance, SUMMAC$_{Conv}$ achieves the highest Spearman and Kendall correlations, with values of 0.445 and 0.297 respectively, which are considerably lower than its Pearson correlation of 0.565. This disparity indicates that the metric scores of the best-performing metrics do not correlate well with human judgments, highlighting the limitations of existing metrics in scenarios involving fine-grained levels of support.

5.2 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, with the highest ROC-AUC score of 92.65, exemplifies significant discriminability between full support and no support instances. This performance can be attributed to its extensive parameters, comprising 11 billion parameters, in contrast to the hundreds of millions of other metrics; 2) a pronounced decline in classification performance is observed across the other two settings. For instance, when comparing the FS-VS-NS and PS-VS-NS settings, the ROC-AUC score of AUTOAIS diminishes from 92.65 to 74.21. This decline indicates that even the best-performing metric struggles with granular sensitivity to varying levels of support; and 3) while entailment-based AUTOAIS generally surpasses other metrics in overall performance, it is outperformed by similarity-based BERTSCORE in the PS-VS-NS setting. Interestingly, while most metrics exhibit their lowest performance in this particular setting, BERTSCORE shows its least effectiveness in another setting, FS-VS-PS. This underscores the unique prediction behaviors displayed by different types of metrics across the binary classification settings.

Table 4: Retrieval performance of faithfulness metrics regarding MRR and nDCG@n scores on the VeJudge dataset. Note that we assign relevance labels 2, 1, and 0 to full, partial, and no support, respectively. The best scores are marked in bold.

METRIC	MRR	nDCG@5	nDCG@10	nDCG@20
FACTCC	0.656	0.648	0.689	0.710
SUMMAC$_{ZS}$	0.737	0.729	0.759	0.776
SUMMAC$_{Conv}$	0.776	0.772	0.798	0.811
ALIGNSCORE	0.847	0.842	0.863	0.869
AUTOAIS	0.846	0.846	0.865	0.872
BERTSCORE	0.867	0.865	0.881	0.887
BARTSCORE	**0.881**	**0.878**	**0.891**	**0.897**

5.3 Retrieval Results

Table 4 presents the results of the retrieval evaluation. The key findings are as follows: 1) similarity-based metrics, BARTSCORE and BERTSCORE, outperforms other entailment-based metrics in both MRR and nDCG@n. For instance, entailment-based AUTOAIS exhibits weaker MMR scores than BARTSCORE (0.846 vs. 0.881). This is likely because entailment-based metrics are more sensitive to noisy information than similarity-based metrics, as many irrelevant documents exist in retrieval scenarios. This suggests the need for the robustness improvements of metrics in post-hoc retrieval scenarios; 2) a significant correlation is observed between MRR and nDCG@n scores across all metrics. Notably, nDCG@n effectively captures the performance variations as the number of text chunks increases (i.e., 5, 10, 20). For instance, as the chunk count increases, BARTSCORE shows a marginal performance improvement (from 0.878 to 0.897), while FACTCC—the least performing metric—exhibits a more pronounced enhancement (from 0.648 to 0.710); and 3) we observe an intriguing shift in performance ranking when comparing metrics between classification and retrieval evaluations. For instance, BARTSCORE ascends from the fourth place to the first while AUTOAIS sees a decline from the top to the third. This divergence highlights that these evaluations offer unique insights into the capabilities of metrics.

5.4 Implications

Based on the evaluation results, we observe that no single faithfulness metric consistently excels across three distinct evaluation protocols. For instance, SUMMAC$_{\text{Conv}}$ achieves the highest performance in correlation analysis, yet it under-performs in classification and retrieval evaluations. This disparity suggests that these evaluation protocols are complementary and should be integrated to comprehensively assess the effectiveness of different metrics. This inconsistency also reveals that the best-performing metrics are insufficiently effective in addressing fine-grained support level scenarios. Particularly, they fail to effectively distinguish partial support from either full or no support scenarios. Furthermore, when comparing entailment-based metrics with similarity-based metrics, a notable shift in performance ranking is observed between the classification and retrieval evaluations. Specifically, the similarity-based BARTSCORE advances from fourth to first place, whereas the entailment-based AUTOAIS declines from the top position to third place. This shift may be attributed to the higher sensitivity of entailment-based metrics to noisy data, which is introduced by irrelevant documents in retrieval scenarios. This suggests the need for improving the robustness of entailment-based metrics against irrelevant documents.

Consequently, we 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 in various information retrieval tasks [20].

6 RELATED WORK

This section outlines two lines of related research: faithfulness evaluation metrics and citation evaluation.

6.1 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) [13, 16–18, 49, 50, 52], particularly in abstractive summarization [14, 15, 23, 32, 47, 48]. 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 [5, 10, 24, 44]. Similarity-based metrics, such as BERTScore [45] and BARTScore [42], quantify text similarity and have demonstrated robust performance in faithfulness evaluation [10, 33]. QA-based metrics utilize a combination of question generation and question answering to estimate faithfulness levels [3, 4, 36, 39]. In this work, we exclude QA-based metrics from our work, following recent works suggesting the challenging limitations in these metrics [21]. We focus on the extrinsic evaluation of faithfulness metrics against human judgments in scenarios requiring fine-grained citation support.

6.2 Citation Evaluation

Citation evaluation seeks to enhance the trustworthiness of retrieval-augmented LLMs by verifying the support provided by citations to the generated statements [11, 12, 35, 43, 46]. 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 [19, 26, 28, 37, 38, 41]. 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.

7 CONCLUSION

LLMs are susceptible to generating hallucinated content, motivating the research on the integration of retrieval augmentation mechanisms to enhance the verifiability of generated statements through in-line citations. However, evaluating how well these citations support the statements remains a major challenge due to the labor-intensive nature of manual citation evaluation. Consequently, faithfulness metrics have been adopted to automate this evaluation, primarily determining citation support in a binary classification scenario. This paper proposes a comparative evaluation framework to explore the efficacy of faithfulness metrics beyond the binary scenario by examining three levels of citation support: full, partial, and no support. Our framework assesses the alignment between metric scores and human judgments across three evaluation protocols: correlation analysis, classification evaluation, and retrieval evaluation. 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 in identifying partial support scenarios. Based on these findings, we further provide practical suggestions for the development of more effective metrics in automated citation evaluation.

REFERENCES

[1] Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037* (2022).

[2] 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)*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[3] 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*. 5055–5070. https://doi.org/10.18653/v1/2020.acl-main.454

[4] 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. 2587–2601. https://doi.org/10.18653/v1/2022.naacl-main.187

[5] 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*. 2214–2220. https://doi.org/10.18653/v1/P19-1213

[6] 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. 2023. 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)*. 16477–16508.

[7] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[8] 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*. https://openreview.net/forum?id=XPZIaotutsD

[9] 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*. 1693–1701. https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html

[10] 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*. 3905–3920. https://doi.org/10.18653/v1/2022.naacl-main.287

[11] Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training Language Models to Generate Text with Citations via Fine-Grained Rewards. *CoRR* abs/2402.04315 (2024). arXiv:2402.04315

[12] Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A Key to Building Responsible and Accountable Large Language Models. *CoRR* abs/2307.02185 (2023). https://doi.org/10.48550/arXiv.2307.02185 arXiv:2307.02185

[13] 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*, Vol. 33. 8449–8456.

[14] Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. 2021. GPT2MVS: Generative Pre-trained Transformer-2 for Multi-modal Video Summarization. In *Proceedings of the International Conference on Multimedia Retrieval*. 580–589.

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

[16] Jia-Hong Huang, Chao-Chun Yang, Yixian Shen, Alessio M Pacces, and Evangelos Kanoulas. 2024. Optimizing Numerical Estimation and Operational Efficiency in the Legal Domain through Large Language Models. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

[17] 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. 2021. 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*. 2442–2452.

[18] Jia-Hong Huang, Hongyi Zhu, Yixian Shen, Stevan Rudinac, Alessio M. Pacces, and Evangelos Kanoulas. 2024. A Novel Evaluation Framework for Image2Text Generation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval, LLM4Eval Workshop*.

[19] Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024. Learning Fine-Grained Grounded Citations for Attributed Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*. 14095–14113.

[20] 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 (2022).

[21] 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*. 132–146.

[22] 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)*. 6769–6781.

[23] 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)*. 9332–9346. https://doi.org/10.18653/v1/2020.emnlp-main.750

[24] 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 (2022), 163–177. https://doi.org/10.1162/tacl_a_00453

[25] 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*. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[26] Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024. Towards Verifiable Generation: A Benchmark for Knowledge-Aware Language Model Attribution. In *Findings of the Association for Computational Linguistics ACL 2024*. 493–516.

[27] Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024. 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)*. 5453–5471. https://doi.org/10.18653/v1/2024.naacl-long.305

[28] Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How Hard Is Automatic Attribution Evaluation?. In *Findings of the Association for Computational Linguistics ACL 2024*. 14919–14935.

[29] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7001–7025. https://doi.org/10.18653/v1/2023.findings-emnlp.467

[30] 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* (2019).

[31] 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)*. 12788–12812.

[32] 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*. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[33] 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*. 4812–4829. https://doi.org/10.18653/v1/2021.naacl-main.383

[34] 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 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[35] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics* 49, 4 (2023), 777–840.

[36] 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*. 6594–6604. https://doi.org/10.18653/v1/2021.emnlp-main.529

[37] 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* (2024).

[38] 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 (2023). https://doi.org/10.48550/ARXIV.2312.09075 arXiv:2312.09075

[39] 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*. 5008–5020. https://doi.org/10.18653/v1/2020.acl-main.450

[40] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1112–1122.

[41] 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)*. 6237–6251.

[42] 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*. 27263–27277. https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html

[43] 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*. 4615–4635. https://doi.org/10.18653/v1/2023.findings-emnlp.307

[44] 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)*. 11328–11348.

[45] 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*. https://openreview.net/forum?id=SkeHuCVFDr

[46] Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-Hong Huang, and Evangelos Kanoulas. 2024. Towards Fine-Grained Citation Evaluation in Generated Text: A Comparative Analysis of Faithfulness Metrics. In *International Natural Language Generation Conference (INLG)*.

[47] Weijia Zhang, Jia-Hong Huang, Svitlana Vakulenko, Yumo Xu, Thilina Rajapakse, and Evangelos Kanoulas. 2024. Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization using Large Language Models. In *Proceedings of the 2024 International Conference on Pattern Recognition (ICPR)*.

[48] 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)*. https://doi.org/10.48550/ARXIV.2405.05109

[49] 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* (2021).

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

[51] 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. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR* abs/2309.01219 (2023). https://doi.org/10.48550/ARXIV.2309.01219 arXiv:2309.01219

[52] 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*. 978–987.